Estadística Bayesiana

Manuel Mendoza R. y Pedro Regueiro M.

Departamento de Estadística
Instituto Tecnológico Autónomo de México

2011

Índice general

Capítulo 1

Introducción

1.1. Preliminares

La Estadística, tal como se presenta en los cursos más convencionales, puede definirse como un conjunto de técnicas cuyo propósito es la descripción de fenómenos que se manifiestan a través de datos que presentan *variabilidad.* Esta definición delimita el ámbito de acción de la disciplina -los fenómenos que presentan variabilidad- y al mismo tiempo, establece su objetivo último: la *descripción.* Así, toda la Estadística es descriptiva y, en particular, la *Inferencia Estadística* se ocupa del problema de descripción en el caso en que sólo es posible observar una fracción -o muestra- de la colección completa de datos que el fenómeno de interés puede producir (habitualmente denominada la población). En general, las descripciones que produce la Estadística se llevan a cabo a través del cálculo de resúmenes de la información disponible. Cuando se trata de un problema de inferencia, la descripción que se obtiene siempre es aproximada puesto que se basa sólo en una parte de toda la información que podría, al menos potencialmente, ser utilizada. En esas condiciones, hay dos retos que es necesario enfrentar. En primer lugar, idealmente, la muestra seleccionada debería reproducir exactamente las caraterísticas de la población entera. En los términos habituales en la literatura, la muestra debería ser *representativa.* En la práctica, sin embargo, nunca es posible comprobar la representatividad de una muestra ya que ello implicaría el conocimiento de la población completa. Por tal razón, en el mejor de los casos, se cuenta con muestras que aproximan el comportamiento de la población y conducen, como ya se indicó, a descripciones aproximadas. El segundo reto consiste precisamente en proveer una medida del grado de aproximación que tienen las inferencias.

Para fortuna de los usuarios de los métodos estadísticos, estos dos retos han sido razonablemente resueltos gracias a la introducción de la *selección probabilística* -por sorteo- de muestras. En su versión más simple, este esquema asigna a cada uno de los elementos de la población la misma probabilidad de aparecer en la

muestra y la extracción de cada elemento se realiza independientemente de la de cualquier otro. Como consecuencia, los rasgos más frecuentes en la población son los que aparecen con mayor probabilidad en la muestra. Adicionalmente, se elimina cualquier sesgo intencional al remitir la decisión de la selección a un mecanismo exógeno y, conceptualmente, se obtiene una muestra (*aleatoria*) que está formada por una colección $\underline{X}_{(n)} = \{X_1, X_2, \cdots, X_n\}$ de variables aleatorias independientes e idénticamente distribuídas de acuerdo con una función de distribución común $F(x)$. En estas condiciones, la descripción del fenómeno es equivalente a la descripción de $F(x)$.

Los problemas de inferencia que serán tratados en este texto pertenecen al dominio de lo que se conoce como Inferencia Estadística *Paramétrica*. Este es el entorno que se genera cuando la función de distribución de interés pertenece a una familia $\mathcal{F}_\theta$ de distribuciones donde todos los elementos tienen la misma forma funcional, y se distinguen por el valor de un parámetro (índice) θ que toma valores en un conjunto $\Theta \in \Re^k$ para un valor fijo de k. Esta simplificación es muy importante porque reduce la búsqueda de un elemento, $F(\cdot)$, en un espacio de funciones de dimensión infinita a la búsqueda de un vector, θ, en un espacio euclidiano de dimensión finita. Una vez determinado el valor de θ, se puede identificar el elemento F_θ en $\mathcal{F}_\theta$.

1.2. Los límites de la Estadística Frecuentista

Las técnicas que se presentan en un curso habitual de *Estadística Matemática,* corresponden a lo que se conoce genéricamente como *Estadística Frecuentista* en virtud de que interpreta la Probabilidad como un límite de frecuencias re- lativas. Esta interpretación es evidente, en particular, cuando se definen los conceptos y criterios para evaluar la calidad de las inferencias (significancia, confianza e insesgamiento, por ejemplo). Esta no es la única interpretación posible y la idea de contar con mecanismos de inferencia que consideren un versión más general de la Probabilidad será objeto de discusón en capítulos posteriores.

Por otra parte, es común la percepción de que la Estadística (Frecuentista) se articula a través de una serie de reglas, métodos y algoritmos, cada uno de los cuales tiene sus propios méritos y ventajas pero que no necesariamente constituyen un cuerpo compatible y coherente de piezas de conocimiento. Especialmente entre sus usuarios, ocurre que suelen visualizar a la Estadística como una vasta colección de algoritmos (*fórmulas*) cuyo empleo es apropiado en forma casuística.

Finalmente, existe una importante colección de ejemplos en los que las técnicas estadísticas frecuentistas producen resultados que arrojan una sombra de duda sobre el carácter general de los conceptos en que se basan; en ocasiones estas dudas sugieren precaución y modificaciones, pero en algunos casos extremos, cuestionan la naturaleza misma de los conceptos. Algunos de estos ejemplos pueden considerarse extremos o patológicos, pero otros son sorprendentemente

generales (en la lista de ejercicios de este capítulo se encuentra una pequeña muestra de este tipo de ejemplos). Habiendo referido estos hechos, es prudente aclarar que un Estadístico competente, frecuentista o no, debería ser capaz de navegar con eficacia las regiones de aguas procelosas que pudiese presentar el océano de la Estadística.

1.3. La conveniencia de una Teoría Estadística

El trabajo de un puñado de brillantes académicos, entre los que destacan Karl Pearson (1857-1936), Ronald A. Fisher (1890-1962), Egon Pearson (1895-1980), Jerzy Neyman (1894-1981), Harald Cramér (1893-1985), David Blackwell (1919-2010) y Calyampudi R. Rao (1920-) hizo posible que a lo largo de un periodo de aproximadamente 30 años que inició alrededor de 1915, los métodos de la Estadística fuesen encontrando respaldo en los principios matemáticos. Es entonces cuando propiamente nace la Estadística *Matemática*.

Sin embargo, este notable avance de matematización que consolidó la Estadística Frecuentista, no fructificó en una *Teoría*, en el sentido axiomático del término, como sí ocurrió, en cambio, con la Probabilidad en 1933 cuando Andrei Kolmogorov (1903-1987) postuló un conjunto de axiomas o principios básicos que encapsulan la naturaleza de la disciplina en su totalidad y a partir de los cuales se pueden deducir todos sus resultados organizados en un cuerpo coherente de conocimientos sin contradicciones ni paradojas.

El surgimiento de una Teoría Estadística o una Teoría de la Inferencia Estadística, habría de aguardar un tiempo más, hasta la década de los años 50 cuando aparece el libro *The Foundations of Statistics* de Leonard J. Savage (1917-1971) que recoge el fruto de su propio trabajo y el de otros estadísticos como Frank Ramsey (1903-1930), Bruno de Finetti (1906-1985) y Dennis V. Lindley (1923-). Ahí se presentan los *Postulados de la Teoría de la Decisión Personal* que actualmente son mejor conocidos como *Axiomas de Coherencia,* y a partir de esa base se establecen, como indica el título del libro, los fundamentos de la Estadística. En otras palabras, se construye una Teoría Axiomática de la Inferencia Estadística.

Probablemente, la consecuencia más espectacular de este esfuerzo fue el hecho de que la teoría desarrollada, si bien incluye, como casos particulares, algunas ideas, ciertos conceptos y determinados resultados específicos de la poderosa Estadística Frecuentista, en su gran mayoría esta disciplina sólo tiene cabida en el nuevo marco como un caso límite y, en una variedad de casos se puede probar que sus procedimientos simplemente violan alguno de los axiomas de coherencia. Así, la nueva teoría nació en conflicto con la escuela predominante de pensamiento estadístico. Más aún, retomó y revaloró ideas y conceptos que habían evolucionado desde finales del siglo XVIII y hasta principios del siglo XX para describir la naturaleza de los fenómenos inciertos.

El exponente más brillante de ese enfoque, con 150 años de antiguedad, fue Pie-

rre Simon de Laplace (1749-1827), quien le dió su nombre: *Probabilidad Inversa*. Laplace elaboró durante años sobre el tema, fue su principal promotor y, en particular, discutió con detalle sus ideas al respecto en obras como *Mémoire sur la probabilité des causes par les évènements* de 1774. En algún momento, Laplace dio crédito a un autor que le antecedió en el tratamiento del tema, así fuera muy puntual y sin gran repercusión en su tiempo. Ese autor no vió publicado su trabajo ya que éste apareció en forma póstuma; su nombre era Thomas Bayes (1702-1761).

Como un anotación histórica curiosa es interesante consignar que Savage, en los años 50, desarrollaba su actividad académica en la Universidad de Chicago, donde fue contemporáneo de distintos economistas que habrían de ser muy reconocidos, en particular por distintos trabajos relacionados con la *Teoría de Elección Racional* que, en cierta forma, comparte orígenes con los axiomas de coherencia. De su relación con los economistas de la universidad dan cuenta, por ejemplo, sus publicaciones conjuntas con Milton Friedman (1912-2006) sobre funciones de utilidad, y las anécdotas sobre su papel como profesor de Harry Markowitz (1927-).

Recientemente se ha dado por llamar *Neo Bayesiana* a la Teoría originada por Ramsey, De Finneti, Lindley y Savage que ha tenido un crecimiento espectacular, especialmente a partir de los ochenta. En una primera fase, la investigación en la materia se orientó al refinamiento de los fundamentos; posteriormente, al desarrollo de métodos Bayesianos para la aplicación en la práctica. Fue esta segunda etapa en la que comprobó que la fortaleza metodológica con frecuencia tenía asociada el costo de la dificultad para obtener resultados con expresiones analíticas cerradas. La tercera etapa, que inició en los 90, se ha caracterizado por un crecimiento explosivo de las aplicaciones complejas en las más diversas áreas, gracias a la incorporación de técnicas de aproximación numérica vía simulación, especialmente a través de cadenas de Markov.

El propósito de este texto es presentar una versión simple pero actualizada de los resultados de las dos primeras etapas y una revisión general de las ideas que gobiernan el desarrollo de la tercera fase. El énfasis se concentra en el procedimiento de construcción de esta Teoría de la Inferencia Estadística (ahora conocida como *Bayesiana*) así como en ilustrar las principales implicaciones generales que tiene en la práctica.

1.4. Ejercicios

Ejercicio 1.1. *Suponga que X es una variable aleatoria con distribución Poisson y media λ. Si cuenta con una única observación x de esta variable, demuestre que el único estimador insesgado de $\theta = P(X = 0)$ está dado por*

$$T(x) = \begin{cases} 1 & si\ x = 0 \\ 0 & en\ otro\ caso. \end{cases}$$

Asimismo, demuestre que en este caso el único estimador insesgado para θ^2, está dado por

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

¿Qué opinión le merecen estos estimadores?

Ejercicio 1.2. *Suponga que cuenta con una muestra aleatoria de tamaño n para una variable aleatoria X, cuya distribución es Normal con media μ_1 y varianza σ^2. Suponga que además cuenta con una muestra aleatoria de tamaño m para otra variable aleatoria Y, cuya distribución es también Normal con media μ_2 y la misma varianza.*

Si las dos muestras son independientes entre sí y además por facilidad se considera el caso en que $\sigma^2 = 1$ y $n = m$, encuentre un estimador puntual para el parámetro $\rho = \frac{\mu_1}{\mu_2}$ bajo el supuesto de que $\mu_2 \neq 0$. Encuentre además un intervalo de confianza de nivel $(1-\alpha)100\,\%$ para ρ.¿Cómo se comporta este intervalo cuando α tiende a cero?

Ejercicio 1.3. *Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria de una distribución Normal con media μ y varianza $\sigma^2 = 1$. Considere la situación en la que desea contrastar las hipótesis*

$$H_0 :\ \mu = 0 \quad vs. \quad H_1 :\ \mu = 1$$

con un nivel de significancia $\alpha = 0{,}05$. Construya la región de rechazo óptima para llevar acabo este contraste. Una vez con esta región diga ¿Qué ocurre con las probabilidades de error de tipo I y II si el tamaño de muestra tiende a infinito? ¿Qué opinión le merece este comportamiento?

Ejercicio 1.4. *Considere la situación en la que se cuenta con los resultados de n lanzamientos independientes de una misma moneda, de los cuales r de ellos resultan en sol, que se define como un éxito.*

Determine un estimador insesgado para θ, la probabilidad de éxito, suponiendo que estos datos proceden de un modelo Binomial con parámetros (n, θ) donde se ha observado $X{=}r$.

Ahora, alternativamente suponga que los datos provienen de una variable aleatoria Y con distribución Binomial Negativa (r, θ) para la cual se ha observado $Y = n$. Calcule un estimador insesgado para θ con este modelo alternativo.

¿Coinciden ambos estimadores? ¿Qué opina de este resultado?

Ejercicio 1.5. *En las elecciones federales de julio de 2009, el Instituto Federal Electoral seleccionó, de entre las 139 959 casillas de votación que se instalaron en todo el país, una muestra aleatoria de 900 casillas y en cada casilla seleccionada examinó el material electoral (boletas de votación, actas y tinta indeleble) para verificar que este cumplía con las normas de calidad y marcas de seguridad*

que la normatividad electoral exige. Como resultado, reportó que en todas las casillas de la muestra el material cumplía con las condiciones requeridas.

Con esta información, ¿Cuál sería su estimación puntual para la proporción θ de casillas en todo el país, cuyo material cumplía la normatividad? ¿Cuál resultaríía ser un intervalo de confianza -así sea aproximado- para θ con un nivel $(1-\alpha)100\,\%$? ¿Qué opina de sus resultados?

Ejercicio 1.6. *En el artículo:* When Did Bayesian Inference Become "Bayesian"? *Stephen E. Fienberg presenta una revisión histórica del desarrollo del paradigma Bayesiano en la Inferencia Estadística. Lea este artículo y redacte un resumen crítico.*

Capítulo 2

Problemas de decisión

Definición 2.0.1. *Un* **problema de decisión** *es la situación en la que un personaje (tomador de decisiones) se enfrenta a un conjunto de decisiones, D, de entre las cuales debe seleccionar una y sólo una de ellas.*

Principio básico. *Una solución (decisión) es mejor en la medida que produce más satisfacción al tomador de decisiones. Las decisiones seran juzgadas por sus consecuencias.*

Considere los siguientes objetos:

- $D = \{d_1, d_2, \ldots, d_k\}$, el *conjunto de decisiones.*
- Para cada $d_i \in D$, $E_i = \{E_{i1}, E_{i2}, \ldots, E_{in_i}\}$ una partición del evento seguro (Ω) y $\mathcal{E} = \bigcup_{i=1}^{k} E_i$ el *conjunto de eventos inciertos relevantes.*
- Para cada $d_i \in D$, $C_i = \{c_{i1}, c_{i2}, \ldots, c_{in_i}\}$ donde c_{ij} es la consecuencia de elegir i y que suceda j; y $C = \bigcup_{i=1}^{k} C_i$ el *conjunto de consecuencias.*
- Una relación binaria $\succ$ definida sobre C tal que $c_{ij} \succ c_{kl} \Leftrightarrow c_{ij}$ es más preferido que c_{kl}.

Un problema de decisión está completamente caracterizado por $(D, \mathcal{E}, C, \succ)$.

Notar que:

- $d_i \leftrightarrow C_i \ \forall\, i \in \{1, 2, \ldots, k\}$, i.e. a cada acción se le asocia un conjunto de posibles consecuencias.
- C no necesariamente es un subconjunto de $\Re$.
- E_i partición de $\Omega \Rightarrow E_{ij} \bigcap E_{ik} = \emptyset \quad \forall j \neq k \quad$ y $\quad \bigcup_{j=1}^{n_i} E_{ij} = \Omega \quad \forall i.$

2.1. Problemas de decisión sin incertidumbre

Definición 2.1.1. *Un problema de decisión se dice que es* **sin incertidumbre** *si para cada decisión la consecuencia respectiva es segura. Es decir si C_i consta de un solo elemento para toda i.*

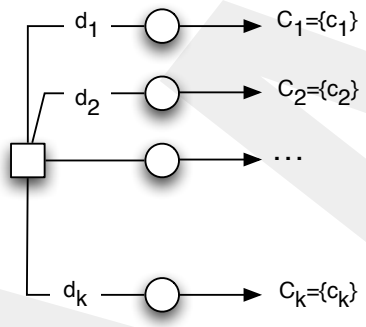


Figura 2.1: Árbol de decisión sin incertidumbre: consecuencias seguras.

Bajo las condiciones de un problema de decisión sin incertidumbre, es posible encontrar un algoritmo de solución:

- Se comparan las consecuencias para identificar la más preferida (para el tomador de decisiones).
- Se identifica la opción asociada a esa consecuencia y se toma como la solución.

Ejemplo 2.1.1. *Suponga que cuenta con una hoja de lámina de acero de superficie S metros cuadrados y que desea construir un contenedor (sin tapa) de base cuadrangular, de lado l y altura h, utilizando todo el material de forma que tenga máxima capacidad.*

Definiendo:

$$D = \{d_{hl} | d_{hl} = \textit{diseño de } l*h, \textit{ con } h,l > 0;\ S = l^2 + 4lh\}$$

y bajo el supuesto de que el beneficio de un diseño depende exclusivamente de su volumen, entonces

$$C = \{c_{hl} | c_{hl} = \textit{volumen del contenedor con diseño } d_{hl};\ c_{hl} = l^2 h\}$$

En este problema las consecuencias son de entrada numéricas y, más aún, sucede que $c_{h'l'} \succ c_{hl} \Leftrightarrow c_{h'l'} > c_{hl}$, lo que conduce a resolver el problema:

$$\begin{aligned} &\max_{(l,h)} f(l,h) = l^2 h \\ s.a. \quad &l^2 + 4lh = S \end{aligned}$$

Cuya solución, en términos de S, puede ser obtenida por métodos estándar de cálculo y está dada por $l^\star = \frac{\sqrt{S}}{\sqrt{3}}$ *y* $h^\star = \frac{\sqrt{S}}{2\sqrt{3}}$.

En este ejemplo es interesante observar que el conjunto D de diseños está parametrizado por h y l y que, de hecho, se puede representar como un conjunto en $\Re^2$ tal como se exhibe en la figura 2.2.

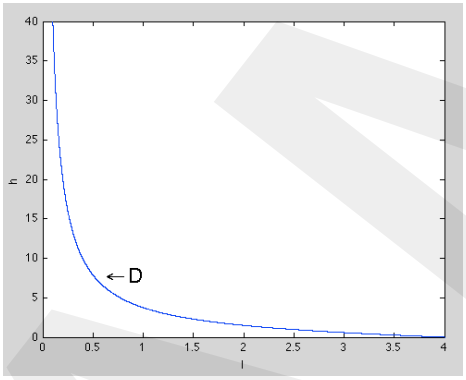


Figura 2.2: Región factible para (l, h) con $S = 16$.

Más aún, dada la restrición que define a estos diseños, las consecuencias en C se pueden expresar como función de h (e implícitamente de l). Este hecho se ilustra en la figura 2.3

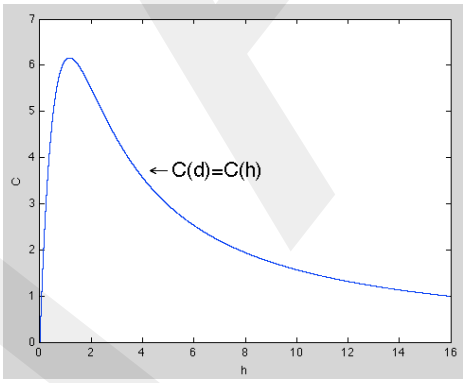


Figura 2.3: Conjunto de consecuencias en función de h, con $S = 16$.

Como se verá a continuación, en la práctica los problemas de decisión más interesantes son aquellos que contienen incertidumbre.

2.2. Problemas de decisión con incertidumbre

Definición 2.2.1. *Un Problema de decisión se dice que es* **con incertidumbre** *cuando para al menos una decisión existe más de una posible consecuencia. Es decir, si existe al menos una i tal que C_i consta de dos o más elementos.*

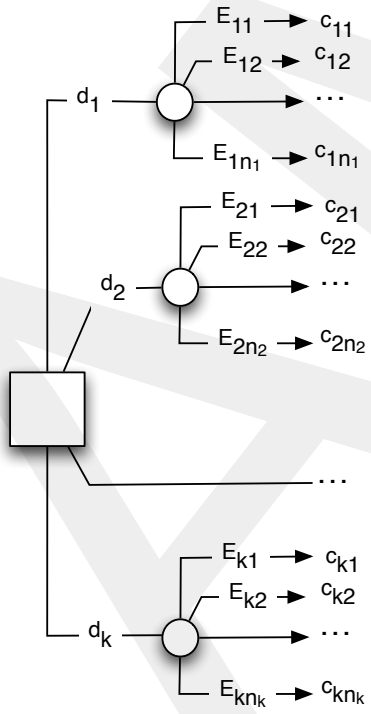


Figura 2.4: Árbol de decisión con incertidumbre: una vez elegido d_j lo único que se sabe es que ocurrirá una y sólo una de las respectivas consecuencias $c_{j1}, c_{j2}, \ldots, c_{jn_j}$.

Bajo las condiciones de un problema de decisión con incertidumbre, no es inmediato que se deba utilizar directamente el algoritmo de solución descrito en la sección 2.1. Una idea general que se ha ensayado para resolver un problema de este tipo, es reducirlo a otro problema sin incertidumbre. Como veremos más adelante, esta idea ha inducido a diversos criterios para la solución de problemas de decisión.

Antes de presentar estos criterios, es interesante considerar la siguiente situación. Suponga que las particiones $E_i : \ i = 1, 2. \ldots, k$ son todas iguales. Esto significa que los eventos inciertos que condicionan las consecuencias de todas las decisiones son los mismos. Esto no es el caso general, pero resulta, como se co-

mentará en su momento, que esta estructura se puede adoptar conceptualmente sin pérdida de generalidad.

Si $E_i = E\ \forall i$, entonces el problema de decisión, además del árbol respectivo, admite una representación gráfica tal como se muestra en la tabla 2.1.

	E_1	E_2	$\dots$	E_n
d_1	c_{11}	c_{12}	$\dots$	c_{1n}
d_2	c_{21}	c_{22}	$\dots$	c_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_k	c_{k1}	c_{k2}	$\dots$	c_{kn}

Tabla 2.1: Representación tabular de las consecuencias para una partición común del espacio de eventos inciertos.

En la tabla anterior surge una idea interesante ¿Qué sucede si al comparar dos renglones (correspondientes a d_i y $d_{i'}$) ocurre que, elemento a elemento, por columna, $c_{ij} \succ c_{i'j}$? Claramente, entonces no importa cual evento incierto se presente, la decisión d_i produce consecuencias más preferidas que $d_{i'}$. Evidentemente $d_{i'}$ no puede ser la opción óptima para el tomador de decisiones, puesto que al menos existe otra que, sin duda, es mejor (d_i). Cuando se presenta esta circunstancia, se dice que d_i domina a $d_{i'}$, y que $d_{i'}$ es inadmisible. Por supuesto, en un problema específico es conveniente identificar las opciones inadmisibles y eliminarlas de D.

Como ya se indicó, en la práctica las particiones $E_i :\ i = 1, 2. \dots, k$ no tienen porqué ser iguales, pero el problema de decisión se puede reformular considerando una partición $E^\star$ donde cada elemento de esta nueva partición se construye como la intersección de k eventos, tomando uno de cada E_i. Es claro que algunos elementos serán iguales al vacío, pero en cualquier caso, por construcción, el resultado es una partición común. Y en esos términos se puede enunciar la siguiente definición general.

Definición 2.2.2. *Una decisión $d \in D$ se dice que es* **inadmisible** *si existe $d' \in D$ tal que, para cualquier evento incierto en $\mathcal{E}$ sucede que $d \preceq d'$ y existe un evento incierto $E_i \subseteq \mathcal{E}$ para el cual $d \prec d'$. Se dice tambien que d' domina a d.*

2.3. Algunos criterios de solución para problemas de decisión con incertidumbre

2.3.1. Criterio optimista

De cada grupo de ramas secundarias en el árbol, se eliminan todas excepto aquella con la consecuencia más preferida. Con las ramas sobrevivientes, se resuelve el problema como se haría en ausencia de incertidumbre.

Este criterio equivale a que el tomador de decisiones se considere tan afortunado, que piense que siempre, sin importar la opción que elija, el evento incierto que ocurrirá será aquel que le produzca el mayor beneficio. De este modo, el árbol en la figura 2.4 sería remplazado por el de la figura 2.5.

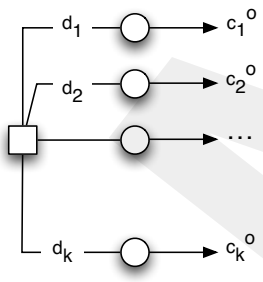


Figura 2.5: Árbol reducido bajo el criterio optimista. Aquí c_i^0 representa la consecuencia más preferida entre las contenidas en $\{c_{i1}, c_{i2}, \ldots, c_{in_i}\}$.

2.3.2. Criterio pesimista (solución minimax)

De cada grupo de ramas secundarias en el árbol, se eliminan todas excepto aquella con la consecuencia menos preferida. Con las ramas sobrevivientes, se resuelve el problema como se haría en ausencia de incertidumbre.

Contrario al criterio anterior, en este criterio el tomador de decisiones se considera tan desafortunado, que piensa que siempre, sin importar la opción que elija, el evento incierto que ocurrirá será aquel que le produzca el menor beneficio. De este modo, el árbol en la figura 2.4 sería remplazado por el de la figura 2.6.

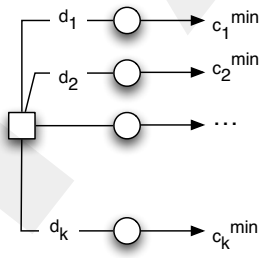


Figura 2.6: Árbol reducido bajo el criterio pesimista. Aquí c_i^{min} representa la consecuencia menos preferida entre las contenidas en $\{c_{i1}, c_{i2}, \ldots, c_{in_i}\}$.

2.3.3. Criterio de la consecuencia más probable

De cada grupo de ramas secundarias en el árbol, se eliminan todas excepto la más probable (moda). Con las ramas sobrevivientes, se resuelve el problema como se haría en ausencia de incertidumbre.

En este caso, el tomador de decisiones actúa como si la consecuencia con más probabilidades de ocurrir se presentara con certeza, sin importar el beneficio que esta implique. Es importante observar que, este criterio requiere una valoración numérica de la credibilidad que el tomador de decisiones le concede a cada uno de los eventos inciertos relevantes en el problema. Esta medida está dada por la probabilidad subjetiva respectiva y el procedimiento para asignarla se discutirá en el capítulo siguiente.

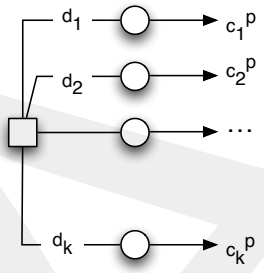


Figura 2.7: Árbol reducido bajo el criterio de la consecuencia más probable. Aquí c_i^p representa la consecuencia con mayor probabilidad de ocurrir entre las contenidas en $\{c_{i1}, c_{i2}, \ldots, c_{in_i}\}$.

2.3.4. Criterio de la utilidad promedio

De cada grupo de ramas secundarias en el árbol, se eliminan todas y se inserta una rama artificial cuya consecuencia equivale al promedio aritmético de las consecuencias. Con las ramas sobrevivientes, se resuelve el problema como se haría en ausencia de incertidumbre.

En el caso de este criterio, se considera que las consecuencias se expresan en términos numéricos y que además su beneficio está determinado por este valor o, más en general, que el beneficio que cada consecuencia reporte puede ser medido con un valor número.

Así, si $u : C \longrightarrow \Re$ es la función que produce ese valor, y c_1 y c_2 son dos consecuencias entonces $c_1 \succ c_2 \Leftrightarrow u(c_1) > u(c_2)$. A este tipo de funciones se les conoce como función de utilidad, y se discutirán en detalle en el capítulo siguiente.

Es interesante observar que con este criterio, el conjunto de ramas asociadas a las consecuencias de una opción es remplazado por una rama nueva, artificial,

que en general no existe en el conjunto original. Esta es una diferencia frente a los criterios expuestos previamente. En todo caso, el árbol de la figura 2.4 se sustituye por

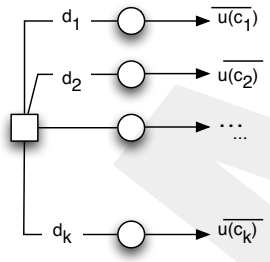


Figura 2.8: Árbol reducido bajo el criterio de la consecuencia más probable. Aquí $\overline{u(c_i)}$ representa la utilidad promedio de las consecuencias contenidas en $\{c_{i1}, c_{i2}, \ldots, c_{in_i}\}$.

2.3.5. Criterio de la utilidad esperada

De cada grupo de ramas secundarias en el árbol, se eliminan todas y se inserta una rama artificial con el promedio ponderado (por la probabilidad) de las consecuencias. Con las ramas sobrevivientes, se resuelve el problema como se haría en ausencia de incertidumbre.

Observe que en este criterio, al igual que en el criterio anterior, será necesario contar con el concepto de utilidad, así como con una medida de credibilidad de los eventos inciertos, tal como se describió en el criterio de la consecuencia más probable. En este caso, la nueva rama artificial tendrá una utilidad que equivale a la utilidad esperada del conjunto de consecuencias original.

