Apuntes de Estadística Bayesiana (Borrador)

Arturo Erdely Ruiz

2004

Índice general

Capítulo 1

Introducción

1.1. Limitaciones de la estadística frecuentista

Una de las principales limitaciones de la estadística frecuentista es que no permite incorporar de manera coherente en el análisis estadístico la *información extra-muestral* disponible, se apoya únicamente en datos muestrales observados. Si no hay datos, la estadística frecuentista está imposibilitada para operar. Si hay muy pocos datos, la estadística frecuentista presenta fuertes problemas también pues muchos de sus métodos se apoyan en resultados asintóticos, esto es, en la ley de los grandes números, el teorema del límite central, etc., y por ello por lo general pide muestras "grandes" para que sus resultados sean "confiables". La estadística bayesiana aprovecha tanto la información que nos proporcionan los datos muestrales así como la *información extra-muestral* disponible, entendiendo por esto último toda aquella información relevante que nos ayude a disminuir nuestra incertidumbre o ignorancia en torno a un fenómeno aleatorio de nuestro interés. Esquemáticamente:

$$\textit{Estadística frecuentista} \quad \longrightarrow \quad \textit{sólo datos}$$

$$\textit{Estadística Bayesiana} \quad \longrightarrow \quad \textit{datos} + \textit{info extra-muestral}$$

Y en un caso extremo en el que no se cuente o no sea posible contar con datos muestrales, si se cuenta con suficiente información extra-muestral, la estadística bayesiana es capaz de hacer inferencias.

En lo que se refiere a *contraste de hipótesis* la metodología de la estadística frecuentista está limitada a contrastar sólo 2 hipótesis, mientras que la metodología bayesiana permite contrastar n hipótesis a la vez. Existen muchas otras limitaciones de la estadística frecuentista que se analizarán a detalle conforme se expongan los distintos aspectos del enfoque bayesiano.

1.2. Enfoques de la Probabilidad

Enfoque clásico. Si un experimento o fenómeno aleatorio puede ocurrir de n maneras diferentes mutuamente excluyentes e igualmente probables tenemos que su espacio de probabilidad queda definido como $(\Omega, \mathcal{F}, \mathbb{P})$ en donde $\Omega = \{\omega_1, \ldots, \omega_n\}$, $\mathcal{F} = 2^\Omega$ y :

$$\mathbb{P}(A) = \frac{|A|}{n}$$

para todo evento $A \in \mathcal{F}$, donde $|A|$ es la cardinalidad de A.

Un ejemplo de experimento aleatorio que cumple con lo anterior es el de lanzar un dado (equilibrado) una vez. En este caso tenemos que $n = 6$. Sea A el evento de que salga número par, esto es $A = \{2, 4, 6\}$, y por lo tanto $\mathbb{P}(A) = \frac{1}{2}$.

El enfoque clásico tiene la limitante de que relativamente pocos problemas reales de interés pueden resolverse mediante un procedimiento tan sencillo como el del ejemplo anterior, en particular en lo que se refiere a la determinación de la medida de probabilidad adecuada.

Enfoque frecuentista. Bajo este enfoque la probabilidad de un evento A está dada por:

$$\mathbb{P}(A) = \lim_{n\to\infty} \frac{f_A(n)}{n}$$

donde $f_A(n)$ es el número de veces que ocurre el evento A en n repeticiones idénticas e independientes del experimento o fenómeno aleatorio.

Este enfoque presume de ser objetivo porque se basa sólo en datos observables pero:

- Tenemos que

$$\lim_{n\to\infty} \frac{f_A(n)}{n} = \mathbb{P}(A) \Leftrightarrow \forall\ \epsilon > 0\ \exists\ k \text{ tal que si } n > k\ \Rightarrow\ |\frac{f_A(n)}{n} - \mathbb{P}(A)| < \epsilon$$

lo cual **NO** se puede garantizar ya que bien puede $\exists\ n_0 > k$ tal que $|\frac{f_A(n)}{n} - \mathbb{P}(A)| > \epsilon$ (que en general resulta poco probable mas no imposible).

- f_A no es una función determinista como las que se utilizan en la teoría del cálculo para definir límites.
- El decir que se tienen repeticiones idénticas e independientes, fuera de los juegos de azar, es una apreciación subjetiva.
- En la práctica n nunca se va a ∞, ni cercanamente, así que no hay manera de comprobar dicho límite.

Enfoque subjetivo. La probabilidad de un evento A es una medida del grado de creencia que tiene un individuo en la ocurrencia de A con base en la información K que dicho individuo posee. Bajo este enfoque toda probabilidad es condicional en la información de la cual se dispone.

Por ejemplo, sea A el evento de que esté lloviendo en el centro de la Ciudad de México. Para un individuo que vive en el Polo Sur totalmente aislado del resto del mundo tendríamos que si K_1 denota la información (total ignorancia en este caso) que tiene el individuo respecto a lo que sucede en la Ciudad de México y al no haber razón alguna para asignar mayor probabilidad al evento A o a su complemento sólo queda establecer $\mathbb{P}(A|K_1) = \mathbb{P}(A^c|K_1)$ y como se debe cumplir $\mathbb{P}(A|K_1) + \mathbb{P}(A^c|K_1) = 1$ esto inmediatamente nos lleva a que $\mathbb{P}(A|K_1) = \frac{1}{2}$.

Si pensamos ahora en un individuo que vive en los suburbios de la Ciudad de México es claro que posee una información K_2 distinta a la del individuo en el Polo Sur y quizás podríamos hablar de algo como:

$$\mathbb{P}(A|K_2) = \begin{cases} \frac{3}{4} & \text{si está lloviendo en los suburbios} \\ \\ \frac{1}{4} & \text{si no está lloviendo en los suburbios} \end{cases}$$

Si bien es cierto que el hecho de que esté lloviendo en los suburbios de la Ciudad de México no es garantía de que esté lloviendo en el centro de la ciudad, dada la cercanía es más probable que así sea. Podemos decir que K_2 representa un mayor nivel de información que K_1. Y si ahora pensamos en un individuo que vive justamente en el centro de la Ciudad de México tenemos entonces que este inidividuo posee un nivel de información K_3 que

de hecho es el máximo nivel de información que se puede tener respecto al evento A y por lo tanto dicho individuo esta en posición de reportar uno de dos resultados: $\mathbb{P}(A|K_3) = 1$ o bien $\mathbb{P}(A|K_3) = 0$.

Lo importante a destacar en este ejemplo es el hecho de la existencia de distinitas medidas de probabilidad para un mismo evento, dependiendo de la cantidad de información con la que se cuente.

Son muy diversos los factores que incrementan nuestro nivel de información en relación a un fenómeno o experimento aleatorio, van desde la información que proveen datos históricos observados hasta la apreciación y experiencia de especialistas en dicho fenómeno.

Este enfoque de la probabilidad es ampliamente aprovechado por la metodología bayesiana y es por ello que podemos decir que la estadística bayesiana va más allá que la estadística frecuentista al buscar aprovechar **toda la información disponible**, así se trate de datos observados o de información de otro tipo que nos ayude a disminuir de manera coherente nuestra incertidumbre en torno a un fenómeno aleatorio de interés. Un buen ejemplo para ilustrar que efectivamente la pura experiencia de las personas puede contener información muy valiosa consiste en el siguiente ejercicio. En un salón de clase se solicita a cada estudiante que anote en un papel tres cosas: su estatura y las estaturas máxima y mínima que **él (o ella) creen** que hay en el salón. Aún cuando no se hayan practicado mediciones de estatura en el salón es sorprendente corroborar que en general los alumnos tendrán una idea de las estaturas máxima y mínima bastante cercana a la realidad, lo que nos da idea de la cantidad de información valiosa que puede poseer una apreciación subjetiva.

1.3. La Regla de Bayes

La estadística *bayesiana* toma su nombre del conocido resultado de probabilidad conocido como la *Regla de Bayes* así que brevemente enunciaremos los principales resultados al respecto.

Dado un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ si $A, B \in \mathcal{F}$ y $\mathbb{P}(B) > 0$ entonces la *probabilidad condicional* del evento A dado el evento B se define como:

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}\,.$$

Cabe recordar que aún cuando se tuviera que $\mathbb{P}(B) = 0$ existen resultados

de probabilidad que nos permiten calcular probabilidades condicionales en eventos de medida cero.