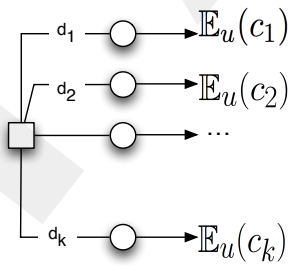


Figura 2.9: Árbol reducido bajo el criterio de la utilidad esperada. Aquí $\mathbb{E}_u(c_i)$ representa la utilidad esperada de las consecuencias contenidas en $\{c_{i1}, c_{i2}, \ldots, c_{in_i}\}$.

Observe que al utilizar este último criterio, el problema de decisión caracterizado por $(D, \mathcal{E}, C, \prec)$ se convierte, computacionalmente, en $(D, \mathcal{E}, C, \prec, u, P)$.

Ejemplo 2.3.1. *En unas elecciones parlamentarias en Inglaterra, competían los partidos Conservador y Laborista. Una casa de apuesta ofrecía las siguientes posibilidades:*

- *A quien apostara a favor de los Conservadores la casa pagaría, en caso de ganar la apuesta,* 7 *libras por cada* 4 *arriesgadas.*
- *A quien apostara a favor de los Laboristas la casa pagaría, en caso de ganar la apuesta,* 5 *libras por cada* 4 *arriesgadas.*

Así, si se definen:

$D = \{d_l, d_c\}$ *donde* $d_l =$*apostar k libras por los Laboristas y*
$d_c =$*apostar k libras por los Conservadores,*
$\mathcal{E} = \{E_1, E_2\}$ *donde* $E_1 =$*ganan los Conservadores y*
$E_2 =$*ganan los Laboristas, y*
$C = \{c_{l1}, c_{l2}, c_{c1}, c_{c2}\}$ *donde* $c_{jk} =$*consecuencia de apostar por j y que gane k,*

el árbol de decisión del problema resulta

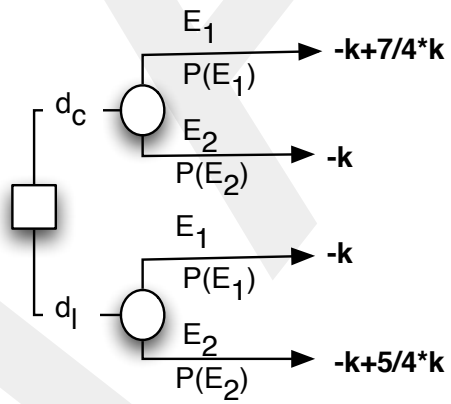


Figura 2.10: Árbol de decisión para el problema de las apuestas.

Por tanto, suponiendo que la utilidad del tomador de decisiones depende únicamente del pago de la apuesta y no de sus preferencias políticas, que no existe la posibilidad del empate, y que sólo compiten los partidos Laborista y Conservador. La solución y el árbol reducido bajo estos diferentes criterios están determinadas como se muestra en las siguientes figuras:

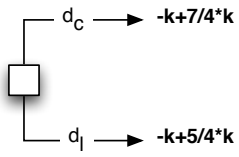


Figura 2.11: Árbol de decisión reducido, bajo el criterio optimista, para el problema de las apuestas. Solución óptima: d_c.

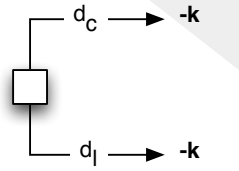


Figura 2.12: Árbol de decisión reducido, bajo el criterio pesimista, para el problema de las apuestas. Solución óptima: cualquiera d_c ó d_l.

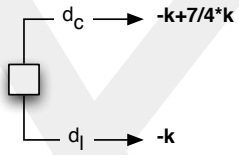


Figura 2.13: Árbol de decisión reducido, bajo el criterio de la consecuencia más probable, para el problema de las apuestas (caso $P(E_1) > 1/2$). Solución óptima: si $P(E_1) > 1/2 \quad d_c$; si $P(E_1) < 1/2 \quad d_l$; si $P(E_1) = 1/2$ este criterio no está definido.

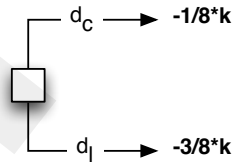


Figura 2.14: Árbol de decisión reducido, bajo el criterio de utilidad promedio, para el problema de las apuestas. Soulución óptima: d_c.

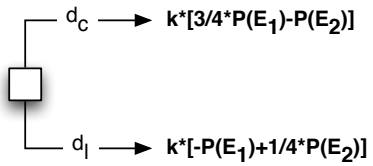
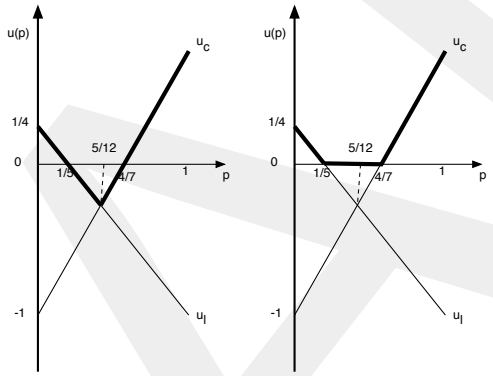


Figura 2.15: Árbol de decisión reducido, bajo el criterio de utilidad esperada, para el problema de las apuestas. Definiendo $p \equiv P(E_1) = 1 - P(E_2)$ la solución óptima resulta $d_c \Leftrightarrow \mathbb{E}_{u_c}(p) \equiv 7/4 * p - 1 \geq 1/4 - 5/4 * p \equiv \mathbb{E}_{u_l}(p) \Leftrightarrow p \geq 5/12$.



(a) (b)

Figura 2.16:
(a)Utilidad esperada en función de $P(E_1)$. Tomando $k = 1$.
(b)Estrategia óptima para el problema de las apuestas con D' y $k = 1$.

Es inmediato verificar los cuatro primeros resultados. Para el quinto criterio, que también se resuelve con facilidad, es interesante observar la figura 2.16.

En la figura 2.16a la línea gruesa identifica la opción óptima como función de la probabilidad subjetiva que el tomador de decisiones le asigna a la victoria de los conservadores. En segundo lugar, aparece un elemento muy revelador, la utilidad esperada de la decisión óptima es positiva en todos los casos excepto si $p \in (1/5, 4/7)$. *Esta información es sumamente valiosa si se observa que, en el caso de que existiera la opción de no apostar, la utilidad esperada de dicha opción sería cero.*

De hecho, este resultado sugiere la conveniencia de considerar el conjunto de opciones modificado $D' = \{d_c, d_l, d_0\}$ *donde* d_0 *corresponde a no apostar, en cuyo caso se obtendrá la utilidad esperada óptima (no negativa en todos los casos) que aparece marcada por la línea gruesa en la figura 2.16b .*

2.4. Ejercicios

Ejercicio 2.1. *El propietario de la tienda de ropa* Los Trapos, *ante el inminente arribo del otoño, debe decidir cuántos suéteres ha de adquirir de sus proveedores. Puede solicitar el material en múltiplos de* 100 *unidades. Si ordena* 100 *suéteres, estos tienen un costo unitario de* 150 *pesos, si ordena* 200 *el costo unitario es de* 120 *pesos y si ordena* 300 *suéteres o más, el costo unitario es de* 90 *pesos.*

En cualquier caso, Los Trapos *vende suéteres, en temporada otoño-invierno, a un precio de* 180 *pesos y si sobran algunos los remata, después de la temporada, a la tercera parte del precio. Además, el propietario del almacén considera que cada suéter que le sea solicitado en temporada y que no pueda vender (por falta de existencias) le supone una pérdida equivalente a 15 pesos, en términos económicos.*

Si por facilidad, el propietario supone que la demanda de suéteres en temporada será de 100, 150, 200, 250 *ó* 300 *unidades con probabilidades* $\frac{1}{6}$, $\frac{2}{9}$, $\frac{1}{3}$, $\frac{2}{9}$ *y* $\frac{1}{18}$ *respectivamente y que la utilidad que le reporte cada consecuencia se mide exclusivamente en términos monetarios, entonces*

a) Describa la estructura del problema de decisión.

b) Identifique las decisiones inadmisibles en este problema. ¿Le parece razonable la eliminación de estas decisiones? ¿Por qué?

c) Resuelva el problema vía minimax.

d) Resuelva el problema vía utilidad esperada máxima.

Ejercicio 2.2. *El día de su cumpleanños número* 20*, un paciente es admitido en el hospital con síntomas que sugieren la enfermedad A (con probabilidad* 0,4*) o bien la enfermedad B (con probabilidad* $0,6$*). Cualquiera que sea la enfermedad que en realidad padece, si no se trata morirá ese mismo día (con probabilidad* 0,8*) o bien sobrevivirá sin consecuencias para su salud (con probabilidad* 0,2*). El médico que recibe a este paciente tiene tres opciones que son alternativas:*

1. *No administrar tratamiento alguno.*
2. *Tratar al paciente con el fármaco F.*
3. *Realizar cirugía al paciente.*

Tanto la cirugía como la administración del farmaco entrañan riesgos. Sin importar la enfermedad, el paciente puede morir durante la operación (con probabilidad 0,5*). De la misma manera, sin importar la enfermedad, la droga puede ocasionar una reacción alérgica mortal (con probabilidad* 0,2*).*

Si el paciente sobrevive a los efectos adversos de la droga, y tenía la enfermedad A, entonces le puede curar (con una probabilidad de 0,9*) o puede que no tenga*

efecto alguno. Si el paciente tenía la enfermedad B, seguro la droga no tiene efecto curativo.

Por otra parte, si el paciente sobrevive a la operación, esta lo cura (con una probabilidad de 0,5*) si padecía la enfermedad A, en caso contrario no tiene efecto. Si padecía la enfermedad B, la operacación lo cura (con probabilidad* 0,6*), y no tiene efecto con probabilidad* 0,4*.*

En cualquier caso, si se recupera, el paciente tiene una esperanza de vida de 50 *años más. ¿Qué tratamiento es más conveniente?*

Ejercicio 2.3. *Considere el siguiente problema estadístico: Sea $x_{(n)}$ una m.a. de una variable aleatoria X tal que $\mathbb{E}(X^2) < \infty$. Encontrar el Mejor Estimador Lineal Insesgado (MELI) de $\mathbb{E}(X)$.*

a) Formule este problema como uno de decisión.

b) Identifique todos los componentes de la estructura.

c) Pruebe que la solución es la misma para cualquiera de los criterios discutidos en este capítulo.

d) ¿ Por qué ocurre este fenómeno?

Ejercicio 2.4. *Después de leer el siguiente artículo, que apareció en el periódico La Jornada el día sábado 7 de julio de 1990, diga usted si la interpretación que hace el periódico de las apuestas a las que se refiere en el último párrafo es correcta. Argumente formulando un problema de decisión como el analizado en el ejemplo 2.3.1*

■ ***De no gustarle su nueva actividad, podría dirigir en Italia***

Franz Beckenbauer, "aburrido del futbol", anunció ayer su retiro

■ ***Aprendimos la lección; no cometeremos los mismos errores del 86, afirman los alemanes***

Agencias, Roma, 6 de julio □ Seis años después de hacerse cargo de la dirección del equipo de Alemania Occidental, Franz Beckenbauer no sólo está más cerca que nunca de coronarse campeón mundial como técnico, sino también de su retiro.

El Kaiser anunció hoy que está aburrido y que iniciará una nueva actividad al margen del futbol, pero no dio detalles. Fuentes de la delegación aseguran que será consultor de mercado de una empresa automotriz alemana. Había versiones de que entrenaría al equipo de Estados Unidos en 1994, lo que al parecer también quedó desechado.

Beckenbauer dijo que si no le gusta su nueva actividad, regresará al mundo del futbol y que en este caso le gustaría trabajar en la liga italiana, "la mejor del mundo y la más profesional".

En cuanto a la final, señaló que espera que no se defina por penales y ratificó su optimismo al declarar que a diferencia de 1986, Maradona "quizá ya no es el número uno del mundo" y ha perdido mucho en estos cuatro años. Sin embargo, comentó que sigue siendo uno de los mejores jugadores del mundo y deben marcarlo de cerca.

En su acostumbrada conferencia de prensa, el entrenador declaró que en México cometieron el error de centrar su atención sobre Maradona y olvidaron a Burruchaga y Valdano, pero aclaró que "ya aprendimos la lección y esta vez abordaremos el problema de manera diferente", y ante la imposibilidad de controlar a Diego durante los 90 minutos, dijo que tratarán de limitar su "radio de acción" y que su probable marcador saldrá de Kohler, Berthold o Buchwald.

Expresó que las armas para el triunfo son la concentración y agresividad, factores que estuvieron en 86 con sus rivales.

No reveló la alineación de su equipo, pero son probables los cambios en el medio campo, para donde cuenta con cuatro elementos como Haessler, Littbarski, Bein y Thon. Pero al parecer ya se recuperó el delantero Rudi Voeller. Acerca de la alineación del rival, declaró tener mucha curiosidad por saber cómo se las ingeniará Bilardo para armar su equipo, ante las suspensiones y lesionados.

Es probable que los aficionados italianos apoyen a la RFA, además de que esta ciudad se apresta para la invasión de unos 30 mil aficionados alemanes.

Los teutones confían en Lothar Matthaeus, que el año anterior fue el frustrado marcador de Maradona, pero que ahora es el líder del equipo y, según Beckenbauer, puede consagrarse como "el mejor jugador del mundo". Matthaeus admitió que esa es su intensión y expresó "llegó la hora que ganemos la final". El volante, hasta el momento autor de cuatro goles, mandó un mensaje al Pibe, en el que advierte: "la lucha será a muerte".

Los jugadores alemanes afirman que serán apoyados por los romanos. Berthold se hace propaganda al declarar que "jugar mi segunda final del Mundial en Roma, en mi ciudad, la más bella del mundo, es un sueño". Y por otra parte comentó que no repetirán los errores de hace cuatro años, que no le asusta el lema de "no hay dos sin tres", sino que prefiere el de "la tercera es la vencida" y que esta vez están en condiciones de ganarle a cualquiera.

Pese a sus errores anteriores, los magos italianos insisten y ahora dicen que Alemania ganará el Mundial. En tanto, para los corredores de apuestas británicos, el equipo alemán es sólo favorito por un margen mínimo: por cada 10 libras que se arriesguen se pagarán 11, en caso de un triunfo alemán.

Capítulo 3

Elementos de la Teoría de Decisión

3.1. Axiomas de coherencia $(D, \mathcal{E}, C, \prec)$

Cuando una Teoría se desarrolla a partir de una base axiomática, la idea fundamental es plasmar en los axiomas los principios básicos que dan sustento a dicha Teoría. Así, todo resultado será consecuencia de los axiomas y por tanto, si algún resultado resultara inapropiado o controvertible el origen para ello debe encontrarse en los propios axiomas. Ahora bien, en general, una misma Teoría se puede desarrollar a partir de distintos conjuntos de axiomas. Habitualmente, la elección de uno de estos conjuntos en particular se realiza en función de la capacidad de síntesis que tenga, así como de su interpretabilidad.

Los axiomas de coherencia que aquí se presentan, son una versión simplificada de los que aparecen en Bernardo y Smith (1994), y que fueron propuestos, por ejemplo, en Bernardo (1981). Estos axiomas, constituyen una base para la Teoría de la Decisión, y describen los principios que debería cumplir un tomador de decisiones que no quiere incurrir en errores sistemáticos (quiere ser coherente). Estos cuatro axiomas son especialmente fáciles de interpretar y, como puede observarse, tres de ellos se refieren a las características del orden de preferencias, mientras que el cuarto establece un sistema de medición de la incertidumbre en términos de preferencias.

Finalmente, si como es el caso en la Teoría de la Decisión que se discute aquí, de los axiomas se sigue un procedimiento único para la solución de los problemas que aborda la Teoría, entonces cualquier otro procedimiento tiene sólo dos posibilidades: o bien es equivalente al que establecen los axiomas, o se contrapone con éstos.

Axioma 1. *Comparabilidad: para todo par de elementos $d_i, d_j \in D$ es cierta una y sólo una de las tres siguientes afirmaciones:*

- $d_i \succ d_j$ *(d_i es mas preferible a d_j)*
- $d_i \sim d_j$ *(d_i es igualmente preferible a d_j)*
- $d_j \succ d_i$ *(d_j es mas preferible a d_i)*

Además, es posible encontrar c_* y c^* en C t.q. $c_* \preceq c \preceq c^* \quad \forall c \in C$.

En términos llanos, este axioma establece que si el tomador de decisiones quiere elegir una opción en D, entonces debe partir del supuesto de que es posible comparar los elementos en de D. La negación de este axioma equivale a renunciar, de entrada, a la solución del problema.

Axioma 2. *Transitividad: si $d_i, d_j, d_k \in D$ y sucede que $d_i \succ d_j$, y además que $d_j \succ d_k$ entonces se cumplirá necesariamente que $d_i \succ d_k$.*

Suponga por un momento que este axioma no se cumple. Esto es, suponga que existen d_i, d_j, d_k en D tales que el tomador de decisiones considera que $d_i \succ d_j$ y que $d_j \succ d_k$, pero que $d_k \succeq d_i$. Si tal fuera el caso, usted podría ofrecerle gratis cualquiera de las tres opciones. Si por ejemplo, el tomador de decisiones eligiese d_k, entonces usted podría ofrecerle la sustitución por d_j (que es más preferible que d_k para él) a cambio de una suma S_1 positiva pero muy pequeña, de manera que la permuta a él le parezca conveniente. Acto seguido, Ud. puede hacer lo mismo para sustituirle d_j por d_i a cambio de una suma S_2. Por último, como el tomador de decisiones considera que $d_k \succeq d_i$ puede ofrecerle el cambio de d_i por d_k gratis. Así, el tomador de decisiones vuelve a la posición original después de haber pagado $S_1 + S_2$. No importa que tan pequeños sean S_1 y S_2, si son positivas usted puede repetir este procedimiento indefinidamente, y habrá convertido al tomador de decisiones en una máquina perpetua de regalar dinero.

Axioma 3. *Sustitubilidad: si $d_i, d_j \in D$ y A es un evento incierto t.q. $d_i \succ d_j$ cuando ocurre A y $d_i \succ d_j$ cuando ocurre A^c entonces $d_i \succ d_j$. Análogamente si $d_i \sim d_j$ cuando ocurre A y $d_i \sim d_j$ cuando ocurre A^c entonces $d_i \sim d_j$.*

Este es un axioma de congruencia. Si por ejemplo, el tomador de decisiones prefiriera invertir en valores de renta fija y no en la bolsa de valores cuando hay recesión en el país y tambien prefiriera los valores de renta fija sobre la bolsa cuando no hay recesión, entonces, simplemente prefiere la renta fija sobre la bolsa. Observe que el axioma no establece que $d_i \succ d_j$ en ambos casos (cuando ocurre A y cuando ocurre A^c). Lo que afirma es que si $d_i \succ d_j$ en los dos escenarios (que forman una partición del evento seguro) entonces los escenarios son irrelevantes.

Axioma 4. *Eventos de referencia: Independientemente de los eventos inciertos relevantes, el tomador de decisiones puede imaginar un procedimiento para generar puntos en el cuadrado unitario I de manera que para cualesquiera dos regiones R_1 Y R_2 en I, el evento $A_1 = \{z \in R_1\}$ es más creíble que el evento $A_2 = \{z \in R_2\} \Leftrightarrow \acute{A}rea(R_1) > \acute{A}rea(R_2)$.*

Este cuarto axioma es de una naturaleza distinta a los tres anteriores. Simplemente define un patrón de referencia, y establece un mecanismo para la medición cuantitativa de la incertidumbre. En términos de irrefutabilidad, lo único que en realidad establece es que el tomador de decisiones sea capaz de imaginar un mecanismo para simular observaciones de una distribución Uniforme en el cuadrado unitario de $\Re^2$.

Ahora, dado un problema de decisión con incertidumbre $(D, \mathcal{E}, C, \succ)$ con $d \in D$ y $d = \{c_1|E_1, c_2|E_2, \ldots, c_k|E_k\}$, considere el siguiente procedimiento:

1. Modificar el conjunto D incluyendo nuevas decisiones artificiales de la forma $d_c = \{c \mid \Omega\}$ $\forall c \in C$, además de todas las decisiones originales. Llame a este conjunto modificado D_1.

2. Modificar el conjunto D_1 incluyendo nuevas decisiones artificiales de la forma $d_E = \{c_*|E^c, c^*|E\}$ $\forall E \subseteq \mathcal{E}$. Llame a este conjunto D_2.

3. Modificar el conjunto D_2 incluyendo nuevas decisiones artificiales del tipo $d_R = \{c_*|R^c, c^*|R\}$ $\forall R \subseteq I$. Llame a este conjunto D_3. Observe que D_3 será no numerable y $d_c \in D_3$ $\forall c \in C$, en particular $d_{c_*}, d_{c*} \in D_3$.

Así para cualesquiera dor regiones R_1, R_2 en el cuadrado unitario, y las decisiones $d_{R_1} = \{c_*|{R_1}^c, c^*|R_1\}$ y $d_{R_2} = \{c_*|{R_2}^c, c^*|R_2\}$, se cumplirá necesariamente que $d_{R_1} \prec d_{R_2} \Longleftrightarrow R_2$ es más creíble que R_1 ($\acute{A}rea(R_2) > \acute{A}rea(R_1)$).

Adicionalmente, para el caso particular de $d_\emptyset = d_{c_*}$ y $d_\Omega = d_{c^*}$, por el axioma 1 se tiene que $c_* \preceq c \preceq c^*$ de donde se sigue que $d_\emptyset = d_{c_*} \preceq d_c \preceq d_{c*} = d_\Omega$ $\forall c \in C$.

Como complemento de los primeros cuatro axiomas que son los que en realidad definen la naturaleza de Teoría de la Decisión que se presenta en este capítulo, en este punto es conveniente introducir un axioma adicional cuya utilidad es fundamentalmente técnica. A partir de los axiomas 1 a 4, la medición cuantitativa tanto las preferencias como de la incertidumbre puede llevarse hasta el extremo de confinar su valor numérico en un intervalo arbitrariamente pequeño, y en la práctica esta aproximación puede ser suficiente. Sin embargo, para poder asignarle un valor preciso y único, condición que es conveniente para efectos conceptuales, es necesario introducir el siguiente axioma.

Axioma 5. *Densidad: La colección de decisiones $D_I = \{d_R \,|R \subseteq I\}$ es densa en D_3, i.e. $\forall d \in D_3$ $\exists R \subseteq I$ t.q. $d \sim d_R$.*

3.2. Definición de utilidad

Definición 3.2.1. *Sea $c \in C$, se define la* **utilidad canónica** *$u_0(c)$ como el de una región $R \subset I$ t.q. $d_c \sim d_R$.*

Observe que, puesto que la utilidad canónica se define en términos del área de una región en I, entonces necesariamente $u_0(c) \in [0,1]\ \forall c \in C$.

Teorema 3.2.1. *$\forall c \in C$, $u_0(c)$ existe y es único.*

Demostración. La existencia de u_0 es consecuencia directa del axioma de densidad. Ahora,

$$\begin{aligned} &sean\ \ u_0(c) = \acute{A}rea(R_1)\ \ t.q.\ d_c \sim d_{R_1} = \{c_*|{R_1}^c, c^*|R_1\} \\ &y\ \ {u_0}'(c) = \acute{A}rea(R_2)\ \ t.q.\ d_c \sim d_{R_2} = \{c_*|{R_2}^c, c^*|R_2\}. \end{aligned}$$

Por el axioma 1: $d_{R_1} \sim d_{R_2}$, *i.e.* R_1 es igualmente creíble que R_2

$$\underset{\text{axioma 4}}{\Longrightarrow}\ \acute{A}rea(R_1) = \acute{A}rea(R_2)\ \therefore\ u_0(c) = {u_0}'(c). \qquad \square$$

Teorema 3.2.2. *u_0 es creciente con respecto a la relación de preferencia $\prec$.*

Demostración. Sean c_1 y c_2 t.q. $c_1 \prec c_2$. Se sabe entonces que existen R_1 y R_2 tales que

$$u_0(c_1) = \acute{A}rea(R_1)\ \text{ y }\ u_0(c_2) = \acute{A}rea(R_2)$$

por tanto, existen también d_{c_1} y d_{c_2} tales que

$$d_{R_1} \sim d_{c_1} \prec d_{c_2} \sim d_{R_2}\ \Longrightarrow\ d_{R_1} \prec d_{R_2}.$$

Ahora, si se supone que $\acute{A}rea(R_1) \geq \acute{A}rea(R_2)$ resulta que

$$d_{R_1} \succeq d_{R_2}\ \text{¡}\ \Longrightarrow\ \ \acute{A}rea(R_1) < \acute{A}rea(R_2)\ \therefore\ u_0(c_1) < u_0(c_2). \qquad \square$$

Corolario. *$u_0(c_*) = 0$ y $u_0(c^*) = 1$.*

Demostración. Como $c_* \in C$, entonces $u_0(c_*) = \acute{A}rea(R)$ donde $d_R \sim d_{c_*}$. De donde se sabe que $\acute{A}rea(R) = 0$, lo que implica que

$$u_0(c_*) = 0.$$

Análogamente para $c^* \in C$ se tiene que $u_0(c^*) = \acute{A}rea(S)$ donde $d_S \sim d_{c*}$ y $\acute{A}rea(S) = 1$, por tanto

$$u_0(c^*) = 1. \qquad \square$$

Observe que si $E_1, E_2 \in \mathcal{E}$ y $d_{E_1} = \{c_*|E_1^c, c^*|E_1\},\ d_{E_2} = \{c_*|E_2^c, c^*|E_2\}$, entonces $d_{E_1} \prec d_{E_2} \Longleftrightarrow E_2$ es más creíble que E_1. Esto es, $d_{E_1} \prec d_{E_2}$ si y sólo si $E_1 \prec^* E_2$. Donde $\prec^*$ se utiliza para definir una nueva relación en $\mathcal{E} \times \mathcal{E}$ que establece el orden de credibilidad entre los eventos inciertos. Adicionalmente, si $E \in \mathcal{E}$ es tal que $d_E = \{c_*|E^c, c^*|E\}$ se sabe, por el axioma 5, que existe un R en I tal que $d_E \sim d_R$, lo que implica que E y R son igualmente creíbles ($E \sim^* R$).

3.3. Definición de probabilidad

Definición 3.3.1. *Sea $E \subset \mathcal{E}$ un evento incierto relevante, se define* **la probabilidad subjetiva** *de E en las condiciones H como $P(E|H) = \acute{A}rea(R)$ donde R cumple que $d_R \sim d_E$ bajo las condiciones H.*

Teorema 3.3.1. *Para todo evento incierto relevante $E \subseteq \mathcal{E}$ y condiciones H, $P(E \,|\, H)$ existe y es única.*

Demostración. La existencia de $P(E|H)$ es consecuencia directa del axioma de densidad. Ahora, sean

$$\begin{aligned} P(E|H) =& \acute{A}rea(R) \quad \text{t.q. } d_E \sim d_R = \{c_*|R^c, c^*|R\} \text{ y} \\ P'(E|H) =& \acute{A}rea(S) \quad \text{t.q. } d_E \sim d_S = \{c_*|S^c, c^*|S\}. \end{aligned}$$

Por lo que utilizando el axioma 1, se tiene que $d_R \sim d_S$. Esto es, R es igualmente creíble que S.

$$\underset{\text{axioma 4}}{\Longrightarrow} \acute{A}rea(R) = \acute{A}rea(S) \quad \therefore P(E|H) = P'(E|H) \qquad \square$$

Teorema 3.3.2. *(Propiedades de la probabilidad subjetiva).*

Sean E y F dos eventos inciertos relevantes en $\mathcal{E}$ y las condiciones H, la función de probabilidad subjetiva cumple las siguientes cuatro propiedades:

1. $0 \leq P(E|H) \leq 1$
2. $P(\emptyset|H) = 0$
3. $P(\Omega|H) = 1$
4. *Si* $E \cap F = \emptyset \implies P(E \cup F|H) = P(E|H) + P(F|H)$

Demostración.

1. $P(E|H) = \acute{A}rea(R)$ donde $R \subset I \;\; \therefore \; 0 \leq P(E|H) \leq 1$
2. Por la propiedad 1, $0 \leq P(E|H) \leq 1$. Y por definición $P(\emptyset|H) = \acute{A}rea(R)$ donde $d_\emptyset \sim d_R = \{c_*|R^c, c^*|R\}$. Lo que implica que $\acute{A}rea(R) = 0$ y por tanto, $P(\emptyset|H) = 0$.
3. La demostración es análoga a la de la propiedad 2.
4. Para el caso en que $E = \emptyset$ o $F = \emptyset$, la conclusión se sigue directamente de la propiedad 2. En caso contrario, si $E, F \neq \emptyset$, sean

$$\begin{aligned} P(E|H) =& \acute{A}rea(R) \quad \text{t.q. } d_E \sim d_R = \{c_*|R^c, c^*|R\} \text{ y} \\ P(E \cup F|H) =& \acute{A}rea(S) \quad \text{t.q. } d_{E\cup F} \sim d_S = \{c_*|S^c, c^*|S\}. \end{aligned}$$

Puesto que E y F son disjuntos y $F \neq \emptyset$, debe suceder que $E \subset E \cup F$, y por tanto $E \cup F$ es más creíble que E ($E \prec^* E \cup F$), lo que implica que $\acute{A}rea(S) > \acute{A}rea(R)$. Así, es posible tomar $R' \subset S \subseteq I$ tal que $\acute{A}rea(R') = \acute{A}rea(R)$, de manera que $d_{R'} \sim d_E$ ($E \sim^* R'$).