Si $\{B_n\}$ es una partición del espacio muestral Ω y para toda n tenemos que $B_n \in \mathcal{F}$ y $\mathbb{P}(B_n) > 0$, entonces:

$$\mathbb{P}(A) = \sum_n \mathbb{P}(A|B_n)\mathbb{P}(B_n) \quad \text{, para toda } A \in \mathcal{F}\,.$$

Un corolario importante del resultado anterior es:

$$\text{Si } B \in \mathcal{F} \Rightarrow \mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)\,.$$

Bajo los supuestos anteriores tenemos la *Regla de Bayes*:

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(A|B_k)\mathbb{P}(B_k)}{\sum_n \mathbb{P}(A|B_n)\mathbb{P}(B_n)} \quad \text{, } A \in \mathcal{F}\,,\, \mathbb{P}(A) > 0\,.$$

Y como corolario importante:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)}\,.$$

Si bien las demostraciones y ejemplos de lo anterior son propios de un curso de probabilidad bien vale la pena abordar un ejemplo que nos servirá más adelante para ilustrar el por qué algunos cálculos de la estadística frecuentista resultan cuestionables.

Supongamos que un grupo de ingenieros biomédicos mexicanos ha diseñado un nuevo aparato para diagnóstico del SIDA (en realidad es para diagnóstico de presencia de VIH pero coloquialmente nos referimos a esta inmunodeficiencia simplemente como SIDA, aunque los médicos nos regañen). Haremos el experimento de escoger a un individuo para probar dicho aparato y consideremos los eventos de interés A y B en donde A sea el evento de que el aparato diagnostique SIDA y B el evento de tener efectivamente SIDA. Los ingenieros que diseñaron este aparato presumen de que éste es muy bueno pues nos reportan que lo probaron con un grupo de 100 portadores del virus del SIDA y en 99 % de los casos el aparato dio positivo y que también lo probaron con 100 individuos sanos y que también en el 99 % de los casos el aparato dio negativo. Probabilísticamente esto se expresa:

$$\mathbb{P}(A|B) = \frac{99}{100} = \mathbb{P}(A^c|B^c)$$

Sea p la proporción de mexicanos con virus del SIDA. Tenemos entonces que $\mathbb{P}(B) = p$. Con la información anterior y utilizando la Regla de Bayes estamos en posición de calcular $\mathbb{P}(B|A)$, que quedará expresada en función de p :

$$\varphi(p) := \mathbb{P}(B|A) = \frac{\frac{99}{100}}{\frac{98}{100} + \frac{1}{100p}}\,.$$

Ahora tabulamos algunos valores $(p, \varphi(p))$:

p	$\mathbb{P}(B\|A)$
0.001	0.0902
0.010	0.5000
0.100	0.9167
0.500	0.9900

De lo anterior se observa que únicamente en el caso de que $p = \frac{1}{2}$ se cumple $\mathbb{P}(B|A) = \mathbb{P}(A|B)$ y que conforme $p \to 0$ estas dos probabilidades se alejan (siendo más precisos, $\mathbb{P}(B|A)$ se aleja de $\mathbb{P}(A|B)$ que permanece constante en 0.99). A primera vista pareciera que estamos discutiendo una trivialidad ya que es bien sabido que por lo general y en diversos contextos $\mathbb{P}(B|A)$ suele ser distinta de $\mathbb{P}(A|B)$ pero si nos detenemos un poco a analizar el ejemplo anterior esto tiene consecuencias terribles para los ingenieros que inventaron un aparato que creen es muy efectivo y en realidad no lo es. En México se estima que hay algo así como cien mil portadores del virus del SIDA, esto es, p =0.001, lo cual significa que la probabilidad de que un individuo tenga SIDA dado que el aparato dice que lo tiene es de tan solo 0.0902!

¿Qué sucedió? Sucede que el aparato fue probado con personas de las cuales conocíamos previamente su estado de salud, pero ya en la práctica cotidiana esto no sucede, las personas que llegan a practicarse un análisis lo hacen porque justamente desconocen cuál es su estado de salud (es decir si tienen o no el virus) y es por ello que el que $\mathbb{P}(A|B)$ sea de 99 % no implica necesariamente que $\mathbb{P}(B|A)$ sea igualmente alta. El ejemplo anterior nos será de mucha utilidad para comprender por qué algunos métodos de inferencia de la estadística frecuentista son cuestionables, en particular cuando se busca hacer inferencias sobre un parámetro θ y en lugar de calcular la probabilidad de que θ tome ciertos valores dada una muestra, es decir $\mathbb{P}[\,\theta \in \Theta_0|(x_1, \ldots, x_n)\,]$, la estadística frecuentista utiliza erróneamente la probabilidad de observar una determinada muestra bajo el supuesto de un valor

específico del parámetro, es decir $\mathbb{P}[\,(x_1,\ldots,x_n)|\,\theta=\theta_0\,]$ mejor conocida como la "verosimilitud", y esto es tanto como pretender que $\mathbb{P}(B|A)=\mathbb{P}(A|B)$ siempre se cumple.

1.4. La filosofía bayesiana

La teoría de la probabilidad se ocupa del estudio de la incertidumbre, de los fenómenos o experimentos aleatorios. La probabilidad depende de dos elementos: el evento incierto y las condiciones bajo las cuales es considerado, por lo que desde este punto de vista la probabilidad es siempre condicional. La estadística es una herramienta para la toma de decisiones bajo condiciones de incertidumbre.

Un enfoque científico sobre la incertidumbre es la medición de la misma. Kelvin dijo que sólo asociando números con el concepto científico es como se puede comprender adecuadamente dicho concepto. La razón de querer medir no es sólo para ser más precisos respecto a la intensidad de la incertidumbre sino también para combinar incertidumbres: En un problema típico de estadística encontramos combinadas la incertidumbre de los datos y la del parámetro.

La medición de la incertidumbre puede realizarse por medio del cálculo de probabilidades. En sentido inverso, las reglas de la probabilidad se reducen de manera simple a las reglas sobre proporciones. Esto explica por qué los argumentos frecuentistas son en muchos casos útiles: la combinación de incertidumbres puede ser estudiada por medio de proporciones o frecuencias. El objetivo de recolectar datos es precisamente reducir el nivel de incertidumbre, pero bajo la perspectiva bayesiana se aprovechan tanto los datos muestrales como otro tipo de informaciones que de manera coherente nos ayuden también a reducir nuestro nivel de incertidumbre en torno a los parámetros de interés.

En resumen:

- La estadística es una herramienta para la toma de decisiones bajo condiciones de incertidumbre.
- La incertidumbre debe ser medida por probabilidad.
- La incertidumbre sobre los datos debe ser medida condicionalmente en los parámetros.

- La incertidumbre sobre los parámetros es similarmente medida por probabilidad.
- La inferencia se lleva a cabo mediante cálculo de probabilidades, haciendo uso particular de la Regla de Bayes.

Las discusiones en contra del enfoque bayesiano se centran en el punto de medir la incertidumbre sobre el parámetro probabilísticamente, esto es, darle tratamiento de variable aleatoria. Para la estadística frecuentista existe algo que llaman "el verdadero valor del parámetro" que consideran fijo y que "sólo Dios conoce" pero que resulta desconocido para nosotros los mortales. Lo anterior, además de que los estadísticos frecuentistas rechazan la utilización de cualquier otro tipo de información que no provenga de los datos muestrales para hacer inferencias.

Para profundizar más a detalle en las ideas anteriores se recomienda la lectura del artículo de D.V. Lindley (2000) *The Philosophy of Statistics* publicado en *The Statistician*, Vol.49, pp.293-337.

Capítulo 2

El paradigma bayesiano

2.1. El modelo general

Para referirnos a un *modelo probabilístico paramétrico* general lo denotamos $p(x \mid \theta)$ en donde la función $p(\cdot \mid \theta)$ puede ser una función de masa de probabilidades de una variable (o vector) aleatoria (v.a.) discreta o bien una función de densidad de una v.a. continua. El escribir dicha función condicional en el parámetro (o vector de parámetros) θ se debe al hecho de que una vez dado un valor específico de θ la función de probabilidad queda totalmente determinada. Para referirnos a una muestra aleatoria (m.a.) utilizamos la notación $\mathbf{X} := (X_1, \ldots, X_n)$ y para referirnos a una observación muestral utilizamos $\mathbf{x} := (x_1, \ldots, x_n)$. Por *espacio paramétrico* entendemos el conjunto Θ de todos los valores que puede tomar θ y por *familia paramétrica* entendemos un conjunto $\mathcal{P} = \{p(x \mid \theta) : \ \theta \in \Theta\}$.

Al empezar a estudiar un fenómeno o experimento aleatorio recurrimos a la teoría de la probabilidad para escoger o definir alguna familia paramétrica que modele razonablemente el fenómeno. Una vez hecho esto queda la incertidumbre sobre el parámetro del modelo (no olvidemos que el parámetro puede ser un vector) pues de entre todos los elementos de la familia paramétrica $\mathcal{P} = \{p(x \mid \theta) : \ \theta \in \Theta\}$ ¿Cuál utilizamos para hacer inferencias?

La Estadística Bayesiana modela la incertidumbre que tenemos sobre θ probabilísticamente, esto es, consideramos a θ como una variable (o vector) aleatoria (v.a.) con una ***distribución de probabilidad a priori (o inicial)*** $p(\theta)$. Se trata de una distribución basada en experiencia previa

(experiencia de especialistas, datos históricos, etc.) antes de obtener datos muestrales.

Luego procedemos a observar datos (obtención de muestra) $\mathbf{x} := (x_1, \ldots, x_n)$ y combinamos esta información con la distribución a priori mediante la Regla de Bayes y obtenemos una ***distribución de probabilidad a posteriori (o final)*** :

$$p(\theta \,|\, \mathbf{x}) = \frac{p(\mathbf{x}, \theta)}{p(\mathbf{x})} = \frac{p(\mathbf{x} \,|\, \theta) p(\theta)}{\int_\Theta p(\mathbf{x} \,|\, \tilde{\theta}) p(\tilde{\theta})\, d\tilde{\theta}} \tag{2.1}$$

Tenemos que $p(\theta \,|\, \mathbf{x})$ es también una distribución de probabilidad de θ pero que a diferencia de la distribución a priori $p(\theta)$ toma en cuenta tanto la información contemplada en $p(\theta)$ así como la información contenida en los datos observados $\mathbf{x} = (x_1, \ldots, x_n)$. La distribución a posteriori de θ es la base para hacer inferencias sobre θ.

Es importante tener presente que, por un lado, $p(\mathbf{x} \,|\, \theta)$ y $p(\theta)$ son distribuciones de probabilidad, y por otro:

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

es la probabilidad (o densidad) conjunta de la muestra $\mathbf{x} = (x_1, \ldots, x_n)$ observada a partir del vector aleatorio $\mathbf{X} = (X_1, \ldots, X_n)$. Pero lo más importante es estar consciente de que $p(\mathbf{x})$ es constante respecto a θ, por lo que podemos escribir:

$$p(\theta \,|\, \mathbf{x}) \propto p(\mathbf{x} \,|\, \theta) p(\theta) \tag{2.2}$$

Respecto a $p(\mathbf{x} \,|\, \theta) = p((x_1, \ldots, x_n) \,|\, \theta)$ tenemos que se trata de la probabilidad conjunta de la muestra condicional en θ (usualmente llamada *verosimilitud*). En el caso de que los componentes del vector aleatorio $\mathbf{X} = (X_1, \ldots, X_n)$ resulten ser independientes (esto es, observaciones independientes) tenemos que:

$$p(\mathbf{x} \,|\, \theta) = \prod_{j=1}^{n} p(x_j \,|\, \theta)$$

Aunque será hasta el capítulo V en donde veamos a detalle la metodología para la inferencia bayesiana, conviene adelantar un poco al respecto para tener una idea a grosso modo. Bajo un determinado esquema que se discutirá en su momento, podemos proponer como *estimador puntual* de θ :

$$\hat{\theta} := \mathbb{E}(\theta) = \int_{\Theta} \theta p(\theta \,|\, \mathbf{x})\, d\theta$$

Y aún en el caso de que no se cuente con infomación muestral se puede calcular $\hat{\theta}$ utilizando $p(\theta)$ en vez de $p(\theta|\mathbf{x})$. Para hacer *estimación por regiones*, por ejemplo si deseamos calcular la probabilidad de que el vector de parámetros θ pertenezca a una región $A \subset \Theta$:

$$\mathbb{P}(\theta \in A) = \int_{A} p(\theta \,|\, \mathbf{x})\, d\theta$$

Y de igual modo, aún en el caso de que no se cuente con infomación muestral se puede calcular lo anterior utilizando $p(\theta)$ en vez de $p(\theta \,|\, \mathbf{x})$. Cabe aclarar que si $\dim \theta = 1$ las *regiones* son subconjuntos de $\mathbb{R}$ y que un caso particular de estas regiones son los intervalos. En este sentido la estimación por regiones en estadística bayesiana es más general que la estimación por intervalos de la estadística frecuentista. Y ya que estamos dando ideas preliminares de lo que es la inferencia bayesiana podemos introducir a un nivel muy simple cómo se hace el *contraste de k hipótesis*. Supongamos que se desea contrastar las hipótesis:

$$\begin{array}{rl} H_1: & \theta \in \Theta_1 \\ H_2: & \theta \in \Theta_2 \\ \vdots & \vdots \\ H_k: & \theta \in \Theta_k \end{array}$$

Una manera de hacerlo es calcular directamente la probabilidad de cada hipótesis y escoger aquella que tenga la más alta probabilidad (que quede claro que esta es una manera muy simple de hacer contraste de hipótesis, más adelante se verá que este esquema se puede enriquecer mucho mediante el uso de funciones de utilidad), y calcular la probabilidad de una hipótesis H_j puede ser tan simple como:

$$\mathbb{P}(H_j) = \int_{\Theta_j} p(\theta \,|\, \mathbf{x})\, d\theta$$

Sin embargo, muchos son los casos en los que, más que estar interesados en el vector paramétrico θ lo que queremos es describir el comportamiento de observaciones futuras del fenómeno aleatorio en cuestión, esto es, hacer ***predicción***.

Dado un valor de θ, la distribución que describe el comportamiento de la observación futura X es $p(x \mid \theta)$. El problema es que por lo general el valor de θ es desconocido. La estadística frecuentista ataca este problema estimando puntualmente a θ con base en la muestra observada y dicho estimador $\hat{\theta}$ es sustituido en $p(x \mid \theta)$, es decir, utilizan $p(x \mid \hat{\theta})$. Desde la perspectiva bayesiana el modelo $p(x \mid \theta)$ junto con la distribución a priori $p(\theta)$ inducen una distribución conjunta para el vector aleatorio (X, θ) mediante el concepto de probabilidad condicional:

$$p(x, \theta) = p(x \mid \theta)p(\theta)$$

y marginalizando la distribución de probabilidad conjunta anterior obtenemos:

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

Combinando los dos resultados anteriores:

$$p(x) = \int_{\Theta} p(x|\theta)p(\theta)\, d\theta \tag{2.3}$$

A $p(x)$ la denominamos ***distribución predictiva a priori (o inicial)***, y describe nuestro conocimiento acerca de una observación futura X basado únicamente en la información contenida en $p(\theta)$. Nótese que $p(x)$ no depende ya de θ.

Una vez obtenida la muestra, el modelo $p(x \mid \theta)$ y la distribución a posteriori $p(\theta \mid \mathbf{x})$ inducen una distribución conjunta para (X, θ) condicional en los valores observados $\mathbf{x} = (x_1, \ldots, x_n)$:

$$\begin{aligned} p(x, \theta \,|\, \mathbf{x}) &= \frac{p(x, \theta, \mathbf{x})}{p(\mathbf{x})} \\ &= \frac{p(x \,|\, \theta, \mathbf{x}) p(\theta, \mathbf{x})}{p(\mathbf{x})} \\ &= p(x \,|\, \theta, \mathbf{x}) p(\theta \,|\, \mathbf{x}) \\ &= p(x \,|\, \theta) p(\theta \,|\, \mathbf{x}) \end{aligned}$$

En lo inmediato anterior $p(x \,|\, \theta, \mathbf{x}) = p(x \,|\, \theta)$ se justifica por la independencia condicional de X y $\mathbf{X} = (X_1, \ldots, X_n)$ dado θ. Marginalizando la distribución conjunta condicional anterior:

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

Combinando los dos resultados anteriores:

$$p(x \,|\, \mathbf{x}) = \int_\Theta p(x \,|\, \theta) p(\theta \,|\, \mathbf{x}) \, d\theta \tag{2.4}$$

A $p(x \,|\, \mathbf{x})$ la denominamos ***distribución predictiva a posteriori (o final)***, y describe nuestro conocimiento acerca de una observación futura X basado tanto en la información contenida en $p(\theta)$ como en la información muestral $\mathbf{x} = (x_1, \ldots, x_n)$. Nótese nuevamente que $p(x \,|\, \mathbf{x})$ no depende de θ.