Por otro lado, considere $(S \backslash R') = \{x \in I \,|\, x \in S, x \notin R'\}$. Observe que, puesto que $R' \subset S$, resulta que $(S \backslash R') \cup R' = S$ y $(S \backslash R') \cap R' = \emptyset$, y por ende $\acute{A}rea(S \backslash R') + \acute{A}rea(R') = \acute{A}rea(S)$. Además, es posible expresar la siguientes relaciones:

$$\begin{aligned}
d_F =& \{c_*|F^c, c^*|F\},\\
d_{(S\backslash R')} =& \{c_*|(S\backslash R')^c, c^*|(S\backslash R')\},\\
d_{E\cup F} =& \{c_*|(E\cup F)^c, c^*|(E\cup F)\} = \{c^*|E, \{c_*|F^c, c^*|F\}|E^c\} \text{ y}\\
d_{(S\backslash R')\cup R'} =& \{c_*|[(S\backslash R')\cup R']^c, c^*|(S\backslash R')\cup R'\}\\
=& \{c^*|R', \{c_*|(S\backslash R')^c, c^*|(S\backslash R')\}|R'^c\}.
\end{aligned}$$

Ahora, dado que E es igualmente creíble que R', suponer que $d_F \prec d_{S\backslash R'}$ implica que $d_{(S\backslash R')\cup R'} \prec d_{E\cup F}$. Por lo que $d_S \sim d_{(S\backslash R')\cup R'} \prec d_{E\cup F.}$ Pero por construcción $d_{E\cup F} \sim d_S$, lo que constituye una contradicción. Análogamente, tampoco es posible que $d_F \succ d_{S\backslash R'}$. Entonces, utilizando el axioma de comparabilidad se cumplirá necesariamente que $d_F \sim d_{S\backslash R'}$, y así $P(F|H) = \acute{A}rea(S\backslash R')$, de donde se sigue la conclusión. □

El teorema 3.3.2 es extraordinariamente importante. Lo que implica es que en el marco de la Teoría de Decisión, los célebres axiomas Kolmogorov para la Probabilidad ya no son axiomas puesto que se derivan de principios más básicos (los axiomas de coherencia). Este resultado da cuenta de la potencia que tiene la Teoría de la Decisión.

3.4. Principio de la utilidad esperada máxima

Como se discutió en el capítulo previo, los métodos para resolver problemas de decisión en ambiente de incertidumbre suelen recurrir a la idea de "podar" el árbol de decisión y tratar el problema como si fuera uno sin incertidumbre. En los ejemplos con los que estos procedimientos han sido ilustrados ha quedado claro que estos criterios no necesariamente conducen a una misma solución, y más aún, que distintos métodos requieren insumos de información diferentes por parte del tomador de decisiones.

Del análisis comparativo entre los criterios considerados, resulta que el el de utilidad esperada máxima es el más costoso en términos de información. Esto podría sugerir que en ese sentido es un "mejor" método. En esta sección se prueba, a partir de los axiomas de coherencia, que el criterio de utilidad esperada máxima no es solamente una "idea razonable"; sino que es el único criterio

compatible con estos axiomas. De hecho, cualquier otro mecanismo de solución, o coincide con este o viola alguno de los axiomas.

Para tal fin, en el contexto de un problema de decisión con un número finito de posibles consecuencias $(D, \mathcal{E}, C, \prec)$, sea d un elemento en D. Esta decisión puede ser representada como

$$d = \{c_1|E_1, c_2|E_2, \ldots, c_k|E_k\},$$

y considere la primera consecuencia involucrada, $c_1 \in C$. Como ya se ha mostrado, debe existir $R_1 \subset I$ tal que

$$c_1 \sim d_{R_1} = \{c_*|{R_1}^c, c^*|R_1\},$$

pero entonces, si se considera la opción

$$d^{(1)} = \{d_{R_1}|E_1, c_2|E_2, \ldots, c_k|E_k\}$$

se tiene que si ocurre E_1^c, d y $d^{(1)}$ producen exactamente la misma consecuencia y por tanto, son igualmente preferibles. Si por el contrario ocurre E_1, d produce c_1 mientras que $d^{(1)}$ produce d_{R_1}, pero $c_1 \sim d_{R_1}$ y por tanto $d \sim d^{(1)}$. Utilizando entonces el axioma de sustitubilidad, debe ocurrir que simplemente

$$d \sim d^{(1)} = \{c_*|E_1 \cap {R_1}^c, c^*|E_1 \cap R_1, c_2|E_2, \ldots, c_k|E_k\}.$$

Procediendo análogamente para cada i en $\{1, 2, \ldots, k\}$ debe ocurrir que si

$$d^{(i)} \sim \{d_{R_1}|E_1, d_{R_2}|E_2, \ldots, d_{R_i}|E_i, c_{i+1}|E_{i+1}, c_{i+2}|E_{i+2} \ldots, c_k|E_k\}$$

entonces $d^{(i)} \sim d^{(i+1)}$ para $i = \{1, 2, \ldots, k-1\}$. Recurriendo al axioma de transitividad, se tiene necesariamente que $d \sim d^{(k)}$, es decir

$$\begin{aligned} d \sim & \{c_*|E_1 \cap {R_1}^c, c^*|E_1 \cap R_1, \ldots, c_*|E_k \cap {R_k}^c, c^*|E_k \cap R_k\} \\ & = \left\{ c_*|(\bigcup_{i=1}^{k}(E_i \cap R_i))^c, c^*|\bigcup_{i=1}^{k}(E_i \cap R_i) \right\}. \end{aligned}$$

De esta forma, ha quedado establecido que para cualquier $d \in D$ existe otra opción $(d^{(k)})$ tal que

I) $d \sim d^{(k)}$

II) $d^{(k)}$ es una opción con sólo dos consecuencias, (c_* y c^*).

En estas condiciones, sean las decisiones $d_1, d_2 \in D$, y los conjuntos definidos por $A = \bigcup_{i=1}^{k_1}(E_{1i} \cap R_{1i})$ y $B = \bigcup_{i=1}^{k_2}(E_{2i} \cap R_{2i})$, entonces

$$d_1 \sim d_A \text{ y } d_2 \sim d_B$$

donde $d_A = \{c_*|A^c, c^*|A\}$ y $d_B = \{c_*|B^c, c^*|B\}$. Ahora bien, $d_1 \prec d_2$ si y sólo si $d_A \prec d_B$. Sin embargo,

$$d_A \prec d_B \Longleftrightarrow \{c_*|A^c, c^*|A\} \prec \{c_*|B^c, c^*|B\}$$

y puesto que esta última desigualdad equivale a que A sea menos creíble que B ($A \prec^* B$), necesariamente se tiene que

$$d_1 \prec d_2 \Longleftrightarrow P(A|H) < P(B|H).$$

Por último, observe que

$$P(A|H) = P\left(\bigcup_{i=1}^{k_1}(E_{1i} \cap R_{1i})\middle| H\right) = \sum_{i=1}^{k_1} P(E_{1i} \cap R_{1i}|H)$$

pero puesto que los eventos de referencia son independientes de los eventos inciertos relevantes,

$$P(A|H) = \sum_{i=1}^{k_1} P(E_{1i}|H)P(R_{1i}|H) = \sum_{i=1}^{k_1} u_0(c_{1i})P(E_{1i}|H).$$

En otras palabras, $P(A|H)$ coincide con la utilidad canónica esperada asociada a d_1. Procediendo análogamente, es posible verificar que

$$P(B|H) = \sum_{i=1}^{k_2} u_0(c_{2i})P(E_{2i}|H).$$

En consecuencia,

$$\begin{aligned} d_1 \prec d_2 &\Longleftrightarrow \sum_{i=1}^{k_1} u_0(c_{1i})P(E_{1i}|H) < \sum_{i=1}^{k_2} u_0(c_{2i})P(E_{2i}|H) \\ &\Longleftrightarrow \mathbb{E}_H\{u_0(d_1)\} < \mathbb{E}_H\{u_0(d_2)\}. \end{aligned}$$

Es decir, la opción más preferible es la que produce la utilidad esperada máxima y, por tanto, el único criterio congruente con la axiomática de Teoría de la Decisión es el de utilidad esperada. Así, el resultado que se deriva de los axiomas de coherencia se puede precisar en tres etapas:

- Toda forma de incertidumbre debe y puede ser descrita con una medida de probabilidad.
- Para toda consecuencia en el problema se debe y puede asignar un valor numérico de utilidad.
- Una decisión es más preferible que otra si y sólo si su utilidad esperada es mayor a la utilidad esperada de la otra.

De esta forma, la teoría conduce a que cualquier problema de decisión en ambiente de incertidumbre se pueda resolver con un algoritmo único y general:

1. Se asignan las probabilidades de todos los eventos inciertos.
2. Se asigna la utilidad de todas las posibles consecuencias en el problema.
3. Se calcula la utilidad esperada para cada d en D.

Y la solución es la opción $d^\star \in D$ tal que $\mathbb{E}_H\{u_0(d^\star)\} \geq \mathbb{E}_H\{u_0(d)\}\ \forall d \in D$.

Por supuesto, aun resta la discusión sobre la manera de asignar las probabilidades y utilidades en un problema concreto. Este es un tema que se atenderá en el capítulo siguiente.

3.5. Incorporación de información adicional

Un aspecto muy importante que no se ha discutido hasta ahora, es el hecho de que la Teoría es de naturaleza estática. Es decir, establece la manera de resolver los problemas de decisión en un momento específico del tiempo. El asunto no es menor, puesto que si un mismo problema $(D, \mathcal{E}, C, \prec)$ se enfrenta después de que ha ocurrido algún tiempo, puede ocurrir que la solución originalmente óptima deje de serlo. Esto ocurre, por ejemplo, porque algunas opciones que se consideraban factibles ya no lo son, porque otras opciones que ni siquiera existían aparecen, o porque las preferencia o el nivel de incertidumbre del tomador de decisiones se ha modificado.

Ahora bien, debe observarse que la misma "receta" de utilidad esperada máxima sigue siendo la única posibilidad compatible con los axiomas para resolver el problema. Por tanto, si en efecto se producen cambios, esto se debe a que han cambiado los "ingredientes".

En general, estos cambios se pueden interpretar como resultado del arribo de información adicional. Y en general, dicha información adicional puede tener dos tipos de impacto en el problema:

- De impacto estructural, *i.e.* se modifica $D, \mathcal{E}$ ó C
- De impacto en creencias o preferencias. *i.e.* se modifica P ó u

Ahora bien, los cambios estructurales habitualmente ocurren independientemente de la voluntad del tomador de decisiones. Por lo que toca a los cambios en preferencias, estos suelen presentarse de manera esporádica y, generalmente, sin la intervención consciente del tomador de decisiones.

El tipo de cambio que se distingue de los demás, porque suele ser mucho más cumún y provocado en forma deliberada por el tomador de decisiones es el de las creencias. Ya se ha discutido, con amplitud, que un problema de decisión

es mucho más complejo cuando involucra incertidumbre. En consecuencia, es razonable que los tomadores de decisiones procuren eliminar, o al menos disminuir, la incertidumbre con la que se enfrentan a un problema de decisión. Entonces, el camino obvio es obtener información adicional sobre los eventos inciertos relevantes.

Así pues, en ocasiones la información adicional (Z) se puede registrar como fruto de la observación de una colección de variables aleatorias. Esto es, existe una distribución o función de probabilidad $P(Z) \equiv P(Z|E,H)$que describe a $Z \in \mathcal{Z}$. En este caso, es posible utilizar la regla de Bayes para actualizar las creencias pues

$$P(E|Z,H) = \frac{P(E \cap Z|H)}{P(Z|H)} \quad y \quad P(Z|E,H) = \frac{P(E \cap Z|H)}{P(E|H)},$$

lo que implica que

$$P(E|Z,H) = \frac{P(Z|E,H)P(E|H)}{P(Z|H)}.$$

Adicionalmente, por la ley de probabilidades totales

$$P(Z|H) = \sum_{i=1}^{r} P(Z|H,E_i)P(E_i)$$

Donde $E_1, E_2 \ldots, E_r$ es una partición de Ω. Y por tanto

$$P(E_i|Z,H) = \frac{P(Z|E_i,H)P(E_i|H)}{\sum_{j=1}^{r} P(Z|H,E_j)P(E_j)}.$$

Observe que $P(E_i|Z,H)$ tiene como argumento al evento incierto E_i y que, puesto que $E_1, E_2 \ldots, E_r$ forman una partición del evento seguro, su suma debe ser igual a 1. Entonces, $P(Z|H)$ puede ser tratado como una constante de normalización, y es posible escribir

$$P(E_i|Z,H) \boldsymbol{\alpha} P(Z|E_i,H)P(E_i|H)$$

donde el símbolo $\boldsymbol{\alpha}$ se lee como "es proporcional a".

La interpretación de esta última expresión es reveladora. $P(E_i|H)$ es la probabilidad que describe el estado de incertidumbre antes de la información Z, que se conoce como la *inicial* o *a priori*. $P(E_i|Z,H)$ es la probabilidad que describe el estado de incertidumbre después de conocer la información Z, y se le llama final o a posteriori. Así, resulta que la final es proporcional al producto de la inicial y el factor $P(Z|E_i,H)$ que, a su vez, se conoce como la *verosimilitud* de E_i dado Z. Este nombre no es casual; como se verá más adelante, en efecto $P(Z|E_i,H)$ coincide con la muy conocida función de verosimilitud que aparece en los textos de Inferencia Estadística.

Es importante observar también que el procedimiento de actualización o aprendizaje que transforma una inicial en una final, es de hecho, un proceso secuencial que se puede repetir cuando después de una primera pieza de información adicional Z se recibe otra más Z'. Así, la inicial $P(E|H)$ se transforma en la final $P(E|H,Z)$, que en el segundo ciclo juega el papel de inicial para actualizarse con Z', y poder llegar a la final $P(E|H,Z,Z')$. Esto bajo las reglas de actualización

$$P(E|H,Z)\boldsymbol{\alpha}P(Z|E,H)P(E|H) \quad \text{y}$$
$$P(E|H,Z,Z')\boldsymbol{\alpha}P(Z'|E,H,Z)P(E|H,Z).$$

donde lo más notable es que, en general, la verosimilitud en el segundo caso es condicional en la primera pieza de información. Por supuesto, si Z y Z' son condicionalmente independientes dado E, entonces

$$P(E|H,Z)\boldsymbol{\alpha}P(Z|E,H)P(E|H) \quad \text{y}$$
$$P(E|H,Z,Z')\boldsymbol{\alpha}P(Z'|E,H)P(E|H,Z).$$

Ahora, resulta importante señalar que, dado que $P(E|H,Z,Z') = P(E|H,Z',Z)$ el orden en el que llegue la información adicional es irrelevante.

Finalmente, debe resultar evidente que, al igual que a priori una solución de Bayes $\delta^\star$ es tal que

$$\mathbb{E}_{P(E)}\{u_0(d,E)\} \leq \mathbb{E}_{P(E)}\{u_0(\delta^\star,E)\} \quad \forall\, d \in D,$$

análogamente, a posteriori una solución de Bayes será $\delta^\star(Z)$ tal que

$$\mathbb{E}_{P(E|Z)}\{u_0(d,E)\} \leq \mathbb{E}_{P(E|Z)}\{u_0(\delta^\star(Z),E)\} \quad \forall\, d \in D.$$

Es interesante insistir en la interpretación de la última expresión; una vez que se cuenta con los datos observados y fijos Z, la solución óptima (de Bayes) es $\delta^\star(Z)$. Naturalmente, si los datos hubieran sido distintos la decisión óptima podría haber sido diferente. Esta idea da lugar a la noción de regla de decisión, que se discute brevemente en la siguiente sección.

3.6. Reglas de decisión

Considere el espacio de opciones originales D, y denomine $\mathcal{D}$ el espacio de todas las funciones que van de $\mathcal{Z}$ a D. Ante la eventualidad de contar con una posible pieza de información $Z \in \mathcal{Z}$ el tomador de decisiones puede preguntarse: ¿Cuál es la función o regla que debiera aplicar a los nuevos datos con fin de seleccionar la opción que resuelva el problema original? De hecho, este es un problema en principio más general cuyo árbol de decisión aparece en la figura 3.1. A una función de este tipo se le conoce como regla de decisión y su definición formal es la siguiente.

Definición 3.6.1. *Sea $\mathcal{Z}$ el espacio de resultados de un experimento, cualquier función $\delta : \mathcal{Z} \to D$ se conoce como* **Regla de Decisión**

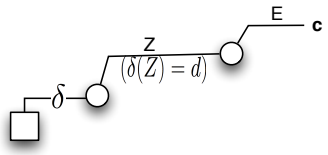


Figura 3.1: Rama típica del árbol de decisión bajo reglas de decisión.

Como se puede observar en la figura 3.1, una vez que se elige una regla de decisión (δ) un nodo de incertidumbre determina la información con la que se contará y, como resultado, la decisión que será seleccionada. En ese momento, otro nodo de incertidumbre produce el evento incierto relevante y este, a su vez, conducirá a la consecuencia. Conceptualmente, se puede pensar en que el problema original $(D, \mathcal{E}, C, \prec)$ se transforma en el nuevo problema $(\mathcal{D}, \mathcal{E} \times \mathcal{Z}, C, \prec)$ este último, con representado por el árbol de decision de la figura 3.2.

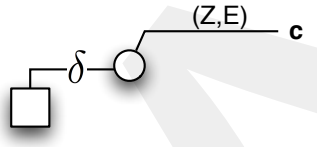


Figura 3.2: Rama típica del árbol de decisión bajo reglas de decisión para el problema modificado.

De esta forma, es claro que $\delta^\star$ es óptima si y sólo si maximiza, sobre $\mathcal{D}$, la utilidad esperada

$$\mathbb{E}\{u(\delta, Z, E)\} = \int\int u\left(\delta\left(Z\right), E\right) P(E, Z)\, dEdZ.$$

Al respecto, es de interés observar que

$$\begin{aligned}\int\int u(\delta(Z), E)P(E, Z)\, dEdZ &= \int\int u(\delta(Z), E)P(E|Z)P(Z)\, dEdZ \\ &= \int P(Z)\left[\int u(\delta(Z), E)P(E|Z)\, dE\right] dZ,\end{aligned}$$

de manera que, si para cada Z fija en $\mathcal{Z}$ se elige la opción $d_Z^\star$ en D tal que

$$\int u(d, E)P(E|Z)dE \leq \int u(d_Z^\star, E)P(E|Z)dE\ \forall d \in D,$$

entonces, si se define $\delta^\star(Z) = d_Z^\star\ \forall Z \in \mathcal{Z}$ se tiene que

$$\int P(Z)\left[\int u(d(Z), E)P(E|Z)\, dE\right] dZ \leq \int P(Z)\left[\int u(d_Z^\star, E)P(E|Z)\, dE\right] dZ.$$

Es decir $\mathbb{E}\{u(\delta, Z, E)\} \leq \mathbb{E}\{u(\delta^\star, Z, E)\}$, y por tanto $\delta^\star$ es la solución de Bayes para el problema de elegir una regla de decisión óptima. Este resultado es particularmente interesante porque significa que el problema de encontrar la mejor

regla de decisón $\delta^\star$ se puede resolver encontrando, para cada Z, la mejor decisión $d^\star_Z$.

Existe otro concepto que es conveniente explorar en este punto. Considere un problema de decisión $(D, \mathcal{E}, C, \prec)$, e imagine que la correspondiente solución de Bayes no es única. Suponga, por ejemplo, que existen d_1 y d_2 en D tales que $\mathbb{E}\{u(d_1, E)\} = \mathbb{E}\{u(d_2, E)\}$ y que $\mathbb{E}\{u(d_i, E)\} \succeq \mathbb{E}\{u(d, E)\}\ \forall d \in D$ y $i = 1, 2$. Es decir, suponga que d_1 y d_2 son dos soluciones de Bayes para el mismo problema. En estas condiciones, el tomador de decisiones podría seleccionar cualquier decisión, d_1 o d_2, y la solución sería óptima. Una pregunta interesante sería la siguiente: ¿Qué pasa si cada vez que deba tomar una decisión en este contexto, el tomador de decisiones lanza una moneda y, dependiendo del resultado, elige d_1 o d_2?

Puesto que tanto d_1 como d_2 son óptimas, la introducción del volado no debiera impactar la calidad del resultado. De hecho, esta idea se emplea en el ámbito de teoría de juegos, en donde además de maximizar la utilidad se lograría que el oponente no pueda anticipar con precisión la jugada. En el caso de una decisión no hay un oponente, así que ese efecto no es un fin en sí mismo pero, en cualquier caso, la idea puede explorarse con otros propositos.

Así, en el caso en que D es finito, puede definirse el concepto de una regla de decisión de la siguiente forma:

Definición 3.6.2. *Sea D un espacio de decision finito de cardinalidad k con elementos $d_1, d_2, \ldots, d_k$, y sea P una distribución de probabilidad definida sobre $\{1, 2, \ldots, k\}$ tal que $P_i \geq 0\ \forall i$ y $\sum_{i=1}^k P_i = 1$. Entonces, a un mecanismo que selecciona la decisión d_i con probabilidad P_i se le conoce como* **regla de decisión aleatorizada** *sobre D.*

Naturalmente, esta definición se puede extender sin problema al caso en que el espacio de decisión es infinito, e incluso no numerable, pero para el propósito de esta introducción basta considerar el caso finito.

Observe que, una regla de decisión aleatorizada constituye una combinación lineal convexa de elementos en D. Esto es, si a partir del problema original $(D, \mathcal{E}, C, \prec)$ se busca elegir la regla de decisión aleatorizada óptima, entonces, este problema puede representarse como $(\mathcal{D}^{(\mathcal{A})}, \mathcal{E}, C, \prec)$, donde $\mathcal{D}^{(\mathcal{A})}$ denota el conjunto de todas las decisiones aleatorizadas. Así, una rama típica del árbol de decisión correspondiente se observa en la figura 3.3. Como puede observarse en esta figura, elegir ∂ en $\mathcal{D}^{(\mathcal{A})}$ equivale a elegir una distribución de probabilidades P^∂ sobre D, por lo que al aplicar el criterio de utilidad esperada se obtiene que

$$\mathbb{E}\{u(\partial, E)\} = \sum_{i=1}^{k} P_i^\partial \mathbb{E}\{u(d_i, E)\}.$$

de forma que la utilidad esperada de ∂ es una combinación lineal convexa de las utilidades esperadas de $d_1, d_2 \ldots, d_k$.

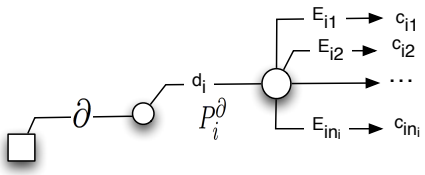


Figura 3.3: Rama típica del árbol de decisión bajo reglas de decisión aleatorizadas.

Algunas preguntas que son de interés cuando se consideran decisiones aleatorizadas son las siguientes: ¿Es posible alcanzar una mayor utilidad esperada? ¿Es posible encontrar nuevas soluciones óptimas al problema? Las respuestas a estas preguntas se examinan en la lista de ejercicios al final de este capítulo.

3.7. Extensiones del espacio del problema de decisión

Hasta ahora se ha examinado el caso en que D es un conjunto finito y, de hecho, también se ha considerado una colección finita de eventos inciertos $\mathcal{E}$. En este apartado se comenta brevemente el tipo de consideraciones adicionales que se requieren para extender los resultados a situaciones más generales.

- Si D es infinito, la búsqueda del máximo de la utilidad esperada debe contemplar el hecho de que la utilidad canónica está acotada y en consecuencia la utilidad esperada también lo está. Así queda garantizada la existencia de, al menos, un supremo de la utilidad esperada, y en el peor de los casos, será posible obtener una opción cuya utilidad esperada sea arbitrariamente cercana al correspondiente supremo. En este caso, dependiendo de si D es discreto o continuo, la función $g(d) = \mathbb{E}\{u(d, E)\}$ puede maximizarse con métodos de optimización discreta, o incluso de cálculo diferencial si g satisface las propiedades necesarias.

- Si la colección de eventos inciertos relevantes $\mathcal{E}$ es infinita, la distribución $P(E)$ puede corresponder a un modelo de probabilidad discreto con soporte numerable o bien a un modelo de probabilidad continuo. Aquí, de nuevo, el hecho de que la utilidad canónica es acotada, garantiza que la utilidad esperada existe para toda $d \in D$.

  Evidentemente, la determinación de la probabilidad de todos, y cada uno, de los elementos de la colección de eventos relevantes no es factible. La alternativa es identificar este valor para un reducido conjunto de eventos, utilizando el mecanismo que se presentará en la sección 4.2, y a partir de esta información, proponer un modelo que produzca una aproximación a todas las probabilidades requeridas.

3.8. Ejercicios

Ejercicio 3.1. *Considere un problema de decisión en ambiente de incertidumbre con estructura:* $D = \{d_1, d_2\}$, $\mathcal{E} = \{E_1, E_2\}$, *y* $C = \{c_{11}, c_{12}, c_{21}, c_{22}\}$. *Suponga que* $P(E_1) = P(E_2)$ *y que las consecuencias se registran con valores numéricos de manera que* c_i *es menos preferible que* c_j $(c_i \prec c_j)$ *si y sólo si* $c_i < c_j$. *Si además se tiene que*

$$c_{11} = \frac{1}{2} - \Delta, \quad c_{12} = \frac{1}{2} + \Delta, \quad c_{21} = \frac{1}{2} - 2\Delta \quad y \quad c_{22} = \frac{1}{2} + 3\Delta$$

Con Δ *una constante estrictamente positiva tal que todas las consecuencias son, a su vez, estrictamente positivas. Si la utilidad de las consecuencias se mide con la función identidad* $(u(c) = c)$, *entonces*

a) Construya el árbol de decisión correspondiente.

b) Demuestre que el criterio optimista conduce a elegir d_2 *como la decisión óptima para cualquier valor de* Δ.

c) Demuestre que el criterio pesimista conduce a elegir d_1 *como la decisión óptima para cualquier valor de* Δ.

d) Demuestre que el criterio de la utilidad esperada conduce a elegir d_2 *como la decisión óptima para cualquier valor de* Δ.

Si alternativamente, la utilidad de las consecuencias en este problema se miden con una función monótona creciente g del valor de las consecuencias, (u(c)=g(c)). Por facilidad, considere el caso en que $g(c) = ln(c)$.

e) En el caso del criterio de la utilidad esperada, demuestre que para diferentes valores de Δ *la decisión óptima puede variar o incluso pueden ser las dos decisiones igualmente preferibles. Explique las causas de este fenómeno y la importancia de este resultado.*

Ejercicio 3.2. *Considere el problema de decisión en donde el espacio de opciones consta de r posibles alternativas. Suponga además que la partición de eventos inciertos es la misma para cada una de ellas (con s elementos). Si tiene una función de pérdida que describe las preferencias entre las posibles consecuencias y una función de probabilidad* $\pi_i = \pi(E_i)$ *para cada* $i = 1, 2, \ldots, s$.

a) Para el caso en que $s = 2$, *represente equemáticamente el llamado conjunto de pérdida, es decir el conjunto de los puntos que representan vectorialmente las pérdidas en las que se puede incurrir para cada opción.*

b) Verifique el hecho de que dos opciones tienen la misma pérdida esperada si y sólo si yacen sobre la misma recta (hiperplano en el caso general) perpendicular al vector (π_1, π_2).

c) ¿Cual es el lugar geométrico de todas las decisiones equivalentes bajo el criterio minimax?

d) Tomado en cuenta el resultado del inciso anterior, ¿Cuáles decisiones en D, en la gráfica, podrían ser soluciones de Bayes para alguna distribución inicial?

e) ¿cuántas posibles decisiones aleatorizadas se podran construir en este caso?

f) Sin olvidar que la partición es la misma para todas las opciones y que el número de eventos inciertos relevantes s, es 2, sugiera una manera de representar en el diagrama que construyó en el inciso a), las decisiones aleatorizadas.

g) ¿Cuáles decisiones aleatorizadas podrían ser soluciones de Bayes si se agranda D para incluirlas junto con las decisiones originales? ¿En dónde se localizarían dentro de su diagrama?

Ejercicio 3.3. *El sentido común sugiere que la solución a un problema de decisión en ambiente de incertidumbre no debiese ser una decisión inadmisible, y entonces surge la conveniencia de eliminar, como un primer paso, las decisiones inadmisibles al resolver este tipo de problemas. Una propiedad de la solución Bayesiana (utilidad esperada máxima) es que, aún si no se eliminan previamente las decisiones inadmisibles, la solución de Bayes es siempre admisible. Demuestre esta afirmación en una situación en donde cada opción tiene un número finito de consecuencias. Suponga además, por facilidad, que los eventos inciertos relevantes son los mismos para cada opción y que todos tienen probabilidad de ocurrencia estrictamente positiva. ¿Cómo modificaría la demostración si las particiones de eventos inciertos no fueran iguales para todas las opciones?*

Ejercicio 3.4. *Considere el siguiente problema de decisión. En un juego, se tiene un conjunto de* 9 *cartas que contienen:* 2 *ases,* 3 *reyes y* 4 *sotas. Al jugador, que paga* 150 *pesos por el derecho a jugar, se le entrega una carta al azar de entre las nueve, una vez con esa primera carta en su poder, puede optar por pedir otra o bien pasar. Si decide pasar pierde su entrada, mientras que si decide pedir otra carta las recompensas se pagan de acuerdo a la siguiente tabla:*

Cartas	Recompensa
2 ases ó 2 reyes	$2,000$
2 sotas ó 1 as y 1 sota	$1,000$
Otras combinaciones	$-1,000$

Describa la estructura del problema y obtenga la decisión óptima (de Bayes) para un jugador que ya pagó su derecho de juego bajo los siguientes escenarios

a) Si resuelve decidir sin mirar la primera carta.

b) Si resuelve decidir sólo después de observar la primera carta.

c) ¿Cómo compararía los resultados de a) y b)?

Diga además

d) ¿Participaría usted en el juego?

e) ¿Con qué estrategia?

Ejercicio 3.5. *Un equipo mexicano de béisbol está sufriendo por la falta de asistencia de fanáticos a sus partidos. Antes de cambiar de sede (como los Tigres), los dueños están intentando decidir si emprenden una campaña de promoción que tiene un costo de* 15 *millones de pesos antes de que inicie el siguiente (y posiblemente último) torneo. Saben que la asistencia del público a los estadios depende, además del efecto de la campaña, del desempeño del equipo. A partir de la experiencia se considera que, si* θ *es la proporción de partidos que el equipo finalmente gana a lo largo del torneo futuro, los ingresos por asistencia serán de* $20 + 20\theta$ *millones de pesos si no lanzan la campaña. En caso contrario (si emprenden la campaña) entonces los ingresos por asistencia a los estadios serán de* $25 + 40\theta$*millones de pesos. Además, tienen el dato de que si el equipo logra ganar al menos el* 75 % *de sus partidos entonces pasará a las finales, en cuyo caso tendrá asegurados ingresos adicionales por* 10 *millones de pesos. Suponga que la utilidad es directamente proporcional al dinero y considerando, como primera aproximación, que la incertidumbre sobre* θ *se describe con una distribución uniforme en* $(0, 1)$.

a) Describa la estructura del problema.

b) Encuentre la solución Minimax.

c) Encuentre la solución de Bayes.

Si ahora la función de densidad de θ *está dada por* $f(\theta) = (a + 1)\theta^a$ *en* $(0, 1)$,

d) ¿Cuáles son los valores del exponente a que conducen a lanzar la campaña publicitaria con el criterio de Bayes?

Ejercicio 3.6. *Si en el problema 3.5 se plantea una función de probabilidad para* θ *totalmente general (en el intervalo* $(0, 1)$*),*

a) ¿Qué aspectos de esa distribución influyen en la solución si el problema se resuelve por el criterio de Bayes?

b) ¿Podría sugerir una decisión robusta, es decir que funcione independientemente de la distribución de la que se trate?

c) ¿Qué tan relevantes son los ingresos que se obtendrían en caso de pasar a las finales?

Ejercicio 3.7. *Suponga que un usuario del Servicio Postal se encuentra con que hay dos tipos de servicio que puede emplear para realizar sus envíos: Ordinario y Express. El costo para el paquete específico que desea enviar es de* 800 *ó* $1,000$ *pesos según elija el servicio Ordinario o Express.*

Además sabe que, de acuerdo a los registros del Servicio Postal, de cada 1000 *envíos que se realizan por servicio Ordinario,* 301 *llegan a su destino la mañana siguiente,* 299 *lo hacen la tarde siguiente, 287 lo hacen la segunda mañana y* 113 *lo hacen la segunda tarde; de igual manera, conoce que los números para el servicio Express son* $451,369,140$, *y* 40 *respectivamente.*

Si en esta ocasión el usuario esta dispuesto a pagar hasta $2,000$ *pesos si su paquete llega con toda seguridad la mañana siguiente, hasta* $1,600$ *pesos si llega con certeza la tarde siguiente,* $1,200$ *si lo hace la segunda mañana y* 800 *la segunda tarde, describa el problema del usuario como uno de decisión y analícelo para obtener una solución óptima.*

Capítulo 4

Probabilidad y utilidad

4.1. Probabilidad subjetiva

A diferencia del punto de vista Frecuentista, en la Teoría Bayesiana no es necesario que un evento sea aleatorio (en el sentido en que sus resultados se presentan con variabilidad) para que se le pueda asignar una probabilidad; el aspecto que es relevante es que exista incertidumbre sobre la eventual ocurrencia del evento.

Ejemplo 4.1.1. *Considerar el evento E = Manuel Mendoza vive a más de 10Km del Instituto Tecnológico Autónomo de México (ITAM).*

El evento E no es aleatorio (suponiendo que Manuel no se muda todos los días). Sin embargo, para alguien que desconoce su dirección el evento E es incierto, y por lo tanto, puede asignarle una probabilidad subjetiva $P(E)$.

4.2. Asignación de la probabilidad inicial

Si se considera un evento incierto E, para determinar la probabilidad $P(E)$ es posible someter al tomador de decisiones a un proceso de decisiones secuenciales a partir de loterías para obtener, así sea aproximado, el valor de esta probabilidad. Tomando $d_E = \{c_*|E^c, c^*|E\}$ y $d_p = \{c_*|1-p, c^*|p\}$ y puesto que $u_0(c_*) = 0$ y $u_0(c^*) = 1$, se tiene que

$$\mathbb{E}\{u(d_E, E)\} = P(E)u_0(c^*) + P(E^c)u_0(c_*) = P(E) \quad \text{y}$$
$$\mathbb{E}\{u(d_P, E)\} = p$$

Así, si además ocurre que $d_E \sim d_p$ entonces, necesariamente $P(E) = p$. Esta condición sugiere un algoritmo para la búsqueda de $P(E)$. Si se toma $p = \frac{1}{2}$ y ocurre que $d_p \prec d_E$ entonces, a partir de los axiomas de coherencia, se puede

asegurar que

$$P(E) \in \left[\frac{1}{2} , 1 \right].$$

Continuando con la misma idea, se puede tomar $P = \frac{3}{4}$ y proceder análogamente hasta que se acote el valor de $P(E)$ en un intervalo suficientemente pequeño para ser útil en la práctica. Finalmente, se puede tomar $P(E)$ igual al punto medio del intervalo obtenido.

En general, se trata de un método de búsqueda de bisección en el que se puede establecer arbitrariamente, y de antemano, el grado de precisión deseado.

4.3. Distribuciones no informativas

Un caso especial y muy interesante en la asignación de probabilidades iniciales ocurre cuando el tomador de decisiones considera que su información subjetiva es muy vaga, o cuando contando con información subjetiva clara desea reportar sus resultados tanto incorporando esta información como excluyendola. Este segundo escenario se puede presentar, por ejemplo, en el ámbito de la investigación científica, cuando se persigue el propósito de transparentar explícitamente el peso relativo en las conclusiones de un estudio que tiene la información adicional Z y la asignación inicial $P(E)$.

La idea original en estas circunstancias fue recurrir al empleo de iniciales que puedan interpretarse como descripciones de un estado de poca información (conocimiento vago). En un extremo, se llegó a denominar a las iniciales de este tipo como mínimo informativas, o incluso no informativas, por razones obvias. Más recientemente, se ha aceptado que el concepto de mínima o nula información no etsá unívocamente definido, y por esta razón se utiliza cada vez más frecuentemente el término de distribuciones de referencia para estas iniciales.

Existe una larga lista de contribuciones en la literatura cuyo objetivo es proponer distribuciones de este tipo. Probablemente el intento más célebre sea el de P. S. Laplace quien introdujo el llamado Principio de la razón insuficiente, que aplica en el caso de un fenómeno con un número finito de posibles resultados, y que establece que ante la ausencia de información no hay razón para que un resultado posible reciba una asignación de probabilidad distinta de otro. En otros términos, la "ignorancia" se representa con una distribución $Uniforme$.

Criterio de la razón insuficiente Si $E_1, E_2 \dots E_k$ son eventos inciertos relevantes y no hay razón para creer más en la ocurrencia de uno sobre otro entonces $P(E_i) = \frac{1}{k} \quad \forall i = 1, \dots, k$.

Es interesante observar que, cuando en los juegos de azar, por ejemplo el lanzamiento de una moneda o un dado, se dice que la moneda o el dado son honestos, lo que se supone es que sus resultados siguen una distribución uniforme, es decir no informativa.

4.4. Utilidad y pérdida

En la sección 3.2 se introdujo el concepto de utilidad canónica, esta función de utilidad es sumamente conveniente pues provee al tomador de decisiones de una forma para calcular la utilidad esperada de una decisión, y por tanto encontrar la solución de Bayes para cualquiera que sea el problema al que se enfrente. Esto es

$$\mathbb{E}\{u(d,E)\} = \sum_{j=1}^{r} u_0(c_j)P(E_j).$$

Sin embargo, en la práctica existen algunos problemas en los que se podría estar interesado en utilizar una función de utilidad distinta a la canónica. Un resultado interesante, que además es fácil de comprobar, es que se puede utilizar cualquier transformación lineal de u_0, es decir

$$\mathbb{E}\{u(d,E)\} = \sum_{j=1}^{r} u(c_j)P(E_j)$$

donde $u(c_j) = au_0(c_j) + b$ con $a, b \in \Re$, y la solución del problema será afectada exclusivamente por el valor de a en la siguiente manera:

- Si $a > 0$ la solución no cambia
- Si $a = 0$ no refleja el problema original
- Si $a < 0$ la solución óptima se obtiene minimizando $\mathbb{E}\{u(d,E)\}$

En el caso en que $a < 0$ a $u(c)$ se le conoce como función de pérdida y generalmente se denota por $L(c)$. De hecho, en muchas ocasiones resulta más fácil o práctico utilizar una función de pérdida en lugar de su correspondiente función de utilidad.

Ejemplo 4.4.1. *Considere el ejemplo 2.3.1 de las elecciones parlamentarias británicas, recuerde que en este caso, debido a que las consecuencias son de entrada numéricas, es posible resolver el problema directamente utilizando la identidad como función de utilidad. Esto es, maximizando sobre $J = \{l, c\}$ la utilidad esperada $E\{u(d_j, E)\} = c_{j1}P(E_1) + c_{j2}P(E_2)$ con $j \in J$, que como ya se ha visto tiene solución dada por $d_c \Leftrightarrow P(E_1) \geq 5/12$.*

Observe que, alternativamente se podría haber resuelto el problema utilizando la función de pérdida $L(c_{jk}) = -u(c_{jk}) = -c_{jk}$ con $j \in J$ y $k \in \{1, 2\}$, lo que lleva a minimizar la pérdida esperada, es decir

$$\begin{aligned} \textit{Apostar por los conservadores} &\iff \mathbb{E}\{L(d_c,E)\} \leq \{L(d_c,E)\} \\ &\iff 1 - 7/4P(E_1) \leq 5/4P(E_1) - 1/4\mathbb{E} \\ &\iff 5/12 \leq P(E_1). \end{aligned}$$

Verificando que la solución se mantiene inalterada a pesar del cambio en la elección de la función de utilidad.

4.5. Asignación de la utilidad

Al igual que con las creencias, para determinar la función de utilidad canónica del tomador de decisiones, es posible someter a este a un proceso interrogatorio mediante loterías que genere un proceso de bisección. Así, para toda consecuencia $c \in C$ tal que $c_* \preceq c \preceq c^*$, se puede enfrentar al tomador de decisiones a la elección entre $d_r = \{c_*|R^c, c^*|R\}$ y $d_c = \{c|R^c, c|R\}$, donde R es un evento de referencia que inicialmente cumple que $Area(R) = r$. De esta manera, si originalmente se toma $r = \frac{1}{2}$ y sucede que $d_r \prec d_c$ entonces, necesariamente se cumple que

$$u(c) \in \left[\frac{1}{2}, 1\right]$$

Continuando con la misma idea, se puede modificar R para que cumpla con que $r = \frac{3}{4}$ y proceder análogamente hasta que se acote el valor de $u(c)$ en un intervalo suficientemente pequeño para ser útil en la práctica. Finalmente, se puede tomar $u(c)$ igual al punto medio del intervalo obtenido.

4.6. Utilidad del dinero

Como ya se ha indicado, existen problemas en los que las consecuencias tienen, directamente, una naturaleza numérica y ese valor preserva el orden de preferencias del tomador de decisión. Un ejemplo es el problema del contenedor que se discutió en la sección 2.1.

Una clase de problemas, con o sin incertidumbre, donde esto ocurre es la que se tiene cuando todas las consecuencias se registran en un valor monetario. Aquí, una práctica común es considerar $u(c) = c$ para cualquier cantidad de dinero c, o $u(c) = -c$ en caso de que se tratase de pérdidas.

Sin embargo, en la práctica se ha observado que esta forma de asignar la utilidad no es, en general, apropiada. Si efectivamente este fuera el caso, todos los tomadores de decisión tendrían la misma actitud de preferencia frente al dinero. Y en particular, si a cada uno se le sometiera al proceso descrito en la sección anterior, el resultado sería una línea recta que pasa por los puntos $(c_*, 0)$, y $(c^*, 1)$, donde c_* y c^* son las cantidades mínima y máxima de dinero involucrado.

Ahora bien, esto significa que si se toma $c = \frac{c_* + c^*}{2}$, entonces necesariamente $u(c) = \frac{1}{2}$. Pero entonces, si se define $d_{\frac{1}{2}} = \left\{c_*|\frac{1}{2}, c^*|\frac{1}{2}\right\}$ debe cumplirse que $c \sim d_{\frac{1}{2}}$ puesto que $\mathbb{E}_u\left(d_{\frac{1}{2}}\right) = \frac{1}{2}$. En otras palabras, todo tomador de decisiones sería indiferente entre una opción que le garantiza la cantidad c, y una con incertidumbre que le ofrece c_* con probabilidad $\frac{1}{2}$ y c^* con esta misma probabilidad.

En contraste, distintos experimentos han mostrado que diferentes personas, sin dejar de ser coherentes, reaccionan diferente frente a las opciones c y $d_{\frac{1}{2}}$. Mien-

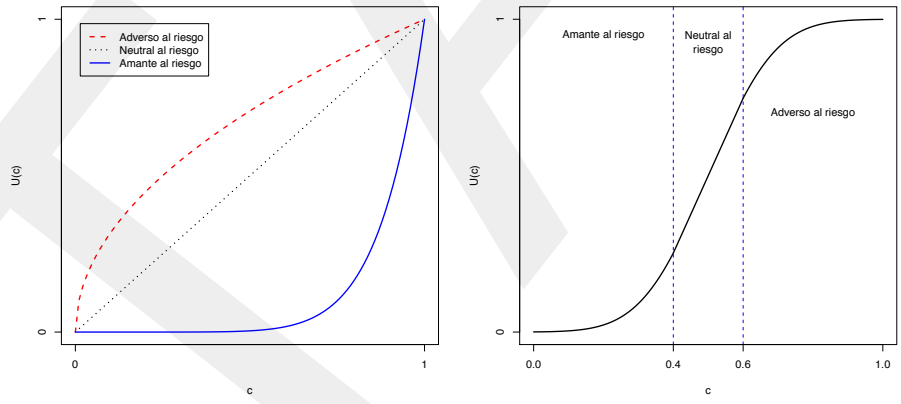
tras algunos prefieren la cantidad segura, otros prefieren arriesgarse con $d_{\frac{1}{2}}$ con la esperanza de obtener c^* e incluso, posiblemente unos más sean efectivamente indiferentes entre estas opciones.

Más aún, un mismo tomador de decisiones coherente puede preferir c, $d_{\frac{1}{2}}$, o incluso ser indiferente entre ambos si los valores de c_* y c^* se modifican lo suficiente.

De hecho, cuando un tomador de decisiones prefiere sistemáticamente las opciones seguras, se le llama adverso al riesgo; cuando por el contrario prefiere la incertidumbre con la ilusión de alcanzar una recompensa mayor, se dice que es amante al riesgo; y si es indiferente en situaciones como la descrita se le denomina neutral al riesgo.

En la gráfica 4.1a se exhiben tres posibles formas de la utilidad del dinero. En estas, cada una tiene un tipo distinto de preferencia por el dinero. En un caso general, incluso estos patrones se pueden combinar en una misma función de utilidad como la presentada en la gráfica 4.1b

De esta forma, se puede observar que los únicos tomadores de decisiones para los cuales es conveniente una utilidad del dinero de la forma $u(c) = c$ son aquellos neutros frente al riesgo. Vale la pena insistir en que la condición de neutralidad (al igual que las otras) depende de la diferencia $c^* - c_*$.



(a) (b)

Figura 4.1: Formas de la función de utilidad del dinero.

4.7. Ejercicios

Ejercicio 4.1. *Sea W la temperatura en grados centígrados que se registrá, al pie del asta bandera de ITAM, el día de mañana a las $17:30$ horas.*

a) *Determine el valor a tal que, en su opinión, $P(W \leq a) = 0{,}5$*

b) *Determine el valor b tal que, en su opinión, $P(W \leq b) = 0{,}25$*

c) *Determine el valor c tal que, en su opinión, $P(W \leq c) = 0{,}9$*

d) *Utilizando la información de los apartados a) y b) encuentre una distribución Normal que se asigne a sus asignaciones. ¿Cuántas existen?*

e) *Confrontando c) con d), ¿Encuentra concordancia? En caso negativo, ¿Cuál cree que sea la causa?*

Ejercicio 4.2. *Considere los siguientes tres eventos: A_1 es el evento de que el primer Doctor en Estadística mexicano (Dr. Basilio Rojas) haya obtenido el grado doctoral antes de 1955; A_2 es el evento de que lo haya obtenido entre 1955 y 1975. Finalmente, A_3 es el evento de que lo haya obtenido después de 1975.*

a) *Por lo pronto, y únicamente con la información al momento de leer el párrafo anterior, asigne sus probabilidades a los eventos $A_i : \ i = 1, 2, 3$.*

b) *Reconsidere sus probabilidades tomando en cuenta la siguiente información: en 1970 obtuvo el grado doctoral el cuarto Doctor mexicano en Estadística (Dr. Ignacio Méndez).*

c) *Reconsidere una vez más. Ahora se le informa que el Dr. Federico O'Reilly, veterano pero plenamente activo en el campo, obtuvo su grado doctoral en Estadística por la North Carolina State University en 1971.*

d) *Comente sobre el impacto de la información de los incisos b) y c) en la asignación inicial en a).*

Ejercicio 4.3. *Considere el siguiente juego de azar: una moneda se lanza tantas veces como sea necesario hasta que aparece un sol por primera vez. Entonces, se registra el número r de lanzamientos que se han efectuado y el jugador, a cambio de un boleto de entrada, recibe como premio la cantidad de $2r$ pesos.*

a) *Calcule el premio esperado del juego.*

b) *Si se plantea el problema de decidir si compra el boleto (d_1) o no lo compra (d_2) para participar en este juego, y la utilidad se mide exclusivamente en términos monetarios, diga cuál sería su decisión en función del precio del boleto.*

Ejercicio 4.4. *Considere nuevamente el juego del problema 4.3, modificado en forma que ahora el jugador recibe un premio de 2^r pesos.*

a) Demuestre que el premio esperado es infinito.

b) Nuevamente, si la utilidad se mide solamente en términos económicos, ¿Compraría el boleto para participar en este juego?

c) ¿Cómo explica que prácticamente ninguna persona está dispuesto a pagar más allá de una pequeña suma de dinero por el boleto?

Ejercicio 4.5. *Luis Enrique contempla la posibilidad de viajar a Belo Horizonte, Brasil, donde espera entrevistarse con la Dra. Rosangela Loschi, una exitosa mujer de negocios radicada en esta ciudad, para proponerle la compra de una plataforma para el desarrollo de aplicaciones de análisis estadístico Bayesiano. Si consigue su objetivo, ganará una comisión de* $60,000$ *pesos.*

Ahora bien, la Dra. Loschi viaja mucho y Luis Enrique considera que con una probabilidad de $0{,}4$ *puede ocurrir que, si viaja a ese país, la Dra. Loschi tenga que salir de Brasil y, por tal causa, no sea posible celebrar la entrevista ni realizar la venta. Por otra parte, aún en el caso en que se produzca la entrevista, Luis Enrique considera que la probabilidad de realizar la venta es de* $0{,}7$*. El viaje a Belo Horizonte cuesta* $8,000$ *pesos y sale del bolsillo de Luis Enrique.*

a) Tomando en cuenta que su interés en estrictamente económico y, por facilidad, suponiendo que en el intervalo de montos considerado, la función de utilidad del dinero se puede considerar lineal ¿Le conviene a Luis Enrique viajar a Belo Horizonte?

Adicionalmente, resulta que una agencia de información, propiedad de Eduardo, ofrece sus servicios a Luis Enrique. Le asegura que le podrá informar, antes de que emprenda el viaje hacia Belo Horizonte, si la Dra. Loschi se encontrará ahí cuando él llegue a esa ciudad. La agencia se autocalifica como altamente confiable y funda esa calificación en su registro histórico de aciertos. De acuerdo a este registro, si una persona efectivamente está en la ciudad designada, la agencia lo informa correctamente un $90\,\%$ *de la veces; por otro lado, si la persona sale de la ciudad, la agencia detecta su ausencia en un* $80\,\%$ *de los casos. El servicio de la agencia cuesta* $5,000$ *pesos.*

b) ¿Le conviene a Luis Enrique contratar el servicio de la agencia de Eduardo?

c) ¿Debe viajar cuando la agencia le dice que la Dra. Loschi sí va a estar en Belo Horizonte?

d) ¿Cuál es el precio máximo que Luis Enrique debería estar dispuesto a pagar a la agencia de Eduardo por la información que le ofrece?

Ejercicio 4.6. *Una editorial está considerando lanzar una revista mensual con artículos de interés para inversionistas. Ya cuenta con un plan de producción, distribución y promoción concreto y, como es habitual, las ganancias del proyecto dependen de la demanda que tenga la revista. El gerente de la editorial considera*

por facilidad, tres escenarios alternativos: una demanda baja (B), una demanda moderada (M) y una demanda alta (A). Además con base en su experiencia asigna las probabilidades $P(B) = 0{,}3,\ P(M) = 0{,}5$ *y* $P(A) = 0{,}2$. *Finalmente, considera un horizonte de un año en el que las ganancias del proyecto en pesos serían:*

Demanda	Ganancia
Baja	$-7,500,000$
Moderada	$1,500,000$
Alta	$9,000,000$

Si por facilidad se considera que la utilidad está convenientemente medida a través del dinero y en caso de no lanzar la revista no hay ganancia ni pérdida,

a) *Identifique la estructura del problema y verifique si existe alguna decisión inadmisible.*

b) *Encuentra la solución minimax y el valor minimax.*

c) *Encuentre la solución de Bayes y el valor de Bayes.*

d) *Represente gráficamente el conjunto de todas las distribuciones de probabilidad para las cuales coinciden las soluciones minimax y de Bayes.*

Suponga ahora que un subgerente se presenta afirmando que sí se debe lanzar la revista, y apoya su aseveración en el hecho de que realizó una prueba de aceptación de la nueva publicación a través de una encuesta y el resultado fue favorable (F). Si se sabe que $P(F \mid B) = 0{,}1,\ P(F \mid M) = 0{,}6$ *y* $P(F \mid A) = 0{,}7$,

e) *Incorporando la información de la encuesta adicional, ¿Usted también lanzaría la revista?*

Capítulo 5

La inferencia como problema de decisión

Los problemas clásicos de la inferencia paramétrica que aparecen en los textos más comunes son: estimación puntual, estimación por regiones y contraste de hipótesis. Tradicionalmente, además, se presentan en ese orden atendiendo una lógica de simplicidad en las técnicas y conceptos necesarios para su solución. Como se verá en lo que resta de este capítulo, desde una perspectiva Bayesiana, el problema que tiene una estructura más sencilla es el de contraste de hipótesis; los problemas de estimación puntual y por intervalos tienen una estructura un poco más compleja, pero con el antecedente de contraste de hipótesis pueden ser abordados sin dificultad.

Una novedad es la introducción de otro problema básico de inferencia que no aparece en los textos introductorios, aquel de pronósticos, puntuales y por intervalos, que resultan casi triviales una vez que ya se han discutido los de estimación correspondientes.

5.1. Contraste de hipótesis

Sea X una v.a. con función de densidad de probabilidad generalizada (f.d.p.g.) $P(x \mid \theta)$, $\theta \in \Theta = \{\theta_0, \theta_1\}$ y $P(x \mid \theta)$ tiene distribución conocida. Se desea contrastar las hipótesis paramétricas simples $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$.

Considere un problema de decisión donde los elementos del conjunto de eventos inciertos relevantes están dados por

$E_0 = P(x \mid \theta_0)$ es el modelo que mejor representa la realidad y
$E_1 = P(x \mid \theta_1)$ es el modelo que mejor representa la realidad

y en el que el conjunto de decisiones está dado por $D = \{d_0, d_1\}$ donde d_0

representa describir a X con $P(x\,|\,\theta_0)$ y d_1 describir a X con $P(x\,|\,\theta_1)$.

Como ya se sabe, este problema puede ser representado gráficamente mediante el árbol de decisión, presentado en la figura 5.1.

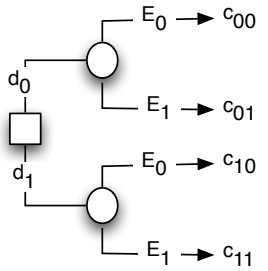


Figura 5.1: Árbol de decisión para el problema de contraste de hipotesis.

Así, la ocurrencia del evento E_0 implicaría que la hipóteis H_0 es verdadera y, por el contrario, si sucediera E_1, entonces H_1 sería correcta. De esta manera, contrastar las hipótesis H_0 v.s. H_1 implica elegir entre d_0 y d_1.

Ahora, observe que θ a pesar de ser un valor fijo, es desconocido y por tanto, como se discutió en el capítulo anterior, es posible asignarle una medida subjetiva de probabilidad $P(\theta)$ que describa la incertidumbre que se tiene sobre el parámetro θ. De esta manera, se denota $P_0 = P(E_0) = P(H_0) = P(\theta = \theta_0)$ y análogamente $P_1 = 1 - P_0 = P(E_1) = P(H_1) = P(\theta = \theta_1)$.

Adicionalmente, sea $l = L(c)$ la función de pérdida, y $l_{ij} = L(c_{ij})$. Y puesto que acertar debe ser preferido a cometer cualquier tipo de error, necesariamente se deberá cumplir que $c_{10} \prec c_{00}$, $c_{01} \prec c_{00}$ y también que $c_{10} \prec c_{11}$, $c_{01} \prec c_{11}$.

		Naturaleza H_0	Naturaleza H_1
Decisión	d_0	Acierto c_{00}	ET2 c_{01}
	d_1	ET1 c_{10}	Acierto c_{11}

Finalmente, se puede suponer que $l_{00} = l_{11} = c$. Es decir, que acertar eligiendo d_0 es igualmente preferido que hacerlo eligiendo d_1, y por tanto $\mathbb{E}\{L(d_0|E)\} = P_0 l_{00} + P_1 l_{01}$ y $\mathbb{E}\{L(d_1|E)\} = P_0 l_{10} + P_1 l_{11}$. Lo que conduce a que a priori

$$
\begin{aligned}
d_1 \text{ es la solución de Bayes} &\iff \mathbb{E}\{L(d_1|E)\} > \mathbb{E}\{L(d_0|E)\} \\
&\iff P_0 l_{10} + P_1 l_{11} > P_0 l_{00} + P_1 l_{01} \\
&\iff (l_{01} - l_{11}) P_1 > (l_{10} - l_{00}) P_0 \\
&\iff k = \frac{(l_{01} - l_{11})}{(l_{10} - l_{00})} > \frac{P_0}{1 - P_0} \\
&\iff \frac{k}{1+k} > P_0
\end{aligned}
$$

Esto es, d_1 es solución de Bayes sólo si P_0 es suficientemente pequeño. Observe que, equivalentemente, es posible realizar la transformación $L' = L - c$ de forma que $l'_{00} = l'_{11} = 0$, y por ende a priori

$$\text{se rechaza } H_0 \Longleftrightarrow \frac{l'_{01}}{l'_{10}} > \frac{P_0}{P_1}$$

Ahora, sea $x_{(n)}$ una m.a. de tamaño n de X. Entonces, utilizando la regla de Bayes

$$P(\theta_0|x_{(n)}) = \frac{P(x_{(n)}|\theta_0)P(\theta_0)}{P(x_{(n)})} \quad y \quad P(\theta_1|x_{(n)}) = \frac{P(x_{(n)}|\theta_1)P(\theta_1)}{P(x_{(n)})},$$

por lo que a posteriori

$$\text{se rechaza } H_0 \Longleftrightarrow \frac{l'_{01}}{l'_{10}} > \frac{P(\theta_0|x_{(n)})}{P(\theta_1|x_{(n)})}.$$

Así,

$$\frac{P(\theta_0|x_{(n)})}{P(\theta_1|x_{(n)})} = \frac{\dfrac{P(x_{(n)}|\theta_0)P(\theta_0)}{P(x_{(n)})}}{\dfrac{P(x_{(n)}|\theta_1)P(\theta_1)}{P(x_{(n)})}} = \frac{P(x_{(n)}|\theta_0)}{P(x_{(n)}|\theta_1)}\frac{P_0}{P_1}$$

lo que implica que

$$\text{se rechaza } H_0 \Longleftrightarrow \boldsymbol{C} \equiv \frac{P_1 l'_{01}}{P_0 l'_{10}} > \frac{P(x_{(n)}|\theta_0)}{P(x_{(n)}|\theta_1)}$$

El hecho más destacado de este resultado, que como puede observarse, es totalmente general (no depende de las particulares hipótesis simples ni del modelo de los datos), es el que establece que la muestra $x_{(n)}$ interviene en la decisión sobre las hipótesis única y exclusivamente a través de *cociente de verosimilitudes*

$$\Lambda = \frac{P(x_{(n)}|\theta_0)}{P(x_{(n)}|\theta_1)},$$

pudiendo así establecer una regla de decisión $\delta : \mathfrak{X}_{(\mathfrak{n})} \to D$ tal que

$$\delta(x_{(n)}) = \begin{cases} d_1 & \text{si } \boldsymbol{C} > \dfrac{P(x_{(n)}|\theta_0)}{P(x_{(n)}|\theta_1)} \\ d_0 & \text{e.o.c} \end{cases}$$

De hecho, si se recupera la idea Frecuentista de región de rechazo (para H_0), entonces el procedimiento Bayesiano establece que H_0 se rechaza si y sólo si $x_{(n)} \in \mathscr{C}$ donde

$$\mathscr{C} = \left\{x_{(n)} \in \mathfrak{X}_{(\mathfrak{n})} \mid \boldsymbol{C} > \Lambda\right\}.$$

En otras palabras, se obtiene una región de rechazo que tiene la misma *forma* que la que se sigue del conocido lema de Neyman-Pearson. Es importante insistir en que sólo se recupera la forma porque la constante $\boldsymbol{C}$ se determina por procedimientos conceptual y técnicamente muy distintos.