Así que para hacer predicción sobre observaciones futuras del fenómeno aleatorio que estemos modelando usamos $p(x)$ o bien $p(x \,|\, \mathbf{x})$, según sea el caso. Y de manera análoga a lo brevemente mencionado sobre inferencia bayesiana, una manera simple de hacer predicción puntual, por ejemplo, de una observación futura X podría ser:

$$\hat{x} := \mathbb{E}(X) = \int_{Ran\, X} x p(x \,|\, \mathbf{x}) \, dx$$

Donde $Ran\, X$ es el rango de la v.a. X. También, una manera de calcular la probabilidad de que una observación futura caiga en un conjunto $A \subset Ran\, X$ sería:

$$\mathbb{P}(\{X \in A\}) = \int_A p(x \,|\, \mathbf{x}) \, dx$$

Y algo análogo para contraste de hipótesis.

Las ecuaciones 2.1, 2.3 y 2.4 constituyen *el modelo general de la estadística bayesiana*. Cualquier problema estadístico tratado bajo el enfoque bayesiano implica la obtención y utilización de las fórmulas mencionadas.

El siguiente ejemplo, si bien adolece de ser muy simple, resulta muy ilustrativo para dos cosas: primero, comenzar a entender por qué es válido modelar nuestra incertidumbre sobre θ probabilísticamente; segundo, para irnos familiarizando con el modelo bayesiano.

Ejemplo 1. Consideremos una urna que contiene dos monedas: una cargada y la otra equilibrada. Supongamos que la moneda cargada está científicamente construida para tener una probabilidad de $\frac{3}{4}$ de que salga águila. Una persona tomará una de las dos monedas de la urna (no necesariamente al azar) y echará un volado con apuesta de por medio. Haremos lo siguiente:

1. Proponer una familia paramétrica para el experimento anterior,
2. proponer una distribución a priori para el parámetro del modelo, tomando especialmente en cuenta que no estamos seguros de que la moneda fue tomada al azar,
3. obtener la distribución predictiva a priori,
4. obtener la distribución a posteriori,
5. obtener la distribución predictiva a posteriori.

En este sencillo ejemplo es fácil identificar que la familia paramétrica Bernoulli es la adecuada:

$$\mathcal{P} = \{Ber(x \mid \theta) : \theta \in \Theta\}$$

donde

$$Ber(x \mid \theta) = \theta^x (1-\theta)^{1-x} \mathbf{1}_{\{0,1\}}(x)$$

Sólo que en este caso el espacio paramétrico se reduce a $\Theta = \{\frac{3}{4}, \frac{1}{2}\}$.

Desde el enfoque bayesiano, nuestra incertidumbre sobre el parámetro θ la modelamos probabilísticamente, es decir, trataremos a θ como variable

aleatoria y en tal caso $Ran\,\theta = \Theta$ y como Θ es numerable entonces tenemos que θ es una variable aleatoria discreta que sólo toma dos valores: $\frac{3}{4}$ o $\frac{1}{2}$.

Sean $\mathbb{P}(\theta = \frac{3}{4}) = \alpha$ y $\mathbb{P}(\theta = \frac{1}{2}) = 1-\alpha$, para algún α entre 0 y 1. Entonces la distribución a priori queda como sigue:

$$p(\theta) = \alpha \mathbf{1}_{\{\frac{3}{4}\}}(\theta) + (1-\alpha)\mathbf{1}_{\{\frac{1}{2}\}}(\theta) \qquad \alpha \in]0,1[$$

La distribución inicial propuesta permite modelar la parte de los supuestos del problema en donde se dijo que se toma una moneda de la urna no necesariamente al azar. En particular es por medio de α que reflejaremos en la distribución a priori nuestro grado de información acerca de cómo fue escogida la moneda de la urna. Así por ejemplo si estamos o nos sentimos seguros de que fue tomada al azar entonces $\alpha = \frac{1}{2}$. Haríamos la misma asignación si carecemos totalmente de información al respecto ya que no habría razón alguna para suponer que alguna de las dos monedas tiene mayor probabilidad que la otra de ser escogida. Y si por alguna razón contamos con cierta información que nos haga pensar que alguna de las monedas tiene mayor probabilidad de ser escogida también por medio de α podemos reflejarlo. Por ejemplo suponiendo que el procedimiento para elegir la moneda de la urna es lanzando un dado y que si sale un seis entonces escogemos la moneda equilibrada. En este caso claramente $\alpha = \frac{5}{6}$. Importante es destacar el hecho de que restringimos a α al intervalo abierto $]0,1[$ ya que si ocurriese que $\alpha = 0$ o bien $\alpha = 1$ entonces querría decir que estamos seguros del valor de θ y en tal caso no tendría sentido hacer inferencias sobre θ.

Por medio de la fórmula (2.3) obtenemos la distribución predictiva a priori:

$$\begin{aligned} p(x) &= \sum_{\theta\in\Theta} p(x\,|\,\theta)p(\theta) \\ &= \alpha\Big(\frac{3}{4}\Big)^x\Big(\frac{1}{4}\Big)^{(1-x)}\mathbf{1}_{\{0,1\}}(x) + (1-\alpha)\Big(\frac{1}{2}\Big)^x\Big(\frac{1}{2}\Big)^{(1-x)}\mathbf{1}_{\{0,1\}}(x) \end{aligned}$$

La expresión anterior se simplifica a:

$$p(x) = \frac{\alpha}{4}\Big(\mathbf{1}_{\{1\}}(x) - \mathbf{1}_{\{0\}}(x)\Big) + \frac{1}{2}\mathbf{1}_{\{0,1\}}(x)$$

es decir:

$$p(1) = \frac{1}{2} + \frac{\alpha}{4} = \text{probabilidad de que salga águila}$$

$$p(0) = \frac{1}{2} - \frac{\alpha}{4} = \text{probabilidad de que salga sol}$$

De lo anterior cabe destacar que si $\alpha \to 1$ (lo cual se interpreta como que nos sentimos muy seguros de que se escogió la moneda cargada) entonces $p(1) \to \frac{3}{4}$, tal cual se esperaría.

Pensando en que se va a hacer una apuesta sobre el resultado del primer volado, para tener un juego justo, definimos la variable aleatoria U como la ganancia/pérdida resultante de apostar en favor de que salga sol:

$$\mathbb{P}[U = u] = p(0)\mathbf{1}_{\{a\}}(u) + p(1)\mathbf{1}_{\{-b\}}(u) \qquad a, b > 0$$

es decir, U es una v.a. que toma el valor $+a$ (ganancia) con probabilidad $p(0)$ o bien el valor $-b$ (pérdida) con probabilidad $p(1)$. Para tener un juego justo se requiere que $\mathbb{E}(U) = 0$:

$$\begin{aligned} \mathbb{E}(U) = 0 &\Leftrightarrow ap(0) - bp(1) = 0 \\ &\Leftrightarrow a = \frac{p(1)}{p(0)}\, b = \frac{2+\alpha}{2-\alpha}\, b \end{aligned}$$

Es decir, que la cantidad justa a apostar en favor de que salga sol debe ser igual a $\dfrac{p(1)}{p(0)}$ veces la cantidad que se apueste en favor de que salga águila. Si bien lo inmediato anterior es más un problema típico de un curso elemental de probabilidad, resultará interesante analizar como se modifica el esquema de apuestas conforme se van lanzando volados, esto es, ante la presencia de información muestral.

Supongamos ahora que ya se escogió una de las monedas de la urna y que se efectuaron n lanzamientos con ella y los resultados de cada lanzamiento se registran como un vector $n-$dimensional de unos y ceros $\mathbf{x} := (x_1, \ldots, x_n)$. La información contenida en la muestra observada $\mathbf{x}$ modifica nuestra incertidumbre sobre θ pasando de la distribución a priori $p(\theta)$ a la distribución a posteriori:

$$p(\theta \,|\, \mathbf{x}) = \frac{p(\mathbf{x} \,|\, \theta)p(\theta)}{p(\mathbf{x} \,|\, \frac{3}{4})p(\frac{3}{4}) + p(\mathbf{x} \,|\, \frac{1}{2})p(\frac{1}{2})}$$

Si resulta razonable suponer que se hacen lanzamientos independientes entonces:

$$\begin{aligned} p(\mathbf{x}\,|\,\theta) &= \prod_{j=1}^{n} p(x_j\,|\,\theta) \\ &= \prod_{j=1}^{n} \theta^{x_j}(1-\theta)^{1-x_j}\mathbf{1}_{\{0,1\}}(x_j) \\ &= \theta^{\sum x_j}(1-\theta)^{n-\sum x_j}\prod_{j=1}^{n}\mathbf{1}_{\{0,1\}}(x_j) \\ &= \theta^{\sum x_j}(1-\theta)^{n-\sum x_j} g(\mathbf{x}) \end{aligned}$$

Por otro lado:

$$p(\mathbf{x}\,|\,\tfrac{3}{4})p(\tfrac{3}{4}) = \frac{3^{\sum x_j}}{4^n}\,\alpha\, g(\mathbf{x}) \qquad p(\mathbf{x}\,|\,\tfrac{1}{2})p(\tfrac{1}{2}) = \frac{1-\alpha}{2^n}\, g(\mathbf{x})$$

De lo anterior:

$$p(\theta\,|\,\mathbf{x}) = \frac{[2(1-\theta)]^n\big(\frac{\theta}{1-\theta}\big)^{\sum x_j}[\alpha\mathbf{1}_{\{\frac{3}{4}\}}(\theta) + (1-\alpha)\mathbf{1}_{\{\frac{1}{2}\}}(\theta)]}{\alpha\big(\frac{3^{\sum x_j}}{2^n} - 1\big) + 1}$$

Si definimos:

$$\nu = \nu(\alpha, \mathbf{x}) := \frac{3^{\sum x_j}}{2^n}\Big(\frac{\alpha}{1-\alpha}\Big)$$

reescribimos $p(\theta\,|\,\mathbf{x})$ como:

$$p(\tfrac{3}{4}\,|\,\mathbf{x}) = \frac{1}{1+\nu^{-1}} \qquad p(\tfrac{1}{2}|\mathbf{x}) = \frac{1}{1+\nu}$$

Supongamos que la moneda con que se lanzarán los volados es tomada de la urna al azar. En este caso tendríamos que $\alpha = \frac{1}{2}$. La probabilidad a priori de que la moneda escogida sea la cargada es:

$$p(\tfrac{3}{4}) = 0.5$$

Se lanza un primer volado y observamos que sale águila. En este caso $n = 1$, $\mathbf{x} = (x_1) = (1)$ y por lo tanto $p(\frac{3}{4}\,|\,(1)) = 0.6$. Es decir, a la luz de la información muestral con la que se cuenta hasta el momento nos vemos obligados a revisar o *actualizar* la probabilidad de que sea la moneda cargada la que

se está utilizando. Cabe destacar que con la información muestral obtenida ahora es más probable que sea la moneda cargada la que se está utilizando. Se podría pensar que no es difícil que salga un águila con una moneda equilibrada pero el que haya salido águila es evidencia más a favor de que se esté usando la moneda cargada que la equilibrada.

Ahora efectuamos un segundo lanzamiento con la moneda en cuestión y resulta que obtenemos un sol. Ahora $n = 2$, $\mathbf{x} = (x_1, x_2) = (1, 0)$ y obtenemos $p(\frac{3}{4} \,|\, (1, 0)) = 0.4286$. Es decir, a la luz de la información muestral con la que se cuenta hasta el momento nos vemos obligados a *actualizar* nuevamente la probabilidad de que sea la moneda cargada la que se está utilizando. Cabe destacar que con la información muestral obtenida hasta ahora es más probable que sea la moneda equilibrada la que se está utilizando. Se podría pensar que no es difícil que salga un águila y luego un sol con una moneda cargada pero el que haya salido águila y luego sol es evidencia más a favor de que se esté usando la moneda equilibrada que la cargada.

Podría pensarse que nos la podemos pasar así oscilando de un valor a otro a capricho de los resultados muestrales, pero no es así pues conforme el tamaño de la muestra n crece el valor de $p(\frac{3}{4} \,|\, \mathbf{x})$ se va estabilizando:

$$\begin{aligned} \text{Si } n \to \infty \quad &\Rightarrow \quad \sum x_j \to \tfrac{3}{4}\, n \quad \text{o} \quad \sum x_j \to \tfrac{1}{2}\, n \\ &\Rightarrow \quad \nu \to \infty \quad \text{o} \quad \nu \to 0 \\ &\Rightarrow \quad p(\tfrac{3}{4} \,|\, \mathbf{x}) \to 1 \quad \text{o} \quad p(\tfrac{3}{4} \,|\, \mathbf{x}) \to 0 \end{aligned}$$

Lo anterior quiere decir que conforme el tamaño de la muestra se vuelve más y más grande iremos acumulando información que irá reduciendo nuestra incertidumbre respecto a θ (es decir, nuestra incertidumbre respecto a qué moneda se está usando) hasta llegar a un nivel muy cercano a la certeza.

El siguiente es el resultado de una simulación del presente ejemplo con $n = 20$ y $\alpha = \frac{1}{2}$:

$$\mathbf{x} = (0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$$

$$\Rightarrow p(\tfrac{3}{4} \,|\, \mathbf{x}) = 0.9762$$

Y de hecho, efectivamente resultó ser la moneda cargada la que se estaba utilizando.

La información contenida en la muestra observada $\mathbf{x}$ modifica nuestra incertidumbre sobre un siguiente lanzamiento X_{n+1} pasando de la distribución predictiva a priori $p(x)$ a la distribución predictiva a posteriori:

$$\begin{aligned} p(x \,|\, \mathbf{x}) &= p(x \,|\, \tfrac{3}{4}) p(\tfrac{3}{4} \,|\, \mathbf{x}) + p(x \,|\, \tfrac{1}{2}) p(\tfrac{1}{2} \,|\, \mathbf{x}) \\ &= \frac{3^x \nu + 2}{4(\nu+1)} \mathbf{1}_{\{0,1\}}(x) \end{aligned}$$

Regresando al esquema de la apuesta justa, ante la presencia de información muestral $\mathbf{x} = (x_1, \ldots, x_n)$ es necesario ir revisando o *actualizando* los cálculos para determinar cuál sería la apuesta justa, esto es, sin tener información muestral y contando tan solo con la información inicial de que la moneda fue escogida al azar teníamos que la apuesta justa en favor de que salga sol era:

$$a = \frac{p(1)}{p(0)}\, b$$

Esto es, que la apuesta en favor de que salga sol debe ser $\dfrac{p(1)}{p(0)}$ veces la apuesta a favor de que salga águila. Después de observar el resultado x_1 de un primer volado la apuesta justa para el segundo volado se debe actualizar a:

$$a = \frac{p(1|x_1)}{p(0|x_1)}\, b$$

Y después de n volados la apuesta justa para el $(n+1)-$ésimo volado se debe actualizar a:

$$a = \frac{p(1|(x_1, \ldots, x_n))}{p(0|(x_1, \ldots, x_n))}\, b \qquad \diamond$$

§ EJERCICIOS

1. Sea $p(x \mid \theta)$ el modelo paramétrico Bernoulli con parámetro deconocido θ y supongamos que la información a priori sobre θ está dada por $p(\theta) = \text{Beta}(\theta \mid \alpha, \beta)$, con α y β conocidos. Suponiendo observaciones muestrales independientes, obtenga la distribución a posteriori de θ, así como las predictivas a priori y posteriori.