Ejemplo 5.1.1. *Sea X una v.a. con distribución $Normal(x|\mu,1)$, se desea contrastar las hipótesis $H_0:\ \mu=0$ vs $H_1:\ \mu=1$.*

En este caso, la función de verosimilitud está dada por

$$L(\mu \mid x_{(n)}) = \prod_{i=1}^{n}(2\pi)^{-1/2}e^{-(x_i-\mu)^2/2} = (2\pi)^{-n/2}e^{-1/2\sum(x_i-\mu)^2}$$

y por tanto, el cociente de verosimilitudes puede ser simplificado de la siguiente manera:

$$\begin{aligned}\frac{L(\mu_0 \mid x_{(n)})}{L(\mu_1 \mid x_{(n)})} &= \frac{(2\pi)^{-n/2}}{(2\pi)^{-n/2}}\frac{e^{-1/2\sum(x_i)^2}}{e^{-1/2\sum(x_i-1)^2}} = e^{1/2\left[\sum(x_i-1)^2-\sum(x_i)^2\right]}\\ &= e^{1/2\left[\sum(x_i^2-2x_i+1)-\sum x_i^2\right]} = e^{n/2-\sum x_i}\\ &= e^{n(1/2-\bar{x})}.\end{aligned}$$

Así, bajo el enfoque Frecuentista

$$\begin{aligned}\mathscr{C} &= \left\{x_{(n)} \in \mathfrak{X}_{(\mathfrak{n})} \mid e^{n(1/2-\bar{x})} < \boldsymbol{K}\right\} = \left\{x_{(n)} \in \mathfrak{X}_{(\mathfrak{n})} \mid n(1/2-\bar{x}) < \ln \boldsymbol{K}\right\}\\ &= \left\{x_{(n)} \in \mathfrak{X}_{(\mathfrak{n})} \mid \bar{x} > \boldsymbol{K}'\right\}\end{aligned}$$

y fijando la probabilidad del error de tipo I ($P(rechazar(H_0 \mid H_0))$, igual a un valor fijo α, se determina por completo la región de rechazo. En este caso,

$$\begin{aligned}P\left\{x_{(n)} \in \mathscr{C} \mid H_0\right\} = \alpha &\implies P\left\{\bar{x} > \boldsymbol{K}' \mid \mu = 0\right\} = \alpha\\ &\implies P\left\{\bar{x} > \boldsymbol{K}' \mid Normal\left(\bar{x} \mid 0, 1/n\right)\right\} = \alpha\\ &\implies P\left\{\bar{x} \le \boldsymbol{K}' \mid Normal(\bar{x}|0,1/n)\right\} = 1-\alpha\\ &\implies P\left\{\frac{\bar{x}}{\sqrt{1/n}} \le \frac{\boldsymbol{K}'}{\sqrt{1/n}}\middle| Normal\left(\frac{\bar{x}}{\sqrt{1/n}}\middle|0,1\right)\right\} = 1-\alpha\\ &\implies \frac{\boldsymbol{K}'}{\sqrt{1/n}} = Z_{1-\alpha} \quad \implies \quad \boldsymbol{K}' = \frac{1}{\sqrt{1/n}}Z_{1-\alpha}.\end{aligned}$$

Es interesante observar que, bajo este enfoque, si se denomina β a la probabilidad del error de tipo II ($P(aceptar(H_0 \mid H_1))$, y $n \to \infty$, entonces $\beta \to 0$ sin embargo $P(error\ tipo\ I) = \alpha\ \forall n$.

Ahora, desde el enfoque Bayesiano la región de rechazo está dada por

$$\begin{aligned}\mathscr{C} = \left\{x_{(n)} \in \mathfrak{X}_{(\mathfrak{n})} \mid \frac{P(x_{(n)}|\theta_0)}{P(x_{(n)}|\theta_1)} < \frac{P_1 l'_{01}}{P_0 l'_{10}}\right\} &= \left\{x_{(n)} \in \mathfrak{X}_{(\mathfrak{n})} \mid e^{n(1/2-\bar{x})} < \boldsymbol{C}\right\}\\ &= \left\{x_{(n)} \in \mathfrak{X}_{(\mathfrak{n})} \mid \bar{x} > -\frac{\ln \boldsymbol{C}}{n} + \frac{1}{2}\right\},\end{aligned}$$

observe que, bajo este enfoque tanto α como β tienden a cero cuando $n \to \infty$.

5.2. Estimación puntual

Uno de los problemas más conocidos y estudiados de la inferencia paramétrica es el de estimación puntual. Como se comentó en el capítulo 1, este pudiera considerarse el problema original de inferencia paramétrica. Identificar *el* valor de θ permite determinar la función de de distribución de la variable aleatoria bajo estudio. Por tanto, se trata de utilizar la información disponible para producir un valor $\hat{\theta}$ que aproxime a θ.

Sea X una v.a. con f.d.p.g. $f(x \mid \theta)$, $\theta \in \Theta$, se desea estimar puntualmente a θ. La idea es proponer un valor de $\hat{\theta}$ como aproximación de el valor desconocido θ. Así, para expresar este problema como uno de decisión se define

$$D = \{d_{\hat{\theta}} \mid \hat{\theta} \in \Theta\}$$

donde $d_{\hat{\theta}}$ =estimar a θ con $\hat{\theta}$. Observe que en este caso el tamaño de D está determinada por la cardinalidad del conjunto Θ, por lo que también la representación gráfica del problema, mediante el árbol de decisión estará afectada por este conjunto. Sin embargo, es posible mostrar una rama *genérica* de este tal como se hace en la figura 5.2.

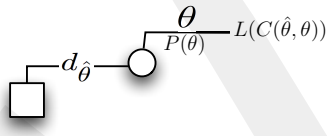


Figura 5.2: Rama típica del árbol de decisión para el problema de estimación puntual.

En este caso

$$\mathbb{E}\{L(d_{\hat{\theta}}, \theta)\} = \int_{\Theta} L(d_{\hat{\theta}}, \theta) P(\theta) d\theta = h(\hat{\theta}).$$

Observe que las consecuencias de una estimación $\hat{\theta}$ dependen de lo bien que se reproduzca el valor desconocido θ. De esta forma, resulta apropiado utilizar funciones de pérdida que dependan de la distancia entre $\hat{\theta}$ y θ, y en este sentido, que entre mayor sea dicha distancia mayor sea la pérdida. En particular, una opción es utilizar la función de pérdida cuadrática $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, de modo que la solución se obtiene de

$$\min_{\hat{\theta} \in \Theta} \mathbb{E}_{P(\theta)}\{L(d_{\hat{\theta}}, \theta)\} = \min_{\hat{\theta} \in \Theta} \mathbb{E}_{P(\theta)}\{(\hat{\theta} - \theta)^2\}$$

donde $\mathbb{E}_{P(\theta)}\{(\hat{\theta} - \theta)^2\}$ se conoce como error cuadrático medio bayesiano, y que se puede desarrollar como

$$\begin{aligned}
\mathbb{E}_{P(\theta)}\{(\hat{\theta} - \theta)^2\} &= \mathbb{E}_{P(\theta)}\left\{(\hat{\theta} - \mathbb{E}(\theta) + \mathbb{E}(\theta) - \theta)^2\right\} \\
&= \mathbb{E}_{P(\theta)}\left\{(\hat{\theta} - \mathbb{E}(\theta))^2\right\} + 0 + \mathbb{E}_{P(\theta)}\left\{(\mathbb{E}(\theta) - \theta)^2\right\} \\
&= \Big(\hat{\theta} - \mathbb{E}(\theta)\Big)^2 + Var(\theta)
\end{aligned}$$

de donde se obtiene que

$$\min_{\hat{\theta}\in\Theta} \mathbb{E}_{P(\theta)}\left\{L(d_{\hat{\theta}},\theta)\right\} \iff \min_{\hat{\theta}\in\Theta} \left(\hat{\theta}-\mathbb{E}(\theta)\right)^2.$$

Observe que, en general $d_{\hat{\theta}}$ de Bayes es el valor en Θ más cercano a $\mathbb{E}(\theta)$. En particular si $\mathbb{E}(\theta)\in\Theta$ el valor de Bayes (a priori) es $Var(\theta)$.

Ahora, sea $x_{(n)}$ una m.a. de tamaño n de X. Utilizando el teorema de Bayes $p(\theta)\overrightarrow{P(x_{(n)}|\theta)} P(\theta\,|\,x_{(n)})$, de modo que si $\mathbb{E}\left\{(\theta\,|\,x_{(n)})\right\}\in\Theta$ entonces se deberá cumplir que $\hat{\theta}_B = \mathbb{E}\left\{(\theta\,|\,x)\right\}$ y el valor de Bayes (a posteriori) resulta $Var(\theta\,|\,x_{(n)})$.

Definición 5.2.1. Sea W una v.a. con varianza σ^2_W, el parámetro $\tau_W \equiv \frac{1}{\sigma^2_W}$ se conoce como la precisión de W.

Ejemplo 5.2.1. *Sea X una v.a. con distribución $Normal(x\,|\,\mu,\sigma^2)$ con σ^2 conocida. Se desea estimar puntualmente a μ (utilizando pérdida cuadrática), suponiendo que a priori se describe el conocimiento sobre μ con un modelo $Normal(\mu\,|\,m,c^2)$.*

Se sabe que a priori la solución de Bayes es $\hat{\mu}_B = m$ con un valor de Bayes de $V_B = c^2$.

Ahora, a posteriori

$$f(\mu\,|\,x_{(n)}) = \frac{f(x_{(n)}\,|\,\mu)f(\mu)}{f(x_{(n)})}, \quad i.e. \quad f(\mu\,|\,x_{(n)})\,\alpha\, f(x_{(n)}\,|\,\mu)f(\mu).$$

Por lo que la verosimilitud está dada por

$$\begin{aligned} f(x_{(n)}\,|\,\mu) &= \prod_{i=1}^{n} f(x_i\,|\,\mu) = \prod_{i=1}^{n}\left\{(2\pi\sigma^2)^{-1/2}\, e^{-(x_i-\mu)^2/2\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-n/2}\, e^{-\sum(x_i-\mu)^2/2\sigma^2}, \end{aligned}$$

y puesto que la inicial cumple que

$$f(\mu) = (2\pi c^2)^{-1/2}\, e^{-(\mu-m)^2/2c^2},$$

la final se determina por

$$f(\mu\,|\,x_{(n)})\,\alpha\, e^{-\sum(x_i-\mu)^2/2\sigma^2}\, e^{-(\mu-m)^2/2c^2} = e^{-\left(\sum(x_i-\mu)^2/2\sigma^2+(\mu-m)^2/2c^2\right)}.$$

Denotando $m_x = \dfrac{\left(\frac{n}{\sigma^2}\right)\bar{x}+\left(\frac{1}{c^2}\right)m}{\left(\frac{n}{\sigma^2}+\frac{1}{c^2}\right)}$, *y analizando el exponente de esta expresión:*

$$\begin{aligned} exp \equiv& \frac{1}{\sigma^2}\sum(x_i-\mu)^2 + \frac{1}{c^2}(\mu-m)^2 \\ =& \frac{1}{\sigma^2}\left[\sum(x_i^2-2x_i\mu+\mu^2)\right] + \frac{1}{c^2}\left[\mu^2-2\mu m+m^2\right] \\ =& \frac{1}{\sigma^2}\sum x_i^2 - \frac{2\mu n\bar{x}}{\sigma^2} + \frac{n\mu^2}{\sigma^2} + \frac{\mu^2}{c^2} - \frac{2\mu m}{c^2} + \frac{m^2}{c^2} \end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{n}{\sigma^2}+\frac{1}{c^2}\right)\mu^2 - 2\mu\left(\frac{n\bar{x}}{\sigma^2}+\frac{m}{c^2}\right) + \left(\frac{\sum x_i^2}{\sigma^2}+\frac{m^2}{c^2}\right) \\
&= \left(\frac{n}{\sigma^2}+\frac{1}{c^2}\right)\left[\mu^2 - 2\mu\frac{\left(\frac{n\bar{x}}{\sigma^2}+\frac{m}{c^2}\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{c^2}\right)}\right] + \left(\frac{\sum x_i^2}{\sigma^2}+\frac{m^2}{c^2}\right) \\
&= \left(\frac{n}{\sigma^2}+\frac{1}{c^2}\right)\left[\mu^2 - 2\mu m_x\right] + \left(\frac{\sum x_i^2}{\sigma^2}+\frac{m^2}{c^2}\right) \\
&= \left(\frac{n}{\sigma^2}+\frac{1}{c^2}\right)\left[\mu^2 - 2\mu m_x + m_x^2\right] - \left(\frac{n}{\sigma^2}+\frac{1}{c^2}\right)m_x^2 + \left(\frac{\sum x_i^2}{\sigma^2}+\frac{m^2}{c^2}\right) \\
&= \left(\frac{n}{\sigma^2}+\frac{1}{c^2}\right)(\mu - m_x)^2 - \left(\frac{n}{\sigma^2}+\frac{1}{c^2}\right)m_x^2 + \left(\frac{\sum x_i^2}{\sigma^2}+\frac{m^2}{c^2}\right),
\end{aligned}
$$

por lo que

$$f(\mu \,|\, x_{(n)})\, \alpha\, e^{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{c^2}\right)(\mu - m_x)^2}\, e^{-\frac{1}{2}K(\sigma^2, c^2, x_{(n)})},$$

donde se observa que $f(\mu \,|\, x_{(n)}) = Normal(\mu \,|\, m_x, c_x^2)$ *con* $c_x^2 = \left(\frac{n}{\sigma^2}+\frac{1}{c^2}\right)^{-1}$.

Entonces, a posteriori la solución de Bayes es $\hat{\mu}_B = m_x$ *con un valor de Bayes de* $V_B = c_x^2$.

De esta forma, en términos de la precisiones $\tau = \frac{1}{\sigma^2}$, $\tau_\mu = \frac{1}{c^2}$ *y* $\tau_x = \frac{1}{c_x^2}$, *y denotando*

$$\alpha \equiv \frac{\left(\frac{n}{\sigma^2}\right)}{\left(\frac{n}{\sigma^2}\right)+\left(\frac{1}{c^2}\right)} = \frac{n\tau}{n\tau+\tau_\mu},$$

resulta que las ecuaciones

$$m_x = \alpha\bar{x} + (1-\alpha)m \tag{5.1}$$

$$\tau_x = n\tau + \tau_\mu \tag{5.2}$$

definen la regla de actulización de los parámetros.

Este ejemplo es particularmente ilustrativo al observar que, el estimador puntual resulta ser una combinación lineal de la media inicial y la media muestral (ecuación 5.1), y semejantemente, la precisión a posteriori se determina mediante una combinación lineal de la precisión inicial y la precisión muestral (ecuación 5.2). Este hecho, se ilustra en la figura 5.3 donde se presenta una gráfica conocida como triplot, que incluye simultaneamente la densidad a priori, la función de verosimilitud, y la densidad a posteriori para el parámetro de interés. En este caso, para fines ilustrativos, se han utilizado los valores $m = 0$, $c^2 = 1$, $n = 6$, $\bar{x} = 2$ *y* $\sigma^2 = 3$.

El aspecto más relevante de esta gráfica es que si bien la inicial (en verde y rayada) y la verosimilitud (en azul y punteada) no son incompatibles, sí poseen información distinta sobre μ *y en esas condiciones la final (en rojo y sólida) resulta en un compromiso entre ambas y que, en particular, para este ejemplo la posterior siempre es más precisa que cualquiera de las dos componentes originales.*

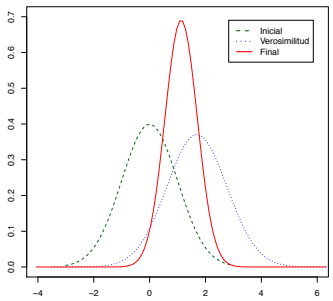


Figura 5.3: Triplot para el ejemplo 5.2.1.

Ejemplo 5.2.2. *Sea X una v.a. con distribución $Normal(x \mid \mu, \sigma^2)$ con σ^2 conocida. Se desea estimar puntualmente a μ (utilizando pérdida cuadrática), y suponiendo que a priori se describe el conocimiento sobre μ con un modelo $Uniforme(x \mid a, b)$, $a < b$.*

En este caso

$$f(\mu \mid x_{(n)}) \propto f(x_{(n)} \mid \mu) f(\mu) \propto e^{-\frac{n}{2\sigma^2}(\mu - \bar{x})^2} I_{[a,b]}(\mu),$$

lo que implica que

$$f(\mu \mid x_{(n)}) = \begin{cases} K e^{-\frac{n}{2\sigma^2}(\mu - \bar{x})^2} I_{[a,b]}(\mu) & si\ \mu \in [a,b] \\ 0 & e.o.c. \end{cases}$$

Observe que $\int_a^b e^{-\frac{n}{2\sigma^2}(\mu - \bar{x})^2} d\mu = K^{-1}$ *determina únicamente la distribución. Así,*

$$\hat{\mu}_B = \mathbb{E}\left[\mu \mid x_{(n)}\right] = \int_a^b K e^{-\frac{n}{2\sigma^2}(\mu - \bar{x})^2} d\mu$$

donde vale la pena comentar que en este ejemplo la distribución final no tiene la misma forma que la inicial, y que por tanto, para calcular $\hat{\mu}_B$ es necesario recurrir a métodos numéricos. Como se comentará más adelante, la propiedad en la que las distribuciones inicial y final comparten la forma puede resultar muy conveniente. En particular, si aquí $p(\mu)$ hubiese sido Normal, como en el ejemplo 5.2.1, $\hat{\mu}_B$ se podría determinar analíticamente.

5.3. Estimación por regiones

Otro problema de inferencia paramétrica muy común es el de estimación por regiones, o en el caso más simple, por intervalos. En una variedad de situaciones el investigador no necesita un valor estimado del parámetro de intéres, sino que

prefiere conocer una región de espacio parametral donde, con algún grado de certeza, se encuentra el valor desconocido del parámetro. Adicionalmente, este tipo de inferencia no sólo ofrece información sobre la localización de θ, sino también sobre la incertidumbre acerca de esa localización.

Sea entonces X una v.a. con f.d.p.g. $P(x \mid \theta)$, $\theta \in \Theta$, se desea estimar θ por regiones. La idea es encontrar una región $A \subseteq \Theta$ que sea "lo más pequeña posible" y que tenga "buenas posibilidades" de incluir a θ.

Así, al igual que en los casos de contraste de hipótesis y estimación puntual es posible expresar este problema en términos de uno de decisión. En este caso $D = \{d_A \mid A \subseteq \Theta\}$, donde comúnmente la región A se restringe a un tipo que permita una interpretación útil en la práctica. Una rama *genérica* de este problema se exhibe en la figura 5.4.

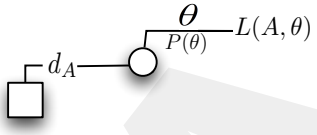


Figura 5.4: Rama típica del árbol de decisión para el problema de estimación por regiones.

En particular si $\Theta \subseteq \Re$ y las regiones son intervalos, entonces el conjunto de decisiones resulta $D = \{d_{ab} \mid d_{ab} = [a, b] \subseteq \Theta\}$. Para este caso, una posible función de pérdida es la que tiene la forma

$$L\left([a, b], \theta\right) = \alpha g(A) + (1 - \alpha) h(A, \theta)$$

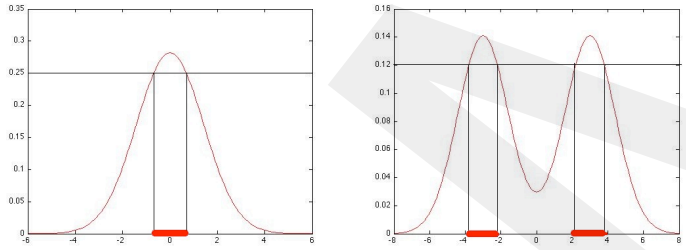
con $g(A) = b - a,\ h(A, \theta) = I_{[a,b]^c}(\theta)$ y $\alpha \in (0, 1)$. De donde resulta que

$$\begin{aligned} \mathbb{E}\left(L\left([a, b], \theta\right)\right) &= \alpha(b - a) + (1 - \alpha)\left(1 - P(\theta \in A)\right) \\ &= \alpha(b - a) + (1 - \alpha) - (1 - \alpha) P(\theta \in A) \\ &= \alpha(b - a) + (1 - \alpha)\left(F_\theta(b) - F_\theta(a)\right). \end{aligned}$$

Encontrar la solución de Bayes para este problema, en general, no es simple. No solo por la forma analítica que pueda tener F_θ y el hecho de que $a < b$, sino porque la especificación de α debe expresar las preferencias del tomador de decisiones, y al mismo tiempo juega un papel para *homogeneizar* las escalas de la longitud $(b - a)$ y la probabilidad $(F_\theta(b) - F_\theta(a))$. Una simplificación a la que se recurre con frecuencia consiste en fijar $F_\theta(b) - F_\theta(a) = \alpha$, con lo que el problema se reduce a uno sin incertidumbre. Así, fijando $P(\theta \in A)$, el problema consiste en minimizar la longitud del intervalo.

Ahora, es fácil concluir que para obtener la menor longitud de los intervalos es conveniente iniciar su construcción a partir de la imagen inversa de la(s) moda(s). De hecho, se puede probar que si se define una región $I \in \Theta$ tal que $P(\theta \in I) = 1 - \alpha$ y de manera que $P(\theta) > P(\theta')\ \forall\ \theta \in I$ y $\theta' \notin I$. Entonces, si A es cualquier otra región de Θ tal que $P(\theta \in A) = 1 - \alpha$ el área de A será al menos el de I.

Una región con estas características, se conoce como región de máxima probabilidad o máxima densidad, y si bien en muchos casos se puede calcular analíticamente, en general se determina numéricamente con métodos como la bisección. En la figura 5.5 se muestran (en rojo) dos posibles formas que puede tomar un intervalo de máxima densidad.



(a) (b)

Figura 5.5:
(a) Región de máxima densidad para una distrubución unimodal
(b) Región de máxima densidad para una distrubución multimodal

Finalmente, es relevante mencionar que para el caso multiparamétrico el procedimiento se puede desarrollar en dos diferentes formas. Si es de interes producir intervalos para cada una de las componentes del vector de parámetros, basta con obtener las marginales correspondientes

$$P(\theta \,|\, x_{(n)}) = \int_{\Phi} P(\theta, \phi \,|\, x_{(n)}) d\phi = \int_{\Phi} P(\theta \,|\, \phi, x_{(n)}) P(\phi \,|\, x_{(n)}) d\phi$$

y proceder como en el caso uniparamétrico. Si por el contrario, interesa determinar una region para un parámetro multidimensional, la idea sigue siendo la misma (fijar el nivel de probabilidad y buscar la region de mínimo volumen con dicho nivel). Sin embargo, el cálculo del volumen puede ser más complicado dependiendo de la geometría de la región.

5.4. Predicción

Es interesante comprobar que el problema de predicción prácticamente no aparece en los textos introductorios a la estadística más comunes (Frecuentistas). Lo habitual, es que este sea un tema que se explora y discute en textos más avanzados (de análisis de regresión y series de tiempo, por ejemplo). Este hecho es, en cierta medida, paradójico si se piensa que el objetivo de la estadística es describir el fenómeno de interés, en este caso describir el comportamiento de la

variable aleatoria X, y que no hay mejor manera de describir a X que siendo capaz de pronosticar los valores que ha de producir.

Aquí, no se discutirá el porqué de dicho tratamiento al tema de pronósticos en los textos Frecuentistas. En lugar de esta discusión, se tratará el problema puesto que es uno central en la inferencia.

Al igual que como se ha hecho con el resto de los problemas de inferencia, es posible expresar y resolver el problema de pronóstico como uno de decisión. Sin embargo una diferencia importante debe ser observada, en este caso es necesario considerar dos escenarios, mutuamente excluyentes, en los que se puede presentar el problema de pronóstico, ya sea puntual como por intervalos, y que conducen a dos variantes distintas del problema. Esto es, el caso en que se conoce el valor de los parámetros de la distribución de probabilidad de la variable que se desea pronosticar, y el caso en el que se desconoce al menos uno de estos parámetros.

5.4.1. Pronóstico puntual

Sea X una v.a. con f.d.p.g. $P(x \mid \theta)$, $\theta \in \Theta$, se desea pronosticar un valor $x_\star$ de una observación futura $x \in \mathscr{X}$.

Así, se trata de elegir una $\hat{x}_\star$ como anticipación del valor $x_\star$ que efectivamente producirá el fenómeno cuando sea observado. Por tanto, en este problema es posible definir $D = \{d_x \mid x \in \mathscr{X}\}$. Ahora, las consecuencias de un pronóstico particular $\hat{x}_\star$ dependen de lo bien que este reproduzca al valor futuro $x_\star$. Así, las funciones de pérdida apropiadas, como en el caso de estimación puntual, en general dependen de alguna forma de la distancia entre $\hat{x}_\star$ y $x_\star$, y son tales que entre mayor sea la distancia, asignen mayor pérdida.

- Si θ es conocido.

  En este escenario, una reprsentación gráfica del problema de resulta en la figura 5.6, donde puede observarse que este problema tiene exactamente la misma estructura que el problema de estimación puntual.

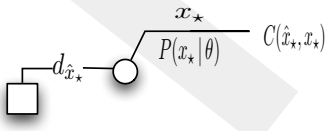


Figura 5.6: Rama típica del árbol de decisión para el problema de pronostico puntual con θ conocido.

  Y en particular, si se utiliza $L\left(C(\hat{x}_\star, x_\star)\right) = a(\hat{x}_\star - x_\star)^2$ donde a es una constante positiva, resulta que a priori la solución está determinada por

$$\hat{x}_{\star_B} = \begin{cases} \mathbb{E}_{P(x_\star \mid \theta)}(x) & \text{si } \mathbb{E}_{P(x_\star \mid \theta)}(x) \in \mathscr{X} \\ \text{el valor más cercano a } \mathbb{E}_{P(x_\star \mid \theta)}(x) & \text{e.o.c.} \end{cases}$$

con un valor de Bayes de $V_B = a\mathbb{V}(x)$

De igual manera, si $x_{(n)}$ es una m.a. de X, y $x_\star$ es condicionalmente independiente de $x_1, x_2, \ldots, x_n$ dado θ, entonces a posteriori se seguirá cumpliendo que

$$\hat{x}_{\star_B} = \underset{x}{\operatorname{argmin}}\, \mathbb{E}_{P(x_\star \,|\, \theta)}\{L(\hat{x}_\star, x_\star)\}.$$

- Si θ es desconocido

  En este caso, a diferencia del caso en el que θ es conocido, la utilidad esperada de la opción d_x no puede calcularse con respecto a $P(x \,|\, \theta)$ puesto que θ es desconocido. De hecho, ocurre que θ siendo desconocido introduce otro factor de incertidumbre, y entonces si tanto $x_\star$ como θ son desconocidos la distribución de probabilidad que debe asignar el tomador de decisiones es necesariamente de la forma de una conjunta $P(x_\star, \theta)$ y ya no la de una condicional $P(x_\star \,|\, \theta)$. De esta manera, el problema tiene asociado un árbol con una estructura como el de la figura 5.7. En esta figura, se hace evidente que existen en el problema dos fuentes de incertidumbre $x_\star$ y θ. Sin embargo, es interesante observar que la función de pérdida involucra a $x_\star$ pero no a θ.

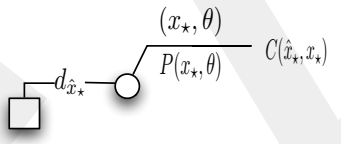


Figura 5.7: Rama típica del árbol de decisión para el problema de pronostico puntual con θ desconocido.

Así, la solución de Bayes estará dada por $\hat{x}_{\star_B} = \underset{x}{\operatorname{argmin}}\, \mathbb{E}_{P(x_\star, \theta)}\{L(\hat{x}_\star, x_\star)\}$.

En esta expresión, vale la pena recordar que la conjunta $P(x_\star, \theta)$ puede expresarse en forma alternativa a través de las dos descomposiciones $P(x_\star \,|\, \theta)P(\theta)$, y $P(\theta \,|\, x_\star)P(x_\star)$.

Volviendo al problema de optimización, observe que

$$\begin{aligned}
\mathbb{E}_{P(x_\star, \theta)}\{L(\hat{x}_\star, x_\star)\} &= \int_{\mathscr{X}} \int_{\Theta} L(\hat{x}_\star, x_\star) P(x_\star, \theta) d\theta dx_\star \\
&= \int_{\mathscr{X}} L(\hat{x}_\star, x_\star) \left\{ \int_{\Theta} P(x_\star, \theta) d\theta \right\} dx_\star \\
&= \int_{\mathscr{X}} L(\hat{x}_\star, x_\star) \left\{ \int_{\Theta} P(x_\star \,|\, \theta) P(\theta) d\theta \right\} dx_\star \\
&= \int_{\mathscr{X}} L(\hat{x}_\star, x_\star) P(x_\star) dx_\star \\
&= \mathbb{E}_{P(x_\star)}\{L(\hat{x}_\star, x_\star)\} \\
&= g(\hat{x}_\star)
\end{aligned}$$

En otras palabras, el hecho de que la función de pérdida no dependa de θ permite expresar esta pérdida esperada con una formulación alternativa en donde efectivamente el único factor de incertidumbre es $x_\star$ como en el caso en que el parámetro es conocido. La diferencia, sin embargo es que el modelo que se utiliza en aquel caso $P(x_\star \,|\, \theta)$ es ahora remplazado por $P(x_\star)$ que se relaciona con el primero a través de la expresión

$$P(x_\star) = \int_\Theta P(x_\star \,|\, \theta) P(\theta) d\theta.$$

Resulta entonces que el problema de predicción es formalmente el mismo problema que estimación puntual utilizando la distribución (que se conoce como predictiva) $P(x_\star)$. Más especificamente, esta distribución predictiva se denomina predictiva a priori si $P(\theta)$ es a su vez una distribución a priori. Por otro lado si $x_{(n)}$ es una m.a. de X y $x_\star$ es condicionalmente independiente de $x_1, x_2, \ldots, x_n$ dado θ, se tiene que

$$P(x_\star, \theta \,|\, x_{(n)}) = P(x_\star \,|\, \theta, x_{(n)}) P(\theta \,|\, x_{(n)}) = P(x_\star \,|\, \theta) P(\theta \,|\, x_{(n)}).$$

De modo que a posteriori la distribución predictiva resulta
$P(x_\star \,|\, x_{(n)}) = \int_\Theta P(x_\star \,|\, \theta) P(\theta \,|\, x_{(n)}) d\theta$.

De hecho, el problema de estimación puntual en todos los casos, incluyendo el de θ conocido, se resuelve minimizando la pérdida esperada calculada respecto a la distribución predictiva correspondiente. Basta observar que $P(x_\star \,|\, \theta)$ cuando θ es conocido, digamos $\theta = \theta_0$ equivale a utilizar una predictiva tal que $P(\theta = \theta_0) = 1$.

Ejemplo 5.4.1. *Sea X una v.a con distribución $Normal(x \,|\, \mu, \sigma^2)$, se desea pronosticar un valor $x_\star$ de una observación futura de X.*

- *Si μ, σ son conocidos.*

  $\hat{x}_{\star_B}$ es el valor que minimiza $\mathbb{E}_{N(x \,|\, \mu, \sigma^2)} \{L(\hat{x_\star}, x_\star)\}$, en particular si se utiliza pérdida cuadrática $\hat{x}_{\star_B} = \mu$.

- *Si σ es conocido y $Normal(\mu \,|\, m, c^2)$*

  A priori: $\hat{x}_{\star_B}$ es el valor que minimiza $\mathbb{E}_{P(x_\star))} \{L(\hat{x}_\star, x_\star)\}$ y tomando $L(\hat{x}_\star, x_\star) = (\hat{x}_\star - x_\star)^2$

$$\begin{aligned}
\hat{x}_{\star_B} =& \mathbb{E}_{P(x_\star)}(x_\star) = \int_{\mathscr{X}} x_\star P(x_\star) dx_\star \\
&= \int_{\mathscr{X}} x_\star \left\{ \int_\Theta P(x_\star \,|\, \mu) P(\mu) d\mu \right\} dx_\star \\
&= \int_\Theta \int_{\mathscr{X}} x_\star P(x_\star \,|\, \mu) P(\mu) dx_\star d\mu \\
&= \int_\Theta P(\mu) \left\{ \int_{\mathscr{X}} x_\star P(x_\star \,|\, \mu) dx_\star \right\} d\mu
\end{aligned}$$

$$= \int_{\Theta} P(\mu)\mathbb{E}(x_\star \,|\, \mu)d\mu = \mathbb{E}_{P(\mu)}\left\{\mathbb{E}_{P(x_\star \,|\, \mu)}(x_\star)\right\}\mathbb{E}_{P(\mu)}(\mu) = m.$$

Ahora,

$$\begin{aligned}
P(x_\star) =& \int_{\mu} (2\pi\sigma^2)^{-1/2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} (2\pi c^2)^{-1/2} e^{-\frac{(\mu-m)^2}{2c^2}} d\mu \\
=& \int_{\mu} (2\pi\sigma^2)^{-1/2} (2\pi c^2)^{-1/2} e^{-\frac{(x-\mu)^2}{2\sigma^2} - \frac{(\mu-m)^2}{2c^2}} d\mu \\
=& \frac{1}{2\pi\sigma c} \int_{\mu} e^{-\frac{(x-\mu)^2}{2\sigma^2} - \frac{(\mu-m)^2}{2c^2}} d\mu.
\end{aligned}$$

Y sean $\mu_0 = \frac{\left(\frac{x}{\sigma^2} + \frac{m}{c^2}\right)}{\left(\frac{1}{\sigma^2} + \frac{1}{c^2}\right)}$ *y* $\sigma_0^2 = \left(\frac{1}{\sigma^2} + \frac{1}{c^2}\right)^{-1}$*, analizando el exponente*

$$\begin{aligned}
exp \equiv& -\frac{1}{2}\left\{\frac{(x-\mu)^2}{\sigma^2} + \frac{(\mu-m)^2}{c^2}\right\} \\
=& -\frac{1}{2}\left\{\frac{x^2}{\sigma^2} - \frac{2x\mu}{\sigma^2} + \frac{\mu^2}{\sigma^2} + \frac{\mu^2}{c^2} - \frac{2m\mu}{c^2} + \frac{m^2}{c^2}\right\} \\
=& -\frac{1}{2}\left\{\mu^2\left(\frac{1}{\sigma^2} + \frac{1}{c^2}\right) - 2\mu\left(\frac{x}{\sigma^2} + \frac{m}{c^2}\right) + \left(\frac{x^2}{\sigma^2} + \frac{m^2}{c^2}\right)\right\} \\
=& -\frac{1}{2}\left\{\left(\frac{1}{\sigma^2} + \frac{1}{c^2}\right)\left[\mu^2 - 2\mu\frac{\left(\frac{x}{\sigma^2} + \frac{m}{c^2}\right)}{\left(\frac{1}{\sigma^2} + \frac{1}{c^2}\right)}\right] + \left(\frac{x^2}{\sigma^2} + \frac{m^2}{c^2}\right)\right\} \\
=& -\frac{1}{2}\left\{\frac{1}{\sigma_0^2}\left[\mu^2 - 2\mu\mu_0\right] + \left(\frac{x^2}{\sigma^2} + \frac{m^2}{c^2}\right)\right\} \\
=& -\frac{1}{2}\left\{\frac{1}{\sigma_0^2}\left(\mu - \mu_0\right)^2 + \left(\frac{x^2}{\sigma^2} + \frac{m^2}{c^2}\right) - \frac{1}{\sigma_0^2}\mu_0^2\right\} \\
=& -\frac{1}{2\sigma_0^2}\left(\mu - \mu_0\right)^2 - \frac{1}{2}\left\{\left(\frac{x^2}{\sigma^2} + \frac{m^2}{c^2}\right) - \frac{1}{\sigma_0^2}\mu_0^2\right\} \\
=& -\frac{1}{2\sigma_0^2}\left(\mu - \mu_0\right)^2 - \frac{\sigma_0^2}{2\sigma^2c^2}\underbrace{\left\{\frac{\sigma^2c^2}{\sigma_0^2}\left(\frac{x^2}{\sigma^2} + \frac{m^2}{c^2}\right) - \frac{\mu_0^2\sigma^2c^2}{\sigma_0^4}\right\}}_{A}.
\end{aligned}$$

Por lo que si se desarrolla A,

$$\begin{aligned}
A =& \frac{\sigma^2c^2}{\sigma_0^2}\left(\frac{x^2}{\sigma^2} + \frac{m^2}{c^2}\right) - \frac{\left(\sigma_0^2\left(\frac{x}{\sigma^2} + \frac{m}{c^2}\right)\right)^2\sigma^2c^2}{\sigma_0^4} \\
=& \frac{1}{\sigma_0^2}\left(c^2x^2 + \sigma^2m^2\right) - \left(\frac{x}{\sigma^2} + \frac{m}{c^2}\right)^2\sigma^2c^2 \\
=& \left(\frac{1}{\sigma^2} + \frac{1}{c^2}\right)\left(c^2x^2 + \sigma^2m^2\right) - \left(\frac{x}{\sigma^2} + \frac{m}{c^2}\right)^2\sigma^2c^2 \\
=& \frac{c^2x^2}{\sigma^2} + x^2 + m^2 + \frac{\sigma^2m^2}{c^2} - \left(\frac{x^2}{\sigma^4} + \frac{xm}{\sigma^2c^2} + \frac{m^2}{c^4}\right)\sigma^2c^2
\end{aligned}$$

$$
\begin{aligned}
=&\frac{c^2x^2}{\sigma^2} + x^2 + m^2 + \frac{\sigma^2 m^2}{c^2} - \frac{x^2c^2}{\sigma^2} - 2xm - \frac{m^2\sigma^2}{c^2} \\
=&x^2 - 2xm + m^2 = (x-m)^2,
\end{aligned}
$$

se obtiene que

$$
\begin{aligned}
exp =& -\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 - \frac{\sigma_0^2}{2\sigma^2c^2}(x-m)^2 \\
=& -\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 - \frac{1}{2\frac{\sigma^2c^2}{\sigma_0^2}}(x-m)^2.
\end{aligned}
$$

Así,

$$
\begin{aligned}
P(x_\star) &= \frac{1}{2\pi\sigma c}\int_\mu e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 - \frac{1}{2\frac{\sigma^2c^2}{\sigma_0^2}}(x-m)^2} d\mu \\
&= \frac{1}{2\pi\sigma c}\frac{\left(2\pi\sigma_0^2\right)^{-1/2}}{\left(2\pi\sigma_0^2\right)^{-1/2}}\int_\mu e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2} e^{-\frac{1}{2\frac{\sigma^2c^2}{\sigma_0^2}}(x-m)^2} d\mu \\
&= \left(2\pi\frac{\sigma^2c^2}{\sigma_0^2}\right)^{-1/2} e^{-\frac{1}{2\frac{\sigma^2c^2}{\sigma_0^2}}(x-m)^2} \underbrace{\int_\mu \left(2\pi\sigma_0^2\right)^{-1/2} e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2} d\mu}_{=1 \quad \left(Normal(\mu\,|\,\mu_0,\sigma_0^2)\right)} \\
&= \left(2\pi\frac{\sigma^2c^2}{\sigma_0^2}\right)^{-1/2} e^{-\frac{1}{2\frac{\sigma^2c^2}{\sigma_0^2}}(x-m)^2} \\
&= \left(2\pi\left(\sigma^2+c^2\right)\right)^{-1/2} e^{-\frac{1}{2(\sigma^2+c^2)}(x-m)^2}.
\end{aligned}
$$

Por lo tanto, a priori, la distribución predictiva es

$$P(x_\star) = Normal\left(x_\star\,|\,m, \sigma^2+c^2\right).$$

Ahora, a posteriori se sabe que $P(\mu\,|\,x_{(n)}) = Normal(\mu\,|\,m_x, c_x^2)$ *donde*

$$m_x = \frac{\left(\frac{n}{\sigma^2}\right)\bar{x} + \left(\frac{1}{c^2}\right)m}{\left(\frac{n}{\sigma^2}+\frac{1}{c^2}\right)} \quad y \quad c_x^2 = \left(\frac{n}{\sigma^2}+\frac{1}{c^2}\right)^{-1},$$

por lo que, procediendo igual que antes, se obtiene que

$$
\begin{aligned}
P(x_\star\,|\,x_{(n)}) &= \int_\mu P(x_\star\,|\,\mu)P(\mu\,|\,x_{(n)})d\mu \\
&= Normal(x_\star\,|\,m_x, \sigma^2+c_x^2).
\end{aligned}
$$

Entonces, bajo pérdida cuadrática, resulta que

$$\hat{x}_{\star_B} = \mathbb{E}_{P(\mu\,|\,x_{(n)})}(\mu) = m_x,$$

y recordando que $m_x = \alpha\bar{x} + (1-\alpha)m$ *se puede observar que, en general, este pronóstico no coincide con el estimador frecuentista habitual* $\bar{x}$.

5.4.2. Pronóstico por regiones

Sea X una v.a. con f.d.p.g. $P(x \mid \theta)$, $\theta \in \Theta$, se desea pronosticar por regiones un valor $x_\star$ de una observación futura de $x \in \mathscr{X}$.

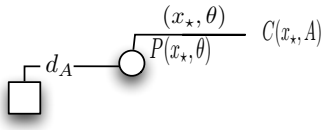


Figura 5.8: Rama típica del árbol de decisión para el problema de pronostico por regiones.

De la misma forma que ocurre con el pronóstico puntual, el problema de pronóstico por regiones resulta ser completamente análogo a su contraparte de estimación por regiones. Aquí, el espacio de opciones es $D = \{d_A \mid A \subseteq \mathscr{X}\}$, con $\mathscr{X}$ el soporte de X, y la distribución de probabilidades relevante es la predictiva para x, sea con θ desconocida (a priori o a posteriori) o con θ conocida.

5.5. Ejercicios

Ejercicio 5.1. *Sea X una variable aleatoria Normal con media μ y varianza $\sigma^2 = 1$. Si se cuenta con una muestra aleatoria de tamaño 10 de X, tal que su media muestral es 0,35 y resulta que es de interés contrastar las hipótesis*

$$H_0 : \mu = 0 \quad vs. \quad H_1 : \mu = 1,$$

compare las decisiones a las que se llega se se utiliza por una parte, el procedimiento frecuentista usual con $\alpha = 0{,}05$ y, por otra parte, el procedimiento Bayesiano cuando no se tiene información sobre la validez de las hipótesis y el error de tipo I se considera 5 veces más grave que el error de tipo II.

¿Qué ocurre si la muestra, con el mismo valor de la media muestral, proviniese de una muestra de tamaño 30 en lugar de 10?

Comente las diferencias que se observan en los cuatro casos considerados y explique las causas de este comportamiento.

Ejercicio 5.2. *Sea X la variable aleatoria que describe el tiempo que hay que esperar a un autobús en cierta parada un día determinado de la semana, a una hora particular. Suponga que X sigue una distribución Uniforme en $(0, \theta)$ y que se desea contrastar las hipótesis*

$$H_0 : 5 \leq \theta \leq 15 \quad vs. \quad H_1 : \theta > 15,$$

considerando los siguientes elementos:

- *A priori, el conocimiento sobre θ se describe razonablemente con una distribución Pareto con parámetros 5 y 3. Es decir, inicialmente se tiene que $p(\theta) = 3(5)^3\theta^{-4}$ para $\theta > 5$.*

- *Se han observado cinco tiempos de espera con los siguientes valores:* $x_1 = 10$, $x_2 = 3$, $x_3 = 2$, $x_4 = 5$ *y* $x_5 = 14$.

Calcule la probabilidad a priori de cada hipótesis, la probabilidad a posteriori de cada hipótesis y sugiera la manera de tomar la decisión en este problema.

Ejercicio 5.3. *Imagine que cuenta con una observación* x *de una variable aleatoria* X *con distribución Bernoulli con parámetro desconocido* θ *que toma valores en el intervalo* $(0, 1)$. *Si a priori se asigna a* θ *la distribución* $\pi(\theta)$, *demuestre que a posteriori el estimador de Bayes con pérdida cuadrática de* θ *satisface la ecuación*

$$\hat{\theta}_B = \mathbb{E}_\pi(\theta) + \frac{\mathbb{V}_\pi(\theta)}{[\mathbb{E}_\pi(\theta)]^x [\mathbb{E}_\pi(\theta) - 1]^{1-x}}$$

donde $\mathbb{E}_\pi$ *y* $\mathbb{V}_\pi$ *representan respectivamente la esperanza y varianza inicial. ¿Será cierto que la esperanza final es siempre mayor a la esperanza inicial?*

Ejercicio 5.4. *Sea* X *una variable aleatoria con f.d.p.g.* $f(x \mid \theta)$ *con* $\theta \in \Theta \subseteq \Re$ *y suponga que inicialmente el conocimiento sobre* θ *se describe con la distribución* $p(\theta)$. *Sea* $L(d, \theta) = |d - \theta|$ *la función de pérdida asociada a la decisión de estimar* θ *con el valor estimado* d. *Demuestre que la solución de Bayes es una mediana de la distribución* $p(\theta)$.

Ejercicio 5.5. *En el mismo contexto del ejercicio 5.4, considere ahora la función de pérdida*

$$L(d, \theta) = \begin{cases} a(\theta - d) & si\ \theta - d \geq 0 \\ b(d - \theta) & e.o.c \end{cases}$$

donde a *y* b *son constantes positivas. Demuestre que la solución de Bayes al problema de estimación de* θ *está dada por cualquier cuantil de orden* $a/(a + b)$ *de la distribución* $p(\theta)$.

Ejercicio 5.6. *Suponga que desea estimar el parámetro* θ *con un estimador* d *y que la función de pérdida que desea utilizar es la siguiente:*

$$L(d, \theta) = \left(\frac{d - \theta}{d} \right)^2$$

Encuentre la solución de Bayes para este problema e identifique todos los supuestos que sean necesarios para garantizar que exista solución.

Ejercicio 5.7. *Sea* $x_1, x_2, \ldots, x_n$ *una muestra aleatoria de una variable* X *con distribución Uniforme en el intervalo* $(0, \theta)$ *donde* $\theta > 0$. *Si a priori se considera que el conocimiento sobre el parámetro* θ *está adecuadamente descrito con una distribución Uniforme en el intervalo* $(0, c)$ *con* $c > 0$ *una constante conocida,*

a) Plantee el problema de estimar θ *como uno de decisión.*

b) ¿Cuál es la solución al problema planteado en a) sin incorporar la muestra y utilizando la función de pérdida cuadrática?

c) De la misma forma que en el inciso b), es decir, sin datos, ¿Cuál es la solución a este problema si se utiliza $L(d,\theta) = |d-\theta|$*?*

d) Si se incorpora la muestra, ¿Cuál es la distribución final de θ*?*

e) Considerando los datos, ¿Cuál es la solución utilizando $L(d,\theta) = (d-\theta)^2$*?*

f) ¿Cuál es la solución, con datos, utilizando la función de pérdida absoluta?

g) Considerando los datos, encuentre es la solución si

$$L(d,\theta) = \begin{cases} 0 & si\ |d-\theta| < 0{,}001 \\ 1 & e.o.c \end{cases}$$

Ejercicio 5.8. *Sea* X *una variable aleatoria con f.d.p.g.* $f(x\,|\,\theta)$ *con* $\theta \in \Theta \subseteq \Re$*. Si cuenta con una muestra aleatoria de tamaño* n *de* X *y una distribución inicial* $P(\theta)$ *para el valor desconocido del parámetro, ¿Cuál es la solución de Bayes al problema de estimar puntualmente* θ *si, para un valor* ϵ *cercano de cero, la función de pérdida está definida por:*

$$L(d,\theta) = \begin{cases} 1 & si\ |d-\theta| > \epsilon \\ 0 & e.o.c \end{cases}$$

Ejercicio 5.9. *Considere un problema de estimación en el que* $\Theta = D = (0,1)$ *y se utiliza una función de pérdida de la forma* $L(d,\theta) = 100\,(d-\theta)^2$*. Suponga que la distribución inicial sobre* θ *está dada por* $p(\theta) = 2\theta$ *para* $\theta \in \Theta$*. Demuestre que el valor* $d = 2/3$ *es la solución de Bayes a priori, y el valor de Bayes es* $50/9$*.*

Ejercicio 5.10. *En el contexto del problema 5.9, suponga que un estadístico A piensa que la distribución de probabilidad de* θ *es en efecto la descrita en ese problema, mientras que el estadístico B cree que la distribución de probabilidad de* θ *es* $P_B(\theta) = 3\theta^2$ *para* θ *en* $(0,1)$*. ¿En que magnitud cree B que A incrementará su pérdida esperada debido a su conocimiento* incorrecto *de la distribución de probabilidades de* θ*?*

Ejercicio 5.11. *Se ha propuesto un procedimiento para clasificar la sangre de cada individuo distinguiendo entre tipos O, A, B o AB. El procedimiento consiste en extraer una muestra de sangre del individuo en cuestión y medir la cantidad* X *de una cierta sustancia. Se sabe que para cada individuo,* X *sigue una distribución determinada por la densidad*

$$f(x\,|\,\theta) = exp\{-(x-\theta)\}\mathbb{1}_{(0,\infty)}(x)$$

donde θ *es un parámetro que, en cada individuo, determina el tipo de sangre; de hecho, si* $0 < \theta \leq 1$*, la sangre es de tipo AB; si* $1 < \theta \leq 2$*, la sangre es de*

tipo A; si por otra parte, $2 < \theta \leq 3$*, la sangre es de tipo B, y si* $\theta > 3$ *la sangre es de tipo O.*

Ahora bien, en la población de interés el valor del parámetro θ *cambia de individuo a individuo, pero se sabe que su distribución está determinada por la densidad*

$$p(\theta) = exp\{-\theta\}\mathbb{1}_{(0,\infty)}(\theta).$$

¿En qué grupo sanguíneo clasificaría a un individuo particular para el cual se ha observado que $X = 4$ *si la pérdida por clasificación incorrecta viene dada por la siguiente tabla?*

	Clasificación			
Tipo real	AB	A	B	O
AB	0	1	1	2
A	1	0	2	2
B	1	2	0	2
O	3	3	3	0

Ejercicio 5.12. *El número de incendios que se producen semanalmente en una ciudad* X*, sigue una distribución Poisson con media* θ*. Se desea construir el intervalo de máxima densidad de probabilidad a posteriori para* θ*. Puesto que inicialmente no se conoce nada sobre* θ*, parece adecuado utilizar la función* $\pi(\theta) = \theta^{-1}\mathbb{1}_{(0,\infty)}(\theta)$ *para describir esta falta de información. Observe que* $\pi(\theta)$ *no es propiamente una función de distribución (pues no integra a uno), estas funciones se conocen con el nombre de distribuciónes impropias y se discutirán en el siguiente capítulo. Por lo pronto, si durante cinco semanas se observaron:* $x_1 = 0$, $x_2 = 1$, $x_3 = 0$, $x_4 = 0$ *y* $x_5 = 0$ *fuegos respectivamente, ¿Cuál es el intervalo de máxima densidad a posteriori para* θ *con probabilidad* 0,9*?*

Ejercicio 5.13. *Suponga que* X *una variable aleatoria Bernoulli con parámetro* θ *en el intervalo* $(0, 1)$ *y el conocimiento sobre* θ *se describe con la distribución (inicial o final)* $P(\theta)$*. ¿Cuál es la distribución predictiva para una observación futura de* X*?*

Ejercicio 5.14. *Suponga que* X *una variable aleatoria Normal con media* μ *y precisión conocida. Si el conocimiento sobre* μ *se describe con una Normal de media* m *y precisión* τ_μ *¿Cuál es la distribución predictiva para una observación futura de* X*?*

Ejercicio 5.15. *Suponga que* X *una variable aleatoria Normal con media conocida y precisión* τ*. Si el conocimiento sobre* τ *se describe con una Gamma*(α, β) *¿Cuál es la distribución predictiva para una observación futura de* X*?*

Capítulo 6

Inferencia Paramétrica Bayesiana

Una vez que han sido establecidos los elementos generales de la Teoría de Decisión, y que los problemas típicos de la Inferencia Paramétrica, al menos en sus versiones más simples, han sido identificados como casos particulares de problemas de decisión en ambiente de incertidumbre, es conveniente volver al tema de Inferencia Paramétrica en general para establecer sus características, especialmente, cuando se aborda desde la perspectiva Bayesiana.

6.1. Principio de verosimilitud

Habitualmente, un problema de Inferencia Estadística Paramétrica se presenta cuando se cuenta con una muestra aleatoria $x_{(n)}$ de una variable aleatoria X, cuya f.d.p.g. $f(x \mid \theta)$ es totalmente conocida excepto por el valor del parámero θ, que es un elemento del espacio paramétrico $\Theta \subseteq \Re^k$. Así, el problema general de la inferencia consiste en utilizar la información disponible para describir, así sea aproximada, el comportamiento de la variable X.

Por supuesto, si el valor del parámetro θ fuese conocido, el modelo de probabilidad $f(x \mid \theta)$ sería, a su vez, totalmente conocido y la descripción de X sería completa. Más aún, los problemas específicos de producción de pronósticos, al plantearse como problemas de decisión, única posibilidad bajo el enfoque Bayesiano, se habrían de resolver utilizando $f(x \mid \theta)$ como modelo predictivo para X.

En el caso más común, en que θ es desconocido, y sea cual sea la manera en la que se pretende describir a X, este desconocimiento representa una fuente de incertidumbre que debe considerarse al producir la inferencia de interés. En el lenguaje Bayesiano, es una fuente de incertidumbre cuyo efecto en el proceso de

toma de decisiones debe tomarse en cuenta.

En el caso particular de pronósticos, la única forma en que esta incertidumbre puede tomarse en cuenta es utilizando $f(x) = \int f(x \mid \theta)P(\theta)d\theta$ como modelo predictivo para X. En esta expresión $P(\theta)$ representa el modelo de probabilidad que describe la incertidumbre del investigador sobre el valor de θ y, respecto a la muestra $x_{(n)}$, puede ser a priori o a posteriori en cuyo caso la notación apropiada es $f(x \mid x_{(n)}) = \int f(x \mid \theta)P(\theta \mid x_{(n)})d\theta$, que presupone que x y $x_{(n)}$ son condicionalmente independientes dado θ. De cualquier manera, entonces, el impacto de la información muestral $x_{(n)}$ en la producción de pronósticos, se produce a través del efecto que $x_{(n)}$ tenga en la transformación de la inicial $P(\theta)$ en la final $P(\theta \mid x_{(n)})$.

Cundo la descripción de X se refiere al análisis de alguno de sus atributos (momentos, cuantiles, moda, probabilidades específicas, etc.), sólo puede ocurrir que el atributo de interés sea independiente del valor de θ, en cuyo caso el problema es análogo al que se enfrenta cuando el parámetro es conocido, o bien que el atributo, por ejemplo η, sea función de θ, en cuyo caso el valor de $\eta = \eta(\theta)$ es desconocido y constituye la fuente de incertidumbre relevante. El punto aquí es que η es incierto porque θ es incierto, y que la distribución $P(\eta)$ que describe el estado de conocimiento del investigador sobre el atributo η puede, en general, derivarse de la distribución $P(\theta)$. Y entonces, de nuevo, si la asignación de este modelo se produce antes de contar con la muestra $x_{(n)}$, se cuenta con la a priori $P(\theta)$ (y la correspondiente a priori $P(\eta)$), mientras que si ya se observó $x_{(n)}$ se utiliza la posteriori $P(\theta \mid x_{(n)})$ (y su respectiva posterior $P(\eta \mid x_{(n)})$).

De esta manera, en todos los casos de la Inferencia Paramétrica, el efecto de la muestra $x_{(n)}$ en el proceso inferencial se reduce al impacto que esta tiene en la transformación de $P(\theta)$ en $P(\theta \mid x_{(n)})$.

En estas condiciones, es de interés fundamental el estudio del mecanismo a través del cual la inicial se combina con la información muestral, para dar origen a la final. Es decir, resulta del mayor interés comprender cómo operan los elementos de la, aparentemente simple, formula de Bayes:

$$P(\theta \mid x_{(n)}) \boldsymbol{\alpha} P(x_{(n)} \mid \theta)P(\theta)$$

que establece que la final es simplemente el producto de la inicial $P(\theta)$ por la verosimilitud $P(x_{(n)} \mid \theta)$.

Un primer resultado que, aun siendo evidente, es frecuentemente ignorado por los procedimientos estadísticos habituales (no Bayesianos) es el siguiente: Si $x_{(n)}$ influye en las inferencias únicamente a través de su impacto en la transformación de la inicial en la final, y en ese mecanismo $x_{(n)}$ sólo participa mediante la función de verosimilitud, entonces, dada una inicial $P(\theta)$, dos muestras distintas que produzcan la misma verosimilitud (como función de θ) deben dar lugar a las mismas inferencias. Esta idea es tan importante que en la literatura Bayesiana ha alcanzado el rango de *principio.*

Principio de verosimilitud: Considere dos colecciones de variables aleatorias $x_{(n)} = (x_1, x_2, \ldots, x_n)$ y $y_{(m)} = (y_1, y_2, \ldots, y_m)$ con f.d.p.g. conjuntas dadas por $f(x_{(n)} \,|\, \theta)$ y $g(y_{(m)} \,|\, \theta)$, donde θ es el mismo parámetro en ambos modelos. Si como función de θ ocurre que $f(x_{(n)} \,|\, \theta) = k\, g(y_{(m)} \,|\, \theta)$ con k una constante (respectiva a θ), entonces, para una inicial común $P(\theta)$, las distribuciones finales $P(\theta \,|\, x_{(n)})$ y $P(\theta \,|\, y_{(m)})$ coinciden y, por tanto, dan lugar a las mismas inferencias.