2. Sea $p(x \mid \theta) = \text{Unif}(x \mid 0, \theta)$ con θ desconocido y con distribución a priori sobre θ dada por:
$$p(\theta) = 2(\theta - 1)\mathbf{1}_{[1,2]}(\theta)$$
   a) Obtener la distribución predictiva a priori y graficarla.
   b) Obtener la distribución a posteriori de θ así como la predictiva a posteriori, en ambos casos para tamaño de muestra $n \geq 3$.
   c) Obtener nuevamente la distribución predictiva a priori y graficarla, pero ahora utilizando:
$$p(\theta) = 2(2 - \theta)\mathbf{1}_{[1,2]}(\theta)$$
   y compara con lo obtenido en el inciso a).
   d) Utilizando lo obtenido en los incisos a), b) y c) calcula $\hat{\theta} = \mathbb{E}(\theta)$ sin y con información muestral.
   e) Calcula $\mathbb{P}(\theta < \frac{3}{2})$ sin y con información muestral.

3. Calcula la distribución a posteriori del parámetro en cuestión así como las predictivas a priori y a posteriori para los siguientes casos:
   a) $p(x \mid \lambda) = \text{Poisson}(x \mid \lambda)$ con λ desconocida, $p(\lambda) = \text{Gamma}(\lambda \mid \alpha, \beta)$, con α y β conocidos.
   b) $p(x \mid \theta) = \text{Unif}(x \mid 0, \theta)$ con θ desconocida, $p(\theta) = \text{Pareto}(\theta \mid \alpha, \beta)$, con α y β conocidos.

4. Considera una urna con bolas del mismo tamaño y numeradas de la 1 a la N, donde N es desconocido. Sea la variable aleatoria $X \sim \text{Poisson}(\lambda)$, con λ desconocida y sea $N := X + 1$. Se cuenta con la información a priori (inicial) de que el valor más probable para N es un valor k, conocido. Obtener:

a) La distribución a priori de N.

b) La distribución predictiva a priori.

c) La distribución a posteriori de N para una muestra aleatoria de tamaño 1.

d) La distribución predictiva a posteriori para una muestra aleatoria de tamaño 1, suponiendo que las bolas de la muestra son regresadas a la urna antes de hacer predicción.

e) Suponiendo que $k = 3$ y que se tiene la muestra $x_1 = 2$ calcula las probabilidades a priori y a posteriori de que la urna contenga más de 2 bolas y explique por qué una es menor que la otra.

f) Continuando con el inciso anterior, suponiendo que la bola de la muestra se regresa a la urna, calcula las probabilidades a priori y a posteriori de que una bola tomada al azar tenga el número 3.

5. Si los datos muestrales provienen de observaciones independientes utilizamos $p(\mathbf{x} \mid \theta) = \prod p(x_j \mid \theta)$, pero si las observaciones no son independientes el modelo general sigue siendo válido, pero en este caso $p(\mathbf{x} \mid \theta) \neq \prod p(x_j \mid \theta)$. Supongamos que tenemos una urna con bolas numeradas de la 1 a la N y que lo único que sabemos sobre N es que es 4 o 5.

   a) Propón y justifica una distribución a priori razonable para N.

   b) Deduce la distribución predictiva a priori y calcula la probabilidad de que una bola tomada al azar tenga el número 5.

   c) Si se va a tomar una muestra de tamaño 2 *sin reemplazo* deduce la distribución a posteriori de N. Luego, suponiendo que la muestra obtenida fueron las bolas 1 y 3 calcula la probabilidad de que haya 5 bolas en la urna sin y con información muestral, explicando el por qué de la diferencia.

   d) Supongamos ahora que las dos bolas de la muestra se regresan a la urna. Deduce la distribución predictiva a posteriori y con base en toda la información disponible calcula la probabilidad de que una bola tomada al azar tenga el número 5 y compara este resultado con el obtenido en el inciso b), explicando el por qué de la diferencia.

6. Sea $\{p_i(\theta)\}_{i=1}^k$ una sucesión de distribuciones de probabilidad sobre θ. Definimos la siguiente distribución de probabilidad sobre θ :

$$p(\theta) := \sum_{i=1}^{k} \alpha_i \, p_i(\theta) \quad , \quad \sum_{i=1}^{k} \alpha_i = 1 \ , \ \alpha_i > 0$$

Sea la familia paramétrica $\mathcal{P} := \{p(x \,|\, \theta) \ : \ \theta \in \Theta\}$. Si utilizamos como distribución a priori a la $p(\theta)$ definida anteriormente, demuestre que la distribución a posteriori de θ se puede expresar también como la combinación lineal convexa:

$$p(\theta \,|\, \mathbf{x}) = \sum_{i=1}^{k} \beta_i \, p_i(\theta \,|\, \mathbf{x})$$

exhibiendo la fórmula general de β_i y $p_i(\theta \,|\, \mathbf{x})$.

Capítulo 3

Información a priori

3.1. Determinación de la distribución a priori

Utilizamos el enfoque subjetivo de la probabilidad mencionado en el capítulo 1 para especificar la distribución a priori $p(\theta)$ con base en la información que se tiene en un momento dado, como puede ser: información histórica, la experiencia de especialistas, etc. La elección de una u otra distribución para modelar nuestro nivel de incertidumbre (o información) sobre θ no resulta crucial, en tanto cualquiera de ellas (o ambas) tengan la capacidad de reflejar la información que se tiene sobre θ.

Ejemplo 2. Una compañía de seguros va a ser objeto de una auditoría por parte de la Comisión Nacional de Seguros y Fianzas (CNSyF). La auditoría consistirá en revisar los expedientes de los asegurados y determinar qué porcentaje de ellos están incompletos. En caso de que dicho porcentaje exceda el 10 % la CNSyF procederá a multar a la compañía de seguros. Antes de que esto suceda, la mencionada compañía decide apoyarse en su área de auditoría interna para darse idea del porcentaje de expedientes incompletos. Supongamos que la cantidad de expedientes es tal que sólo daría tiempo de revisar el 0.75 % de ellos antes de que los audite la CNSyF. Aquí podemos intentar aprovechar la experiencia de los auditores internos de la compañía, formulando algunas preguntas como:

1. De acuerdo a su experiencia y conocimiento de la compañía ¿Alrededor de qué cantidad estiman se ubica el porcentaje de expedientes incompletos? *Respuesta: El año pasado estimamos dicho porcentaje en* 8 %*;*

sin embargo, este año el volumen de ventas ha superado nuestras expectativas y esto generalmente juega un poco en contra en lo que ha eficiencia administrativa se refiere por lo que para este año estimamos que dicho porcentaje estará alrededor del 9 %.

2. ¿Cuáles serían sus escenarios optimista y pesimista para dicho porcentaje? *Respuesta: En el mejor de los casos ubicaríamos dicho porcentaje en* 8 % *y en el peor de los casos vemos difícil que exceda el* 11 %.

3. ¿Qué cantidad de expedientes da tiempo de revisar antes de la auditoría de la CNSyF? *Respuesta:* 150.

Sea θ el porcentaje de expedientes incompletos. Podemos modelar lo anterior mediante la familia paramétrica Bernoulli, ya utilizada en el Ejemplo 1, pero aquí el espacio paramétrico $\Theta =]\,0, 1\,[$. Las respuestas a las preguntas 1 y 2 nos dan idea de la centralidad y dispersión de θ. Modelaremos dicha información mediante:

$$p(\theta) = \text{Beta}(\theta \,|\, \alpha, \beta)$$

Pudimos haber elegido alguna otra distribución, la elección anterior se debe a dos razones: primero, que es una distribución de probabilidad en el intervalo $]\,0, 1\,[$ tal cual la necesitamos; segundo, cuenta con dos parámetros (que llamaremos *hiperparámetros*) que nos permiten controlar de manera amplia la centralidad y dispersión de la distribución. Habremos de traducir la información a priori que se tiene en $p(\theta)$, esto es, a través de los hiperparámetros α y β asignándoles valores que reflejen la información que se tiene.