Ejemplo 6.1.1. *Sean X e Y dos v.a. con distribución $Binomial(x \,|\, \theta, 10)$ y $BinomialNegativa(y \,|\, \theta, 4)$ respectivamente, y $x_{(1)}$ e $y_{(1)}$ dos m.a. de tamaño uno tales que $x = 4$ e $y = 10$, entonces:*

$$L(\theta \,|\, x = 4) = \binom{10}{4}\theta^4(1-\theta)^6 \quad y \quad L(\theta \,|\, y = 10) = \binom{9}{3}\theta^4(1-\theta)^6.$$

Observe, por tanto, que

$$L(\theta \,|\, x = 4) = \binom{10}{4}\binom{9}{3}^{-1} L(\theta \,|\, y = 10) = C\, L(\theta \,|\, y = 10),$$

es decir,

$$L(\theta \,|\, x = 4)\, \boldsymbol{\alpha}\, L(\theta \,|\, y = 10) \implies L(\theta \,|\, x = 4)P(\theta)\, \boldsymbol{\alpha}\, L(\theta \,|\, y = 10)P(\theta).$$

Lo que implica que

$$P(\theta \,|\, x = 4)\, \boldsymbol{\alpha}\, P(\theta \,|\, y = 10)$$

o bien, que

$$P(\theta \,|\, x = 4) = K P(\theta \,|\, y = 10) \quad K \in \Re.$$

Por otro lado,

$$\int_\theta P(\theta \,|\, x = 4)d\theta = 1 \quad y \quad \int_\theta P(\theta \,|\, y = 10)d\theta = 1,$$

por lo tanto

$$K = 1 \implies P(\theta \,|\, x = 4) = P(\theta \,|\, y = 10).$$

En otras palabras, para efectos de un análisis Bayesiano, es lo mismo observar diez lanzamientos (número fijo) de una moneda habiendo ocurrido cuatro éxitos, que haber fijado de antemano el número de éxitos (4) y que el cuarto éxito ocurra, precisamente, en el décimo lanzamiento. Así, las inferencias sobre la probabilidad de éxito θ son la mismas con una muestra o la otra. Más aún, la estimación puntual óptima para ambos casos debe coincidir.

Este es un ejemplo particularmente interesante si se recuerda que, por ejemplo, el valor del estimador insesgado (frecuentista) para θ en el modelo Binomial negativo, no coincide, en general, con el estimador insesgado para θ en el modelo Binomial.

Ejemplo 6.1.2. *Sean X e Y dos v.a. con distribución $Poisson(x \mid \lambda)$ y distribución $Exponencial(y \mid \lambda)$ respectivamente, y $x_{(r)}$ e $y_{(s)}$ dos m.a. de tamaño r y s respectivamente, entonces:*

$$P(x_{(r)} \mid \lambda) = \prod_{i=1}^{r} \left[\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right] = \frac{\lambda^{\sum x_i} e^{-r\lambda}}{\prod x_i!} = L(\lambda \mid x_{(r)}) \quad y$$

$$P(y_{(s)} \mid \lambda) = \prod_{i=1}^{s} \left[\lambda e^{-y\lambda} \right] = \lambda^s e^{-\lambda \sum y_i} = L(\lambda \mid y_{(s)}).$$

Por tanto, cuando $\sum_{i=1}^{r} x_i = s$ y $\sum_{i=1}^{s} y_i = r$ resulta que $L(\lambda \mid x_{(r)}) \boldsymbol{\alpha} L(\lambda \mid y_{(s)})$. De manera que, de nuevo, si se parte de la misma inicial $P(\lambda)$, entonces necesariamente se tiene que $P(\lambda \mid x_{(r)}) = P(\lambda \mid y_{(s)})$.

Una vez establecido el principio de verosimilitud, existe otro rasgo general en el proceso de aprendizaje que es particularmente interesante desde el punto de vista conceptual. El concepto de suficiencia representa, posiblemente, el único concepto donde los enfoques frecuentista y Bayesiano coinciden plenamente. Por otra parte, en el ámbito exclusivamente Bayesiano, posibilita el desarrollo de un mecanismo para la asignación de distribuciones iniciales que han probado ser muy convenientes en la práctica.

6.2. Suficiencia

Antes de introducir formalmente el concepto de suficiencia Bayesiana, es conveniente recordar la definición de estadística.

Definición 6.2.1. *Se dice que $T_n : \mathfrak{X}_{(\mathbf{n})} \longrightarrow \Re^{\kappa(n)}$ es una* **estadística** *si es una v.a. que es función de la muestra y no involucra en su expresión ningún parámetro desconocido. Se dice además que $T_n\left(x_{(n)}\right)$ es de dimensión fija si $\kappa(n) = k \, \forall n$.*

Ahora sí, una clase de estadísticas especialmente importantes son las que, en un sentido Bayesiano, resultan suficientes para un parámetro θ.

Definición 6.2.2. *Sea $x_{(n)}$ una m.a. de una v.a. X con f.d.p.g. $P(x \mid \theta)$, $\theta \in \Theta$ y sea $T_n\left(x_{(n)}\right)$ una estadística de los datos, se dice que $T_n\left(x_{(n)}\right)$ es* **suficiente** *(desde el punto de vista Bayesiano) para $\theta \Leftrightarrow P(\theta \mid x_{(n)})$ depende de $x_{(n)}$ sólo a través de $T_n\left(x_{(n)}\right)$ $\forall n$ y $\forall P(\theta)$.*

Ejemplo 6.2.1. *Sean X una variable aleatoria con distribución $Bernoulli(x \mid \theta)$ y $x_{(n)}$ una m.a. de X. Esto es*

$$P(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i} \; \forall \; i \in \{1, 2, \ldots, n\}.$$

Así, para cualquier distribución inicial $P(\theta)$, resulta que

$$P(\theta \,|\, x_{(n)}) \,\boldsymbol{\alpha}\, P(x_{(n)} \,|\, \theta)\, P(\theta) = \left[\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}\right] P(\theta)$$
$$= \left[\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\right] P(\theta),$$

por lo que $T_n(x_{(n)}) = \sum_{i=1}^{n} x_i$ *es una estadística suficiente para* θ.

Recuerde que, de acuerdo con la definición tradicional, $T_n\left(x_{(n)}\right)$ es una estadística suficiente (desde el punto de vista frecuentista) para θ si y sólo si $P\left(x_{(n)} \,|\, T_n\left(x_{(n)}\right)\right)$ no depende de θ. Adicionalmente, que una estadística suficiente (desde el punto de vista frecuentista) puede ser caracterizada de acuerdo con el criterio de factorización de Fisher-Neyman.

Teorema 6.2.1. *(Factorización de Fisher-Neyman).*

Sea $x_{(n)}$ una m.a. de una v.a. X con f.d.p.g. $P(x \,|\, \theta)$, y $T_n(x_{(n)})$ una estadística, entonces T_n es suficiente (desde el punto de vista frecuentista) si y sólo si existen funciones $h(x_{(n)})$ y $g(\theta, T_n(x_{(n)}))$, donde h no depende de θ y g depende de la muestra sólo a través de T_n, tales que $P(x_{(n)} \,|\, \theta) = h(x_{(n)})\, g(\theta, T_n(x_{(n)}))$.

Observe que el teorema 6.2.1 es ampliamente general en el sentido en que, si $T_n(x_{(n)})$ es una estadística suficiente, no impone una forma específica sobre la función g. De hecho, cualquier función de θ y $T_n(x_{(n)})$ que dependa de la muestra sólo a través de T_n puede jugar el papel de g, en cuyo caso $h(x_{(n)})$ será la correspondiente constante de normalización para $P(x_{(n)} \,|\, \theta)$. Así, puesto que

$$P(x_{(n)} \,|\, \theta) = P(x_{(n)}, T_n(x_{(n)}) \,|\, \theta) = P(x_{(n)} \,|\, T_n(x_{(n)}), \theta) P(T_n(x_{(n)}) \,|\, \theta),$$

y dado que T_n es una estadística suficiente (desde el punto de vista frecuentista), $P(x_{(n)} \,|\, T_n(x_{(n)}), \theta) = P(x_{(n)} \,|\, T_n(x_{(n)}))$, de forma que

$$P(x_{(n)} \,|\, \theta) = P(x_{(n)} \,|\, T_n(x_{(n)})) P(T_n(x_{(n)}) \,|\, \theta),$$

pero de nuevo, como T_n es suficiente (desde el punto de vista frecuentista), $P(x_{(n)} \,|\, T_n(x_{(n)}))$ no depende de θ. De manera que una posibilidad particular es tomar $g(\theta, T_n(x_{(n)}))$ como la f.d.p.g. $P(T_n(x_{(n)}) \,|\, \theta)$, caso en el que la función $h(x_{(n)})$ resulta ser la condicional $P(x_{(n)} \,|\, T_n(x_{(n)}))$.

Como se comentó previamente, el concepto de suficiencia es quizá el único que transita libremente entre los dos enfoques, Bayesiano y frecuentista. Y la equivalencia queda establecida a través del siguiente teorema.

Teorema 6.2.2. *Sea $x_{(n)}$ una m.a. de una v.a. X, discreta o continua, con f.d.p.g. $P(x \,|\, \theta)$, $\theta \in \Theta$, entonces:*
$T_n\left(x_{(n)}\right)$ es suficiente Bayesiana $\iff$ $T_n\left(x_{(n)}\right)$ es suficiente frecuentista.

Demostración.

$\underline{\Rightarrow}$ Sea $T_n(x_{(n)})$ una estadística suficiente (desde el punto de vista Bayesiano). Es decir,

$$P(\theta \,|\, x_{(n)}) = P(\theta \,|\, T_n(x_{(n)})). \tag{6.1}$$

Ahora, por el teorema de Bayes

$$P(x_{(n)} \,|\, \theta) = \frac{P(\theta \,|\, x_{(n)})P(x_{(n)})}{P(\theta)} = \frac{P(\theta \,|\, T_n(x_{(n)}))P(x_{(n)})}{P(\theta)},$$

en donde la última igualdad se obtiene sustituyendo 6.1. Así, utilizando nuevamente el teorema de Bayes, se tiene que

$$P(x_{(n)} \,|\, \theta) = \frac{P(T_n(x_{(n)}) \,|\, \theta)P(\theta)P(x_{(n)})}{P(T_n(x_{(n)}))P(\theta)} = \underbrace{P(T_n(x_{(n)}) \,|\, \theta)}_{=g(\theta, T_n(x_{(n)}))} \underbrace{\left[\frac{P(x_{(n)})}{P(T_n(x_{(n)}))}\right]}_{=h(x_{(n)})},$$

donde $h(x_{(n)})$ no depende de θ, y $g(\theta, T_n(x_{(n)}))$ depende de $x_{(n)}$ sólo a través de T_n. Por tanto, el teorema de factorización de Fisher-Neyman implica que $T_n(x_{(n)})$ es una estadística suficiente (en el sentido frecuentista).

$\underline{\Leftarrow}$ Sea $T_n(x_{(n)})$ una estadística suficiente (desde el punto de vista frecuentista). Entonces, por el criterio de factorización de Fisher-Neyman,

$$P(x_{(n)} \,|\, \theta) = P(T_n(x_{(n)}) \,|\, \theta)h(x_{(n)}). \tag{6.2}$$

Ahora, por el teorema de Bayes

$$P(\theta \,|\, x_{(n)}) = \frac{P(x_{(n)} \,|\, \theta)P(\theta)}{P(x_{(n)})},$$

en donde sustituyendo 6.2 resulta que

$$\begin{aligned} P(\theta \,|\, x_{(n)}) &= \frac{P(T_n(x_{(n)}) \,|\, \theta)h(x_{(n)})P(\theta)}{P(x_{(n)})} \\ &= \frac{P(T_n(x_{(n)}) \,|\, \theta)P(\theta)}{P(T_n(x_{(n)}))} \left[\frac{h(x_{(n)})P(T_n(x_{(n)}))}{P(x_{(n)})}\right] \\ &= P(\theta \,|\, T_n(x_{(n)})) \left[\frac{h(x_{(n)})P(T_n(x_{(n)}))}{P(x_{(n)})}\right]. \end{aligned}$$

Y por tanto, integrando con respecto a θ

$$\begin{aligned} 1 = \int P(\theta \,|\, x_{(n)})\, d\theta &= \frac{h(x_{(n)})P(T_n(x_{(n)}))}{P(x_{(n)})} \underbrace{\int P(\theta \,|\, T_n(x_{(n)}))\, d\theta}_{=1} \\ &= \frac{h(x_{(n)})P(T_n(x_{(n)}))}{P(x_{(n)})}, \end{aligned}$$

de donde se sigue que

$$P(\theta \,|\, x_{(n)}) = P(\theta \,|\, T_n(x_{(n)}))$$

o equivalentemente, que T_n es una estadística suficiente (desde el punto de vista Bayesiano). □

El teorema 6.2.2, además de establecer una equivalencia interesante, resulta ser muy útil en la práctica, pues permite utilizar el teorema de factorización de Fisher-Neyman en la identificación de estadísticas suficientes bajo un contexto Bayesiano.

Ejemplo 6.2.2. *Sean X una variable aleatoria con distribución Poisson$(x \,|\, \lambda)$ y $x_{(n)}$ una m.a. de X. Esto es*

$$P(x_i \,|\, \lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \;\forall\, i \in \{1, 2, \ldots, n\},$$

de forma que la verosimilitud resulta

$$P(x_{(n)} \,|\, \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \;\forall\, i \in \{1, 2, \ldots, n\} = \left[\frac{1}{x_1!\, x_2! \ldots x_n!}\right] \left[e^{-n\lambda}\lambda^{\sum x_i}\right]$$

y, por el criterio de factorización, es fácil ver que $T_n(x_{(n)}) = \sum_{i=1}^{n} x_i$ es una estadística suficiente para λ.

6.3. Familias conjugadas

Cualquiera que sea el problema específico de inferencia paramétrica que se pretenda resolver, desde la perspectiva Bayesiana la distribución que describe la incertidumbre del investigador sobre el parámetro desconocido juega un papel central. Especialmente cuando existe una muestra aleatoria, caso en el que el análisis Bayesiano hace uso de la correspondiente distribución a posteriori $P(\theta \,|\, x_{(n)}) \boldsymbol{\alpha} P(x_{(n)} \,|\, \theta)\, P(\theta)$ para producir las inferencias óptimas (de pérdida esperada mínima).

Precisamente en ese sentido, y en tanto que $P(x_{(n)} \,|\, \theta)$ y $P(\theta)$ son modelos que, en general, describen aproximadamente el proceso de muestreo y el conocimiento previo sobre θ, es interesante observar que la representación de estos dos elementos no necesariamente es única, y que algunas selecciones pueden resultar más ventajosas que otras en términos de los cálculos involucrados. En particular, si el modelo de muestreo se considera fijo y la inicial $P(\theta)$ se puede elegir de manera que, además de describir razonablemente el conocimiento a priori sobre θ, permita que la final $P(\theta \,|\, x_{(n)})$ resulte en un modelo que conduzca a cálculos simples. Esta idea ha dado lugar a, entre otras, la siguiente propuesta.

Definición 6.3.1. *Sea X una v.a con f.d.p.g $P(x\,|\,\theta)$ $\theta \in \Theta$, entonces la familia de distribuciones $\mathcal{F} = \{p(\theta)\}$ se dice que es* **cerrada** *o equivalentemente* **conjugada bajo muestreo** *$P(x\,|\,\theta)$ si cuando la inicial para θ pertenece a $\mathcal{F}$, la final correspondiente a cualquier muestra aleatoria de X tambien pertenece a $\mathcal{F}$.*

Esta definición es general e incluso, incluye casos que son irrelevantes. Por ejemplo, si $P(\theta_*)$ es la distribución degenerada tal que $P(\theta = \theta_*) = 1$ y $\mathcal{F}$ es la familia de la forma $\{P(\theta)\,|\,P(\theta) = P(\theta_*),\ \theta_* \in \Theta\}$. En este caso, los datos serán irrelevantes y la distribución final $P(\theta\,|\,x_{(n)})$ también será $P(\theta_*)$ de forma que para cualquier esquema de muestreo $\mathcal{F}$ es conjugada. En sentido contrario, si $\mathcal{F}$ es la familia de todas las funciones de distribución, entonces, necesariamente y sin importar el esquema de muestreo, $P(\theta)$ y $P(\theta\,|\,x_{(n)})$ pertenecen a la misma familia $\mathcal{F}$ que es, evidentemente, conjugada e inútil para efectos de simplificar el cómputo.

De hecho, como puede comprobarse en los siguientes ejemplos, la idea iniciales conjugadas es de utilidad práctica cuando $\mathcal{F}$ es una familia paramétrica de modelos con características conocidas.

Ejemplo 6.3.1. *Del ejemplo 5.2.1 se puede ver que $P(\mu) = Normal(\mu\,|\,m, c^2)$ es conjugada para los datos $Normal(x\,|\,\mu, \sigma^2)$ con σ^2 conocido. De hecho, en este caso, la fórmula de Bayes se puede sustituir por dos ecuaciones de actualización que además de ser extraordinariamente simples, hacen posibles la interpretación de la manera en que se combinan la información muestral con la información inicial:*

$$m_x = \frac{\left(\frac{n}{\sigma^2}\right)\bar{x} + \left(\frac{1}{c^2}\right)m}{\left(\frac{n}{\sigma^2}\right) + \left(\frac{1}{c^2}\right)} \quad y \quad c_x^2 = \frac{1}{\left(\frac{n}{\sigma^2}\right) + \left(\frac{1}{c^2}\right)}\ .$$

Y equivalentemente, si se parametriza en términos de la precisión se tiene entonces que

$$m_x = \alpha\bar{x} + (1-\alpha)m \quad y \quad \tau_x = n\tau + \tau_\mu,$$

donde $\alpha = \frac{n\tau}{n\tau+\tau_\mu}$, $\tau = \frac{1}{\sigma^2}$, $\tau_\mu = \frac{1}{c^2}$ y $\tau_x = \frac{1}{c_x^2}$.

Estas ecuaciones de actualización son realmente interesantes, pues permiten observar, por ejemplo, que $\tau_x > \tau$ y que $m_x \to \bar{x}$ cuando $n \to \infty$ o cuando $\tau \to 0$.

En el ejemplo anterior, se observa que la familia Normal es conjugada bajo el muestreo (también) Normal. Sin embargo, como puede observarse en el siguiente ejemplo, esta coincidencia no es un caso general.

Ejemplo 6.3.2. *Suponga que X es una v.a. con distribución $Poisson(x\,|\,\lambda)$ y que el conocimiento inicial sobre λ se describe con una $Gamma(\lambda\,|\,\alpha, \beta)$, de forma que*

$$P(\lambda\,|\,X_{(n)}) \boldsymbol{\alpha} \left(\lambda^{\alpha-1}e^{-\beta\lambda}\right)\left(e^{-n\lambda}\lambda^{\sum x_i}\right) \boldsymbol{\alpha} \lambda^{\alpha+\sum x_i - 1}e^{-(\beta+n)\lambda}.$$

Es decir

$$P(\lambda \,|\, X_{(n)}) = Gamma(\lambda \,|\, \alpha + \sum x_i, \beta + n),$$

por lo que la clase Gamma es cerrada bajo muestreo Poisson. En cambio, si el conocimiento inicial se describe con una distribución Poisson$(\lambda \,|\, \phi)$ *resulta que*

$$P(\lambda \,|\, X_{(n)})\boldsymbol{\alpha}\frac{e^{-\phi}\phi^{\lambda}}{\lambda!}\,\frac{e^{-n\lambda}\lambda^{\sum x_i}}{\Pi\,(x_i!)}\boldsymbol{\alpha}\frac{(\phi e^{-n})^{\lambda}\,\lambda^{\sum x_i}}{\lambda!},$$

de donde se puede ver que la clase Poisson no es conjugada bajo muestreo Poisson. En este caso, de hecho, ni siquiera es razonable en general considerar una distribución inicial discreta para λ *puesto que el espacio parametral es* $\Re^{+}$.

La construcción de familias paramétricas conjugadas ha sido un amplio objeto de estudio, el resultado de este trabajo se resume en el siguiente teorema que es de gran utilidad

Teorema 6.3.1. *Sea* X *una v.a. con f.d.p.g.* $f\,(x \,|\, \theta)\ \ \theta \in \Theta$. *Supongamos que para toda m.a.* $x_{(n)}$ *de* X *existe una estadística* $T_n\left(x_{(n)}\right)$ *de dimensión fija* r *que es suficiente para* θ. *Si como consecuencia del teorema de factorización se tiene que* $P\left(x_{(n)} \,|\, \theta\right) = h\left(x_{(n)}\right) g\left(T_n\left(x_{(n)}\right), \theta\right)$ *y además se cumple que* $\int_{\Theta} g\left(T_n\left(x_{(n)}\right), \theta\right) d\theta < \infty$, *entonces existe una familia paramétrica conjugada (básica) para* θ.

Demostración. Sean $m \in \mathbb{N}$ y $\mathscr{X}^{T}_{(m)}$ el contradominio de $T_m\left(x_{(m)}\right)$, y sean además $\mathcal{X}^{T} = \bigcup_{m=1}^{\infty}\left\{(m, t_m) \,|\, t_m \in \mathscr{X}^{T}_{(m)}\right\}$, $\quad \phi^{t} = (m, t_m)$ con $t_m \in \mathscr{X}^{T}_{(m)}$,
$\phi^{t} \in \mathcal{X}^{T} \subseteq \Re^{r+1}$ y $\mathcal{F}_{\phi} = \left\{P(\theta \,|\, \phi) \,|\, P(\phi \,|\, \theta)\boldsymbol{\alpha} g(t_m, \theta), \phi \in \mathcal{X}^{T}\right\}$.
Si $P(\theta) \in \mathcal{F}_{\phi}$, entonces $P(\phi \,|\, \theta) = cg(t_m, \theta)$ para algún $m \in \mathbb{N}$. Y tomando una m.a $x_{(n)}$, resulta que

$$\begin{aligned} P(\theta \,|\, x_{(n)})\boldsymbol{\alpha}\, & P(x_{(n)} \,|\, \theta)P(\theta)\,\boldsymbol{\alpha}\, h\left(x_{(n)}\right) g\left(T_n\left(x_{(n)}\right), \theta\right) P(\theta \,|\, \phi) \\ & \boldsymbol{\alpha}\, g\left(T_n\left(x_{(n)}\right), \theta\right) g(t_m, \theta). \end{aligned}$$

Ahora, si $y_{(m)}$ es una m.a. tal que $T_{(m)}(y_{(m)}) = t_m$, entonces

$$\begin{aligned} P(\theta \,|\, x_{(n)})\boldsymbol{\alpha}\, & g\left(T_n\left(x_{(n)}\right), \theta\right) g\left(T_m\left(y_{(m)}\right), \theta\right) \\ & \boldsymbol{\alpha}\, P(x_{(n)} \,|\, \theta) g\left(T_m\left(y_{(m)}\right), \theta\right) \\ & \boldsymbol{\alpha}\, P(x_{(n)} \,|\, \theta) P(y_{(m)} \,|\, \theta) \\ & \boldsymbol{\alpha}\, P(z_{(n+m)} \,|\, \theta)\,\boldsymbol{\alpha}\, g\left(T_{n+m}\left(z_{(n+m)}\right), \theta\right), \end{aligned}$$

donde z es una m.a. de tamaño $m + n$. Por lo tanto

$$P\left(\theta \,|\, x_{(n)}\right) = P\left(\theta \,|\, \phi_x\right) \quad \text{con}\, \phi_x = (n + m, t_n + m),$$

y así, $P\left(\theta \,|\, x_{(n)}\right) \in \mathcal{F}_{\phi}$. □

Este teorema se ilustra con el siguiente ejemplo donde, además, se puede observar como la familia conjugada paramétrica básica se puede extender a una familia más general.

Ejemplo 6.3.3. *Sea X una v.a. con distribución Bernoulli$(x\,|\,\theta)$, $\theta \in [0,1]$ de forma que*

$$P\left(x_{(n)}\,|\,\theta\right) = \prod_{i=1}^{n} P(x_i\,|\,\theta) = \theta^{\sum x_i}\,(1-\theta)^{n-\sum x_i}$$

Por el teorema de factorización $T_n(x_{(n)}) = \sum x_i$ es una estadística suficiente para θ, tomando $h(x_{(n)}) = 1$ y $g\left(T_n\left(x_{(n)}\right),\theta\right) = \theta^{\sum x_i}\,(1-\theta)^{n-\sum x_i}$.
Observe que $T_n : \mathscr{X}_{(n)} \longrightarrow \mathcal{T}_n \subseteq \Re$ por lo que es de dimensión fija. Además

$$\int_{\Theta} \theta^{\sum x_i}\,(1-\theta)^{n-\sum x_i}\,d\theta = \int_{\Theta} \theta^{\alpha-1}\,(1-\theta)^{\beta-1}\,d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} < \infty$$

Donde $\alpha = \sum x_i + 1$ y $\beta = n - \sum x_i + 1$. Por lo que existe un familia paramétrica conjugada básica para el muestreo Bernoulli y está dada por la forma

$$\mathcal{F}_\phi = \{P(\theta\,|\,\phi)\,|\,P(\theta\,|\,\phi) \boldsymbol{\alpha}\, g(t_m,\theta),\ \textit{con } m \in \mathbb{N} \textit{ y } t_m \in \mathcal{T}_m\}\,,$$

y donde $g\left(T_n\left(x_{(n)}\right),\theta\right) = \theta^{\sum x_i}\,(1-\theta)^{n-\sum x_i}$, y $\mathcal{T}_m = \{0,1,\ldots,m\}$.
Así, $\mathcal{F}_\phi$ está formado por un subconjunto de todas las distribuciones Beta$(\theta\,|\,\alpha,\beta)$. Específicamente, aquellas tales que $\alpha,\beta \in \mathbb{N}$.

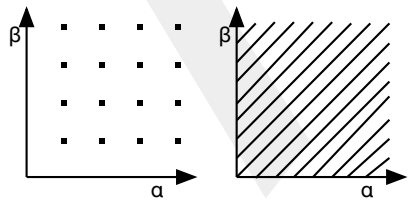


Figura 6.1: Espacio paramétrico de la familia conjugada básica y la familia Beta.

Ahora bien, es un ejercicio elemental probar que si la familia básica $\mathcal{F}_\phi$ se generaliza a la de las distribuciones Beta, la familia resultante sigue siendo paramétrica y conjugada.

En general, con el empleo de familias paramétricas conjugadas, el proceso de aprendizaje que se lleva a cabo a través de la formula de Bayes, se reduce a la actualización de los parámetros de la inicial para obtener los de la final.

Considere $\mathcal{F}_\phi = \{P(\theta) = P(\theta\,|\,\phi) : \phi \in \Phi\}$ y $x_{(n)}$ una m.a. de una v.a. X con f.d.p.g. $P(x\,|\,\theta)$. Si $P(\theta) \in \mathcal{F}_\phi$, y $\mathcal{F}_\phi$ es conjugada para el muestreo $P(x\,|\,\theta)$,

entonces $P(\theta) = P(\theta \mid \phi)$. Y si se calcula la final $P(\theta \mid x_{(n)}) \boldsymbol{\alpha} P(x_{(n)} \mid \theta) P(\theta \mid \phi)$, resulta que $P(\theta \mid x_{(n)}) \in \mathcal{F}_\phi$. Lo que implica que $P(\theta \mid x_{(n)}) = P(\theta \mid \phi_x)$ donde $\phi_x = g(\phi, x_{(n)})$.

Así, la ecuación de actualización

$$\phi_x = g(\phi, x_{(n)})$$

resume todo el proceso de aprendizaje.

6.4. Distribuciones no informativas

Como se mencionó en la sección 4.2, en ocasiones interesa utilizar una distribución inicial que sea no informativa. En ese caso, una pregunta relevante es ¿Cómo encontrar dichas distribuciones? Este problema ha recibido mucha atención en la literatura estadística Bayesiana, y a través de las familias conjugadas se puede proveer una posible respuesta.

6.4.1. Distribuciones conjugadas mínimo informativas

Como se discutió, en el caso de las familias conjugadas el proceso de aprendizaje y la combinación de la información inicial con la muestral, quedan plasmados en la ecuación de actualización paramétrica $\phi_x = g(\phi, x_{(n)})$. En donde es claro que la distribución final es un elemento de la familia en cuestión, que al estar determinada por el parámetro ϕ_x tiene influencia de la información muestral, específicamente, a través de una estadística suficiente de dimensión fija, y de la distribución inicial a través del parámetro ϕ.

Así, la idea de una final que fundamentalmente dependa de los datos se puede llevar al terreno operativo si en la expresión para ϕ_x, el parámetro ϕ se fija o se hace tender a un límite convencional que, en algún sentido, elimine la contribución de la inicial en la distribución final. Un ejemplo que puede clarificar esta idea se presenta a continuación.

Ejemplo 6.4.1. *Sea X una v.a. con distribución Bernoulli$(x \mid \theta)$, $\theta \in [0,1]$ de forma que*

$$P(x \mid \theta) = \theta^x (1-\theta)^{1-x},$$

y por tanto la verosimilitud cumple que

$$P\left(x_{(n)} \mid \theta\right) = \theta^{\sum x_i} (1-\theta)^{n-\sum x_i}.$$

Por otro lado, se sabe del ejemplo 6.3.3 que la familia paramétrica dada por $\mathcal{F}_{(\alpha,\beta)} = \left\{ P(\theta) \boldsymbol{\alpha} \theta^{\alpha-1} (1-\theta)^{\beta-1} \right\}$ es conjugada bajo muestreo Bernoulli(θ).