La respuesta a la pregunta 1 nos permite establecer la siguiente ecuación:

$$\mathbb{E}(\theta) = 0.09$$

y la respuesta a la pregunta 2 la podemos expresar como:

$$\mathbb{P}[\,0.08 < \theta < 0.11\,] = 0.95$$

así que para asignar valores a α y β que reflejen lo anterior basta resolver el siguiente sistema de ecuaciones:

$$\frac{\alpha}{\alpha+\beta} = 0.09$$

$$\int_{0.08}^{0.11} \text{Beta}(\theta \,|\, \alpha, \beta)\, d\theta = 0.95$$

y obtenemos $\alpha = 193.090$ y $\beta = 1952.354$. Basándonos únicamente en la información a priori disponible podemos calcular la probabilidad de que el porcentaje de expedientes incompletos rebase el 10 % :

$$\mathbb{P}[\,\theta > 0.10\,] = \int_{0.10}^{1} \text{Beta}(\theta \,|\, 193.09, 1952.354)\, d\theta = 0.0561$$

De la pregunta 3 tenemos que sólo queda tiempo para revisar 150 expedientes que representan tan solo el 0.75 % de un total de veinte mil expedientes que tiene la compañía por lo que aprovecharemos esta otra fuente de información (información muestral) escogiendo al azar 150 expedientes y obtendremos la información muestral $\mathbf{x} = (x_1, \ldots, x_n)$, en donde $x_j \in \{0, 1\}$ y $x_j = 1$ representa un expediente incompleto. Utilizando el resultado del Ejercicio 1 del Capítulo 2 obtenemos la distribución a posteriori de θ :

$$p(\theta \,|\, \mathbf{x}) = \text{Beta}(\theta \,|\, \alpha + r, \beta + n - r)$$

en donde

$$r := \sum_{j=1}^{n} x_j$$

esto es, r es el número de expedientes incompletos de una muestra aleatoria de tamaño $n = 150$. Ya con toda la información disponible (a priori y muestral) *actualizamos* la probabilidad de que el porcentaje de expedientes incompletos rebase el 10 % :

$$\mathbb{P}[\,\theta > 0.10\,] = \int_{0.10}^{1} \text{Beta}(\theta \,|\, 193.09 + r, 2102.354 - r)\, d\theta \qquad \diamond$$

Al proceso de traducir información a priori en una distribución a priori se le conoce en inglés como *to elicit a prior distribution*. Aunque una traducción de la palabra *elicit* con la misma raíz etimológica no existe (aún) en español, en el resto del texto definimos con este mismo sentido *elicitar*.

En el ejemplo anterior, resultó relativamente sencillo elicitar una distribución a priori para θ, especialmente por el hecho de que la familia paramétrica

es univariada, pero tratándose de modelos multivariados elicitar una distribución a priori puede resultar bastante complicado. De hecho, Lindley (2000) pronostica que uno de los temas más importantes en la investigación estadística del nuevo milenio será el desarrollo de metodologías adecuadas para la asignación de probabilidades [subjetivas], y caso particular de esto es el cómo elicitar una distribución a priori.

Es importante destacar en el ejemplo anterior que al haber elegido una distribución beta para modelar la información a priori sobre θ bajo la familia paramétrica Bernoulli nos arrojó como resultado que la distribución a posteriori sobre θ es también una distribución beta, aunque con hiperparámetros distintos. En ocasiones y bajo ciertas familias paramétricas la elección de ciertas distribuciones a priori trae como consecuencia que la distribución a posteriori del parámetro sea de la misma familia que la distribución a priori (por ejemplo, Ejercicio 3, Capítulo 2), pero esto no siembre es así (Ejercicios 2, 4 y 5 del Capítulo 2). De esto nos ocupamos en la siguiente sección.

3.2. Familias conjugadas

Tanto $p(\theta)$ como $p(\theta \,|\, \mathbf{x})$ son distribuciones de probabilidad sobre θ : la primera sólo incorpora información a priori y la segunda *actualiza* dicha información con la información muestral que se pueda obtener. Si bien dijimos que la elección de una u otra distribución de probabilidad para modelar nuestra incertidumbre sobre θ no resulta crucial en tanto sea factible elicitar con cualquiera de ellas una distribución a priori, resulta conveniente tanto para el análisis como desde un punto de vista computacional el que $p(\theta)$ y $p(\theta \,|\, \mathbf{x})$ pertenezcan a la misma familia.

3.1. Definición. Sea $\mathcal{P} := \{p(x \,|\, \theta) \,:\, \theta \in \Theta\}$ una familia paramétrica. Una clase (o colección) de distribuciones de probabilidad $\mathcal{F}$ es una *familia conjugada* para $\mathcal{P}$ si para todo $p(x \,|\, \theta) \in \mathcal{P}$ y $p(\theta) \in \mathcal{F}$ se cumple que $p(\theta \,|\, \mathbf{x}) \in \mathcal{F}$.

Como ejemplos de lo anterior están los resultados de los Ejercicios 1 y 3 del Capítulo 2. Es inmediato notar que si $p(\theta)$ es conjugada para una familia paramétrica $\mathcal{P}$ entonces las distribuciones predictivas a priori y a posteriori pertenecen a una misma familia de distribuciones $\mathcal{F}'$.

A continuación se presentan algunos modelos paramétricos univariados con sus respectivas familias conjugadas:

Cuadro 3.1: Algunas familias conjugadas

Fam. paramétrica	Fam. conjugada
Bernoulli(θ)	Beta $(\theta \mid \alpha, \beta)$
Poisson (λ)	Gamma $(\lambda \mid \alpha, \beta)$
Geométrica (θ)	Beta $(\theta \mid \alpha, \beta)$
Exponencial (λ)	Gamma $(\lambda \mid \alpha, \beta)$
Uniforme $(0, \theta)$	Pareto $(\theta \mid \alpha, \beta)$
Normal (μ)	Normal $(\mu \mid \mu_0, \lambda_0)$
Normal (λ)	Gamma $(\lambda \mid \alpha, \beta)$
Normal (μ, λ)	Normal-Gamma $(\mu, \lambda \mid \mu_0, n_0, \alpha, \beta)$

En lo anterior, para el caso de la Normal usamos como λ el inverso de la varianza y por ello la llamamos *precisión*. Se hace este cambio para utilizar la distribución gamma en vez de la gamma invertida.

Ejemplo 3. Sea la familia paramétrica $\mathcal{P} := \{\text{Poisson}(x \mid \lambda) : \lambda \in \mathbb{R}^+\}$. Si utilizamos como distribución a priori $p(\lambda) \in \mathcal{F} := \{Gamma(\lambda \mid \alpha, \beta) : \alpha, \beta \in \mathbb{R}^+\}$ entonces para una muestra aleatoria $\mathbf{x} = (x_1, \ldots, x_n)$:

$$p(\lambda \mid \mathbf{x}) = \text{Gamma}(\lambda \mid \alpha + r, \beta + n)$$

en donde

$$r := \sum_{j=1}^{n} x_j$$

y además

$$p(x) = \text{Pg}(x \mid \alpha, \beta, 1)$$

$$p(x \mid \mathbf{x}) = \text{Pg}(x \mid \alpha + r, \beta + n, 1)$$

en donde Pg se refiere a la distribución Poisson-gamma:

$$\text{Pg}(x \,|\, \alpha, \beta, n) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha + x)}{x!} \frac{n^x}{(\beta + n)^{\alpha+x}} \mathbf{1}_{\{0,1,\ldots\}}(x)$$

cuya esperanza y varianza están dadas por $\mathbb{E}(X) = n\frac{\alpha}{\beta}$ y $\mathbb{V}(X) = \frac{n\alpha}{\beta}\left[1 + \frac{n}{\beta}\right]$, respectivamente. $\diamond$

En el ejemplo anterior, α y β son los hiperparámetros de la distribución a priori de λ, y por tanto se les debe asignar valores que reflejen la información a priori que se tenga, como se hizo en el Ejemplo **??**. La distribución a posteriori $p(\lambda \,|\, \mathbf{x})$ es una gamma con parámetros $\alpha + r$ y $\beta + n$, lo cual ilustra cómo se combinan información a priori e información muestral. Aunque se verá a mayor detalle más adelante cómo hacer estimación puntual, en el Capítulo 2 se mencionó que una manera de hacer estimación puntual sobre el parámetro es calculando su esperanza, aprovechando el hecho de que se tiene una distribución de probabilidad sobre λ y en este caso:

$$\hat{\lambda} = \mathbb{E}(\lambda \,|\, \mathbf{x}) = \frac{\alpha + r}{\beta + n}$$