Por tanto,

$$P\left(\theta \,|\, x_{(n)}\right) \boldsymbol{\alpha} \left[\theta^{\sum x_i}\left(1-\theta\right)^{n-\sum x_i}\right]\left[\theta^{\alpha-1}\left(1-\theta\right)^{\beta-1}\right]$$
$$\boldsymbol{\alpha}\ \theta^{\alpha+\sum x_i-1}\left(1-\theta\right)^{\beta+n-\sum x_i-1} = Beta(\theta \,|\, \alpha_x, \beta_x)$$

con $\alpha_x = \alpha + \sum x_i$ *y* $\beta_x = \beta + n - \sum x_i$. *De hecho, en la familia paramétrica conjugada básica se tiene que*

$$\theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{\sum y_i}\left(1-\theta\right)^{m-\sum y_i},$$

donde $y_{(m)}$ *es una muestra aleatoria hipotética de tamaño* m *de la misma v.a.* X. *De esta forma, a priori* θ *sigue una distribución* $Beta(\theta \,|\, \sum y_i+1, m-\sum y_i+1)$, *esto es,* $\alpha = \sum y_i + 1$ *y* $\beta = m - \sum y_i + 1$. *Así,* α *se puede interpretar como el número de éxitos (más uno) en una muestra hipotética de tamaño* m, *mientras que* β *equivale al número de fracasos (más uno) en la misma muestra hipotética.*

La idea entonces, para encontrar una distribución de referencia, es tomar (α_x, β_x) *de manera tal que anulen la presencia de la distribución inicial. En este sentido, si la inicial se puede interpretar como proveniente de la información contenida en una muestra hipotética de tamaño* m, *entonces una manera de "minimizar" o "eliminar" la información inicial es tomar* $m = 0$, *es decir, trabajar como si no hubiese muestra hipotética.*

Claramente, si $m = 0$ *entonces el número de éxitos y fracasos hipotéticos también debe de ser cero y por tanto,* $\alpha = 1$ *y* $\beta = 1$. *En consecuencia, la inicial con "menor" información para* θ *es una* $Beta(\theta \,|\, 1, 1)$, *es decir, una distribución Uniforme en* $[0, 1]$. *A este tipo de distribuciones se les conoce como* **mínimo informativas límite de conjugadas**.

Finalmente, observe que en este caso, cuando se utiliza esta distibución inicial,

$$P\left(\theta \,|\, X_{(n)}\right) \boldsymbol{\alpha} \left[\theta^{\sum x_i}\left(1-\theta\right)^{n-\sum x_i}\right] \mathbb{I}_{[0,1]}(\theta) = Beta\left(\theta \,|\, \sum x_i + 1, n - \sum x_i + 1\right)$$

es decir, en el proceso de actualización, la distribución final adquiere información únicamente a través de la muestra.

En el caso Bernoulli la inicial mínimo informativa límite de conjugadas, es un modelo de probabilidad en toda la extensión del término; sin embargo, este no es el caso en general. Considere el siguiente ejemplo.

Ejemplo 6.4.2. *Sea* X *una v.a. con distribución* $Normal(x \,|\, \mu, \tau)$ *con* τ *conocida. Se sabe entonces que* $\mathcal{F}_{(\phi)} = \{P(\mu) = N(\mu \,|\, m, \tau_\mu)\}$ *con* $\phi = (m, \tau_\mu)$ *es conjugada bajo muestreo Normal y, por tanto,* $P\left(\mu \,|\, x_{(n)}\right)$ *sigue una distribución* $Normal(\mu \,|\, m_x, \tau_x)$ *donde las reglas de actualización de los parámetros están determinadas de la siguiente manera:*

$$m_x = \frac{(n\tau)\bar{x} + \tau_\mu m}{n\tau + \tau_\mu} \quad y \quad \tau_x = n\tau + \tau_\mu.$$

Así, en este caso, una muestra hipotética $(y_{(r)})$ *de tamaño* r *produce la siguiente conjunta:*

$$\begin{aligned} P\left(y_{(r)} \mid \mu\right) &= \prod_{i=1}^{r} P\left(y_i \mid \mu\right) \\ &= \left(\frac{2\pi}{\tau}\right)^{-r/2} e^{-(\tau/2)\sum(y_i-\mu)^2} \\ &= \left(\frac{2\pi}{\tau}\right)^{-r/2} e^{-(\tau/2)\sum(y_i-\bar{y})^2} e^{-(r\tau/2)(\mu-\bar{y})^2}, \end{aligned}$$

de tal manera que, como función de μ, $g(\mu, y_r)\,\boldsymbol{\alpha}\, e^{-(r\tau/2)(\mu-\bar{y})^2}$ *y, en consecuencia la inicial conjugada resulta* $P(\mu) = Normal(\mu \mid \bar{y}, r\tau)$.

Procediendo como en el ejemplo Bernoulli, la manera de producir una inicial que, en un sentido, refleje la ausencia de información a priori, consiste en proceder como si no se contara con la muestra hipotética y_r *o, equivalentemente, tomando* $r = 0$. *En este caso, sin embargo, debe observarse que el hecho de que* $r = 0$ *implica que la distribución a priori* $P(\mu)$ *tenga varianza infinita. Evidentemente, no existe una distribución Normal con varianza infinita. Sin embargo, volviendo a las ecuaciones de actualización que conducen a la posterior, se observa que si* $\tau \to 0$, *entonces* $m_x \to \bar{x}$ *y* $\tau_x \to n\tau$.

Por tanto, cuando $\tau \to 0$, *se tiene que a posteriori* μ *sigue una distribución* $Normal(\mu \mid \bar{x}, n\tau)$. *Lo que equivale, formalmente, a utilizar en la fórmula de Bayes una distribución inicial que cumpla que*

$$P(\mu)\,\boldsymbol{\alpha}\, 1 \;\forall \mu \in \Re,$$

que no es un modelo de probabilidad, en tanto que no tiene una integral finita. De hecho, este tipo de funciones, que se pueden considerar límite de una sucesión de distribuciónes de probabilidad se denominan **distribuciones impropias** *y, aparecen con frecuencia cuando se buscan iniciales no informativas.*

Utilizar el límite de conjugadas para proponer distribuciones mínimo informativas, es sólo uno entre una variedad de procedimientos que se han propuesto en la literatura para este fin. De hecho, estos trabajos se pueden rastrear hasta el propio T. Bayes, quien utilizó la distribución $Uniforme$ como inicial para el modelo *Bernoulli* sin recurrir al concepto de conjugadas, o mejor aún a Laplace que, en su legendario principio de la razón insuficiente, propuso utilizar la distribución $Uniforme$ como representación general de la ignorancia.

Precisamente en contra del empleo indiscriminado de la distribución $Uniforme$ como representación de la falta de información, aparece el método atribuido a Harold Jeffreys que se estudiará, en detalle, a continuación.

6.4.2. Regla de Jeffreys

El criterio de la razón insuficiente de Laplace establece que ante la ausencia de información, no hay razón para que un resultado posible reciba una asignación de probabilidad distinta de otro. Esto es, que la ausencia de información se debe representa mediante una distribución $Uniforme$. Sin embargo, como se observa en el siguiente ejemplo, el uso indiscriminado de la distribución $Uniforme$ como representación de la ignorancia, puede llevar a resultados inconsistentes.

Ejemplo 6.4.3. (La distribución incicial $Uniforme$ no es una representación universal de la ignorancia).
Sea X una v.a. con distribución $Bernoulli(x \mid \theta)$, $\theta \in \Theta = [0,1]$. Esto es

$$P(x \mid \theta) = (\theta)^x (1-\theta)^{1-x}.$$

Ahora, si se considera la reparametrización $\delta = \theta^2$, de modo que $\delta \in [0,1]$ es una función uno a uno de θ, el modelo de muestreo resulta

$$P(x \mid \delta) = \left(\delta^{1/2}\right)^x \left(1-\delta^{1/2}\right)^{1-x}.$$

De esta manera, si se carece de información inicial sobre θ, el principio de la razón insuficiente implica necesariamente que a priori θ debe seguir una distribución $Uniforme(\theta \mid 0,1)$. Pero, puesto que $\delta = \theta^2$, la ignorancia sobre θ implicará la ignorancia sobre δ y, por tanto, nuevamente, el principio de la razón insuficiente implica que a priori δ debe seguir una distribución $Uniforme(\delta \mid 0,1)$. Sin embargo, la distribución de δ puede derivarse a partir de θ, a través del correspondiente cambio de variable. Esto es,

$$P_\Delta(\delta \leq c) = P_\Theta(\theta^2 \leq c) = P_\Theta(\theta \leq \sqrt{c}) = \begin{cases} \sqrt{c} & si\ 0 \leq c \leq 1, \\ 1 & si\ 1 < c. \end{cases}$$

que no es Uniforme ❣

El ejemplo anterior, ilustra el hecho de que la asignación indiscriminada de la distribución $Uniforme$ como representación de una situación mínimo informativa, contraviene el procedimiento de cálculo de probabilidades a través de cambios de variable. En la literatura se dice que la idea de asignar una $Uniforme$ siempre que se enfrenta una situación de poca información no satisface el principio de invarianza.

Existen varias soluciones para este problema. Una de ellas, consiste en utilizar la mínimo informativa límite de conjugadas para θ y definir la mínimo informativa para $\phi = g(\theta)$ como la que se deriva, vía cambio de variable, a partir de la correspondiente θ.

La idea de Jefreys, en alguna forma, extiende la propuesta de Laplace al considerar que una distribución $Uniforme$ es una descripción razonable para un caso de poca información cuando el parámetro θ es de localización, y más en

particular, cuando θ es la media de una distribución $Normal$. Así, una distribución mínimo informativa para $\phi = g(\theta)$ se puede inducir con un cambio de variable. Precisamente, la importancia de la regla de Jeffreys es que extiende este razonamiento a una clase de modelos muy general.

La intención entonces, es buscar una reparametrización ϕ del parámetro original θ, de manera que, al menos asintóticamente, ϕ pueda interpretarse como la media de un modelo $Normal$; una vez identificado el parámetro ϕ, asignarle la distribución inicial mínimo informativa $P(\phi) \propto 1$, y mediante el cambio de variable determinar la distribución mínimo informativa para θ. Esta distribución se conoce como la inicial de Jeffreys (para θ), y se denota mediante $P_J^{\theta}(\theta)$. Por supuesto, una vez determinada $P_J^{\theta}(\theta)$, la inicial de Jeffreys para cualquier función $\nu = h(\theta)$ se obtiene directamente como

$$P_J^{\nu}(\nu) = P_J^{\theta}\left(h^{-1}(\nu)\right)\left|\frac{d\theta}{d\nu}\right|.$$

Así, el procedimiento general es el siguiente:

Sean X una v.a. con f.d.p.g. $P(x \mid \theta)$, $\theta \in \Theta$ y $x_{(n)}$ una m.a. de X. Si $P(\theta)$ es la distribución inicial, entonces

$$P\left(\theta \mid x_{(n)}\right) \propto P\left(x_{(n)} \mid \theta\right) P(\theta).$$

Además, sean $L(\theta)$ la función de verosimilitud, $l(\theta) = log(L(\theta))$ la logverosimilitud y $\hat{\theta}$ el estimador de máxima verosimilitud para θ. Así, aproximando $l(\theta)$ como función de θ a través de la serie de Taylor de orden dos, alrededor de $\hat{\theta}$ se tiene que

$$l(\theta) \backsimeq l(\hat{\theta}) + l'(\hat{\theta})(\theta - \hat{\theta}) + \frac{l''(\hat{\theta})}{2}(\theta - \hat{\theta})^2,$$

pero, puesto que $\hat{\theta}$ es un máximo, se cumplirá que

$$l'(\hat{\theta}) = 0 \quad \text{y} \quad l''(\hat{\theta}) < 0,$$

y así

$$l(\theta) \backsimeq l(\hat{\theta}) + \frac{1}{2} l''(\hat{\theta})(\theta - \hat{\theta})^2.$$

Por tanto, tomando la exponencial en ambos lados

$$\begin{aligned} L(\theta) \backsimeq & exp\{l(\hat{\theta}) + \frac{1}{2} l''(\hat{\theta})(\theta - \hat{\theta})^2\} \\ & = \left[exp\{l(\hat{\theta})\}\right]\left[exp\left\{-\frac{1}{2}\left[-l''(\hat{\theta})\right](\theta - \hat{\theta})^2\right\}\right] \\ & \propto exp\left\{-\frac{1}{2}\left[-l''(\hat{\theta})\right](\hat{\theta} - \theta)^2\right\}, \end{aligned}$$

Esta expresión revela que, asintóticamente, la verosimilitud para θ guarda semejanza con la verosimilitud que se obtiene de una observación de $\hat{\theta}$ con distribución $Normal$ de media θ, salvo que, la variable aleatoria $\hat{\theta}$ también aparece en el término $l''(\hat{\theta})$ que debiera ser constante. Para solventar este inconveniente es oportuno recordar que la aproximación supone que n es grande, y mejora a medida que $n \to \infty$. Por otro lado, si se define $u_i = ln\left(P\left(x_i \,|\, \theta\right)\right)$, se tiene que

$$\begin{aligned} l(\theta) =& ln\left(L\left(\theta\right)\right) = ln\left(P\left(x_{(n)} \,|\, \theta\right)\right) = ln\left(\prod_{i=1}^{n} P\left(x_i \,|\, \theta\right)\right) \\ =& \sum_{i=1}^{n} ln\left(P\left(x_i \,|\, \theta\right)\right) = \sum_{i=1}^{n} u_i = n\bar{u} \xrightarrow[n\to\infty]{} n\mathbb{E}(u). \end{aligned}$$

Y análogamente, tomando $v_i = \frac{\partial^2}{\partial\theta^2} ln\left(P\left(x_i \,|\, \theta\right)\right)$, que

$$\begin{aligned} l''(\theta) =& \frac{\partial^2}{\partial\theta^2} l(\theta) = \frac{\partial^2}{\partial\theta^2} \sum_{i=1}^{n} ln\left(P\left(x_i \,|\, \theta\right)\right) = \sum_{i=1}^{n} \left[\frac{\partial^2}{\partial\theta^2} ln\left(P\left(x_i \,|\, \theta\right)\right)\right] \\ =& \sum_{i=1}^{n} v_i = n\bar{v} \xrightarrow[n\to\infty]{} n\mathbb{E}(v). \end{aligned}$$

Además, en este último caso observe que

$$\mathbb{E}(v) = \mathbb{E}\left\{\frac{\partial^2}{\partial\theta^2} ln\left(P\left(x \,|\, \theta\right)\right)\right\} = -i_x(\theta)$$

donde $i_x(\theta)$ es la información de Fisher contenida en x para θ. Entonces, como consecuencia, se tiene que

$$l''(\hat{\theta}) = -ni_x(\hat{\theta}).$$

Ahora, si se considera la reparametrización $\phi = \phi(\theta)$ se tendrá que cumplir que

$$L(\phi \,|\, x_{(n)}) \backsimeq exp\left\{-\frac{1}{2}\left[-l''(\hat{\phi})\right](\hat{\phi} - \phi)^2\right\},$$

donde $-l''(\hat{\phi}) \to ni_x(\phi) = -n\mathbb{E}\left\{\frac{\partial^2}{\partial\phi^2} ln\left(P\left(x \,|\, \phi\right)\right)\right\}$. Pero, utilizando iteradamente la regla de la cadena se obtiene que

$$\frac{\partial ln(p(x \,|\, \phi))}{\partial\phi} = \frac{\partial ln(p(x \,|\, \theta))}{\partial\theta}\left(\frac{\partial\theta}{\partial\phi}\right) = l'(\theta)\left(\frac{\partial\theta}{\partial\phi}\right),$$

y también que

$$\frac{\partial^2 ln(p(x \,|\, \phi))}{\partial\phi^2} = \frac{\partial}{\partial\phi}\left(l'(\theta)\left(\frac{\partial\theta}{\partial\phi}\right)\right) = l''(\theta)\left(\frac{\partial\theta}{\partial\phi}\right)^2 + l'(\theta)\left(\frac{\partial^2\theta}{\partial\phi^2}\right)$$

expresión que, tomando valor esperado en ambos lados, se convierte en

$$\mathbb{E}\left\{\frac{\partial^2}{\partial\phi^2} ln\left(P\left(x\,|\,\phi\right)\right)\right\} = -\left(\frac{\partial\theta}{\partial\phi}\right)^2 \mathbb{E}\left\{l''(\theta)\right\} - \left(\frac{\partial^2\theta}{\partial\phi^2}\right)\mathbb{E}\left\{l'(\theta)\right\}.$$

Así, utilizando que

$$\begin{aligned}\mathbb{E}\left\{l'(\theta)\right\} =& \mathbb{E}\left\{\frac{\partial}{\partial\theta} ln\left(P\left(x\,|\,\theta\right)\right)\right\} = \mathbb{E}\left\{\frac{1}{P(x\,|\,\theta)}\left[\frac{\partial}{\partial\theta}P\left(x\,|\,\theta\right)\right]\right\}\\ =& \int_{\mathscr{X}} \frac{1}{P(x\,|\,\theta)}\left[\frac{\partial}{\partial\theta}P\left(x\,|\,\theta\right)\right] P(x\,|\,\theta)dx\\ =& \frac{\partial}{\partial\theta}\int_{\mathscr{X}} P\left(x\,|\,\theta\right)dx = \frac{\partial}{\partial\theta}1 = 0,\end{aligned}$$

resulta que

$$\mathbb{E}\left\{\frac{\partial^2}{\partial\phi^2} ln\left(P\left(x\,|\,\phi\right)\right)\right\} = -\left(\frac{\partial\theta}{\partial\phi}\right)^2 \mathbb{E}\left\{l''(\theta)\right\} = \left(\frac{\partial\theta}{\partial\phi}\right)^2 i_x(\theta),$$

y por lo tanto

$$i_x(\phi) = \left(\frac{\partial\theta}{\partial\phi}\right)^2 i_x(\theta).$$

Esta expresión permite la búsqueda de una parametrización en cuyos términos la verosimilitud asintótica se pueda interpretar como la asociada a una media *Normal*. Basta pedir que $i_x(\phi)\,\boldsymbol{\alpha}\,1$ o equivalentemente que

$$\left(\frac{\partial\theta}{\partial\phi}\right)^2 i_x(\theta)\,\boldsymbol{\alpha}\,1,$$

es decir,

$$\frac{\partial\phi}{\partial\theta}\,\boldsymbol{\alpha}\,(i_x(\theta))^{1/2},$$

de modo que $\phi(\theta)$ necesariamente será de la forma $\phi(\theta) = \int (i_x(\theta))^{1/2} d\theta$. Y por tanto, si se toma $P(\phi)\,\boldsymbol{\alpha}\,1$ resulta que

$$P_J^{\theta}(\theta)\,\boldsymbol{\alpha}\,\sqrt{i_x(\theta)}$$

resultado que se conoce como la regla de Jeffreys.

Ejemplo 6.4.4. *Sea $x_{(n)}$ una m.a. de una v.a. X con f.d.p.g. $Normal(x\,|\,\mu,\sigma^2)$ y σ^2 conocida, de forma que*

$$ln\left(P(x\,|\,\mu)\right) = ln\left\{(2\pi\sigma^2)^{-1/2} e^{-1/2\sigma^2(x-\mu)^2}\right\} = -\frac{1}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2,$$

y, adicionalmente, la verosimilitud está dada por

$$\begin{aligned} L(\mu) =& (2\pi\sigma^2)^{-n/2} e^{-1/2\sigma^2 \sum (x_i-\mu)^2} \\ & \boldsymbol{\alpha}\, e^{-1/2\sigma^2 \sum (x_i-\bar{x})^2} e^{-n/2\sigma^2 (\mu-\bar{x})^2} \\ & \boldsymbol{\alpha}\, e^{-n/2\sigma^2 (\mu-\bar{x})^2}\, \boldsymbol{\alpha}\, N\left(\mu \,|\, \bar{x}, \frac{\sigma^2}{n}\right). \end{aligned}$$

Así, tomado la primera y segunda derivada, con respecto a μ, de $ln\left(P(x \,|\, \mu)\right)$ se tiene que

$$\frac{\partial ln\left(P(x \,|\, \mu)\right)}{\partial \mu} = \frac{(x-\mu)}{\sigma^2} \quad y \quad \frac{\partial^2 ln(P(x \,|\, \mu))}{\partial \mu^2} = -\frac{1}{\sigma^2},$$

lo que implica que $\hat{\mu} = \bar{x}$, y que $i_x(\mu) = \frac{1}{\sigma^2}$. Y entonces, la inicial mínimo informativa de Jeffreys para μ está dada por

$$P_J^{\mu}(\mu)\, \boldsymbol{\alpha} \sqrt{\frac{1}{\sigma^2}}\, \boldsymbol{\alpha}\, 1$$

Análogamente, para el caso en que μ en conocida y en términos de la precisión $\tau = \frac{1}{\sigma^2}$, se tiene que

$$\begin{aligned} ln\left(P(x \,|\, \tau)\right) =& ln\left\{(2\pi\tau^{-1})^{-1/2} e^{-\tau/2(x-\mu)^2}\right\} = -\frac{1}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2 \\ =& k + \frac{1}{2} ln(\tau) - \frac{\tau}{2}(x-\mu)^2 \end{aligned}$$

y la verosimilitud

$$\begin{aligned} L(\tau) =& (2\pi\tau^{-1})^{-n/2} e^{-\tau/2 \sum (x_i-\mu)^2}\, \boldsymbol{\alpha}\, \tau^{n/2} e^{-\left[\frac{\sum (x_i-\mu)^2}{2}\right]\tau} \\ & \boldsymbol{\alpha}\, \tau^{\alpha-1} e^{-\beta\tau} = Gamma\left(\alpha, \frac{1}{\beta}\right). \end{aligned}$$

Por lo que

$$\frac{\partial ln\left(P(x \,|\, \tau)\right)}{\partial \tau} = \frac{1}{2\tau} - \frac{(x-\mu)^2}{2} \quad y \quad \frac{\partial^2 ln(P(x \,|\, \tau))}{\partial \tau^2} = -\frac{1}{2\tau^2},$$

lo que implica que $\hat{\tau} = \frac{n}{\sum (x_i-\mu)^2}$, y que $i_x(\tau) = \frac{1}{2\tau^2}$. Y entonces, la inicial mínimo informativa de Jeffreys para τ está dada por

$$P_J^{\tau}(\tau)\, \boldsymbol{\alpha} \sqrt{\frac{1}{2\tau^2}}\, \boldsymbol{\alpha}\, \frac{1}{\tau}.$$

6.5. Ejercicios

Ejercicio 6.1. *Se dice que una variable aleatoria X (discreta o continua) tiene una distribución que pertenece a la familia exponencial si su f.d.p.g. se puede escribir como*

$$f(x \mid \theta) = h(x)w(\theta)exp\left\{\sum_{j=1}^{k} c_j(\theta)u_j(x)\right\}$$

donde el rango de X no depende del parámetro θ y, para toda j, las funciones h, w, c_j y u_j son totalmente conocidas. Demuestre que los modelos $Normal$, $Bernoulli$, $Poisson$ y $Exponencial$ pertenecen a esta familia. ¿Puede sugerir algún otro modelo que también pertenezca a la familia exponencial?

Ejercicio 6.2. *Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria de una variable aleatoria X con una distribución que pertenece a la familia exponencial. Determine una estadística suficiente para θ. ¿Es de dimensión fija?*

Ejercicio 6.3. *Sea X una variable aleatoria continua con distribución $Bernoulli$ de parámetro θ, un valor en el intervalo $(0, 1)$. Suponga que a priori el conocimiento sobre θ se describe con una distribución $Beta(\alpha, \beta)$. Suponga además que se obtiene una muestra de tamaño 10 de X, en la que se registran 7 éxitos y 3 fracasos. Exhiba en una misma gráfica la densidad inicial, la verosimilitud y la final para θ, en cada uno de los siguientes casos para el vector (α, β) : (0,5, 0,5), $(1, 1)$, $(5, 5)$, $(1, 9)$, $(6, 14)$ y $(7, 3)$. Analice y comente estas gráficas.*

Ejercicio 6.4. *Sea X una variable aleatoria continua con distribución $Uniforme$ en $(0, \theta)$. Construya una familia paramétrica conjugada para θ en este caso.*

Ejercicio 6.5. *Sea X una variable aleatoria con distribución $Poisson$ de parámetro $\lambda > 0$. Muestre que la familia paramétrica $Gamma(\alpha, \beta)$ es conjugada para λ en este modelo.*

Ejercicio 6.6. *Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria de una variable aleatoria X con distribución $Exponencial$ de parámetro θ. Esto es*

$$f(x \mid \theta) = \theta exp(-\theta x);\ x, \theta > 0.$$

Suponga que a priori el conocimiento sobre θ se describe con una distribución $Gamma(\alpha, \beta)$. Demuestre que entonces la distribución final de θ es también una $Gamma$, y exhiba explícitamente la regla de actualización de los parámetros.

Ejercicio 6.7. *Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria de una variable aleatoria X con distribución $Normal$ de media μ conocida y precisión τ. Compruebe que si la distribución inicial de τ es $Gamma$ entonces la final también es $Gamma$. Exhiba la relación que guardan los parámetros de la inicial con los de la final.*

Ejercicio 6.8. *Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria de una variable aleatoria X con distribución $Normal$ de media μ y precisión τ. Suponga que la distribución inicial conjunta $P(\mu, \tau)$ se expresa como*

$$P(\mu, \tau) = P(\mu \,|\, \tau) P(\tau),$$

y que, en términos de media y precisión, $P(\mu \,|\, \tau) = Normal(\mu \,|\, m, u\tau)$ con $u > 0$ y m constantes conocidas, y $P(\tau) = Gamma(\alpha, \beta)$ con α y β también conocidas. En este caso, se dice que $P(\mu, \tau)$ sigue una distribución $Normal$–$Gamma$ con parámetros α, β, m y u. Compruebe que esta familia es conjugada para (μ, τ).

Ejercicio 6.9. *Sea X una variable aleatoria con distribución $Normal$ de media μ y precisión τ. Si a priori el conocimiento sobre los parámetros (μ, τ) se describe con una distribución $Normal - Gamma$ de parámetros α, β, m y u. Verifique que, también a priori, la distribución marginal de μ coincide con un modelo muy conocido en la literatura estadística. ¿Cuál es? ¿Cuál es el correspondiente resultado para la distribución final de μ?*

Ejercicio 6.10. *Sea X una variable aleatoria con distribución $Poisson$ de parámetro $\lambda > 0$. Determine la distribución inicial mínimo informativa límite de conjugadas para λ. Determine también la distribución inicial mínimo informativa de Jeffreys para λ. Compare estas iniciales.*

Ejercicio 6.11. *Sea una variable aleatoria X con distribución $Normal$ de media μ conocida y precisión τ. Determine la distribución inicial mínimo informativa límite de conjugadas para τ. Determine también la distribución inicial mínimo informativa de Jeffreys para τ. Compare estas iniciales.*

Ejercicio 6.12. *En el contexto del problema 6.11, ¿Cuál es la distribución inicial mínimo informativa de Jeffreys para σ^2 la varianza de X?*

Ejercicio 6.13. *Sea X una variable aleatoria con distribución $Exponencial$ de parámetro $\lambda > 0$. Determine la distribución inicial mínimo informativa de Jeffreys para $\mathbb{E}(X)$.*

Ejercicio 6.14. *Sea X una variable aleatoria con distribución $Bernoulli$ de parámetro θ, un valor en el intervalo $(0, 1)$. Determine la distribución inicial mínimo informativa de Jeffreys para θ^2.*

Bibliografía

[1] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Second edition. New York: Springer Verlag.

[2] Bernardo, J.M. (1981) *Bioestadística, una Perspectiva Bayesiana.* Barcelona: Vicens Vives.

[3] Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory.* Chichester: Wiley.

[4] Box, G.E.P. & Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis.* Reading: Addison Wesley.

[5] Casella, G. & Berger, R.L. (2001). *Statistical Inference.* Belmont: Duxbury Press.

[6] Congdon, P. (2001). *Bayesian Statistical Modelling.* Chichester: Wiley.

[7] De Groot, M.H. (1970). *Optimal Statistical Decisions.* New York: McGraw-Hill.

[8] De Groot, M.H. (1988). *Probabilidad y Estadística.* México: Addison Wesley Iberoamericana.

[9] Gamerman D. & Lopes, H.F. (2006). *Markov Chain Montecarlo. Stochastic Simulation for Bayesian Inference.* Second edition. London: Chapman & Hall.

[10] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2004). *Bayesian Data Analysis.* Second edition. London: Chapman & Hall.

[11] Lindley, D.V. (1965). *An Introduction to Probability and Statistics from a Bayesian Viewpoint.* Vol 2. Inference. Cambridge: Cambridge University Press.

[12] Lindley, D.V. (1985). *Making Decisions.* Second edition. London: Wiley.

[13] Mignon, H.S. and Gamerman, D. (1999). *Statistical Inference: An Integrated Approach.* London: Arnold.

[14] O'Hagan, A. (1994). *Kendalls Advanced Theory of Statistics.* Vol 2b. Bayesian Inference. Cambridge: Edward Arnold.

[15] Press, S.J. (1989). *Bayesian Statistics. Principles, Models and Applications.* New York: Wiley.

[16] Robert, C.P. (2001). *The Bayesian Choice.* Second edition. New York: Springer Verlag.