Respecto a lo anterior es importante la siguiente observación. Por simplicidad supongamos por un momento que $\beta = \alpha$. Si la magnitud de α es "muy grande" en comparación con r y n tendremos el caso en que la información a priori tendrá más peso que la información muestral en las inferencias que se realicen con la distribución a posteriori; en el ejemplo anterior se tendría que $\mathbb{E}(\lambda \,|\, \mathbf{x})$ sería aproximadamente igual a 1 y $\mathbb{V}(\lambda \,|\, \mathbf{x})$ aproximadamente igual a α^{-1}, varianza que para valores "grandes" de α se acercaría a cero, lo cual nos hablaría de una distribución cuya densidad (o masa) está muy concentrada alrededor de un punto, en cuyo caso diremos que la distribución a priori es *muy informativa*. Si por el contrario, α es cercana a cero tendremos que la distribución a priori tiene una varianza muy grande y en tal caso diremos que se trata de una distribución a priori *poco informativa*. En el ejemplo anterior se tendría que $\mathbb{E}(\lambda \,|\, \mathbf{x})$ es aproximadamente igual a la media de los datos muestrales, que quiere decir que ante tan poca información a priori las inferencias que se hagan se apoyarán prácticamente en la información que provean los datos muestrales. Esto último es materia de la siguiente sección.

3.3. Distribuciones a priori no informativas

La estadística bayesiana proporciona una metodología que permite combinar de manera consistente información a priori con información experimental (i.e. muestral). Ante esto surge la pregunta de cómo realizar inferencias cuando no se dispone de información a priori, o bien cuando dicha información no se quiere o no se puede utilizar.

El problema quedaría resuelto si pudiésemos determinar una distribución a priori que describa la situación en que los datos experimentales contienen toda la información relevante, en lugar de proporcionar tan sólo parte de ella como sucede cuando se dispone de información a priori. Una forma pragmática (pero incompleta) de atacar este problema sería asignar arbitrariamente una distribución a priori con la única condición de que tenga una varianza "muy grande", con todo lo relativo que esto último puede resultar.

Otra forma sería la siguiente. Supongamos que tenemos un número finito de sucesos inciertos (o hipótesis) $E_1, \ldots, E_k$. Una distribución a priori que describe un estado de ignorancia o carencia de información es la siguiente:

$$\mathbb{P}(E_j) = \frac{1}{k}\mathbf{1}_{\{1,\ldots,k\}}(j)$$

esto es, una distribución uniforme discreta.

Thomas Bayes propuso esta distribución con base en lo que él llamó el *Principio de la Razón Insuficiente*: Si no sabemos cosa alguna sobre $\{E_1, \ldots, E_k\}$ no hay razón para asignarle a alguno de los sucesos inciertos una probabilidad diferente que a los otros.

Sin embargo, este principio no es aplicable en situaciones donde el número de sucesos inciertos no es finito. Volviendo al Ejemplo **??**, supongamos ahora que no se tiene información alguna acerca de la proporción θ de expedientes incompletos. Bajo el principio de la razón insuficiente propondríamos como distribución a priori no informativa para θ :

$$p(\theta) = \mathbf{1}_{]\,0,1\,[}(\theta)$$

es decir, una distribución uniforme continua en $]\,0,1\,[$. Supongamos ahora que más que estar interesados en θ directamente estamos interesados en una función uno-a-uno de θ, digamos $\varphi := -\log\theta$. Si nada sabemos acerca de

θ entonces nada sabemos sobre φ tampoco. Bajo el principio de la razón insuficiente asignaríamos también una distribución uniforme continua para φ, pero aquí aparece el primer problema porque φ toma valores en $]\,0,\infty\,[$. De hecho, resultado de probabilidad elemental es que si θ de distribuye como uniforme continua en $]\,0,1\,[$ entonces φ se distribuye exponencial con parámetro 1, la cual, por ejemplo, asigna mayor probabilidad a valores de φ en un intervalo $]\,0,1\,[$ que en el intervalo $]\,1,2\,[$, violando así el principio de la razón insuficiente, por lo que dicho principio no produce distribuciones a priori consistentes en el sentido de que no resultan ser invariantes ante reparametrizaciones uno-a-uno.

3.4. Regla de Jeffreys

Sólo en el caso en que el espacio parámetrico Θ sea finito el principio de la razón insuficiente provee distribuciones a priori no informativas que son invariantes ante transformaciones uno-a-uno del parámetro (o vector de parámetros). Sin embargo, esto es poco útil ya que por lo general nos enfrentamos a espacios paramétricos infinitos.

Jeffreys (1961) propuso una clase de distribuciones a priori no informativas para el caso de espacios paramétricos infinitos. En términos generales, la construcción de esta clase consiste en buscar simultáneamente invariancia ante transformaciones y proveer la menor información a priori en relación a la información muestral, vía la información de Fisher.

Citaremos algunos resultados de probabilidad para recordar el concepto de la información de Fisher. Las demostraciones se pueden consultar en libros como el de Casella (1990) y el de Lehmann (1998), entre muchos otros.

3.2. Teorema. ***(cota inferior de Crámer-Rao)*** *Sean $X_1,\ldots,X_n$ variables aleatorias con función de densidad conjunta $p(\mathbf{x}\,|\,\theta)$, $\theta\in\Theta\subset\mathbb{R}$. Sea $W(\mathbf{X})=W(X_1,\ldots,X_n)$ cualquier función tal que $\mathbb{E}_\theta(W(\mathbf{X}))$ sea una función diferenciable de θ. Suponiendo que $p(\mathbf{x}\,|\,\theta)=p(x_1,\ldots,x_n\,|\,\theta)$ satisface:*

$$\frac{d}{d\theta}\int_{\mathbb{R}^n}\cdots\int h(\mathbf{x})p(\mathbf{x}\,|\,\theta)\,dx_1\cdots dx_n=\int_{\mathbb{R}^n}\cdots\int h(\mathbf{x})\frac{\partial}{\partial\theta}p(\mathbf{x}\,|\,\theta)\,dx_1\cdots dx_n$$

para cualquier función $h(\mathbf{x})$ *tal que* $\mathbb{E}|h(\mathbf{X})| < \infty$*, entonces se cumple que:*

$$\mathbb{V}_\theta\big(W(\mathbf{X})\big) \geq \frac{\Big[\frac{d}{d\theta}\mathbb{E}_\theta\big(W(\mathbf{X})\big)\Big]^2}{\mathbb{E}_\theta\Big[\big(\frac{\partial}{\partial\theta}\log p(\mathbf{X}\,|\,\theta)\big)^2\Big]}$$

3.3. Corolario. *Si además* $X_1, \ldots, X_n$ *son independientes e idénticamente distribuidas con función de densidad común* $p(x\,|\,\theta)$ *entonces:*

$$\mathbb{V}_\theta\big(W(\mathbf{X})\big) \geq \frac{\Big[\frac{d}{d\theta}\mathbb{E}_\theta\big(W(\mathbf{X})\big)\Big]^2}{n\mathbb{E}_\theta\Big[\big(\frac{\partial}{\partial\theta}\log p(X)\,|\,\theta)\big)^2\Big]}$$

El teorema de Crámer-Rao es válido también para variables aleatorias discretas siempre y cuando sea válido intercambiar diferenciación y sumas, así como que la función de masa de probabilidades $p(x\,|\,\theta)$ sea diferenciable respecto a θ.

Cuando se tiene una muestra aleatoria $\mathbf{X} = (X_1, \ldots, X_n)$, a la cantidad $n\mathbb{E}_\theta\Big[\Big(\dfrac{\partial}{\partial\theta}\log p(X)\,|\,\theta)\Big)^2\Big]$ se le conoce como la *información de Fisher de la muestra* y a la cantidad $\mathbb{E}_\theta\Big[\Big(\dfrac{\partial}{\partial\theta}\log p(X)\,|\,\theta)\Big)^2\Big]$ se le conoce como *información de Fisher por unidad muestral* y la denotamos $I(\theta)$. Para facilitar el cálculo de $I(\theta)$ se tiene el siguiente resultado:

3.4. Lema. *Si además de lo anterior* $p(x\,|\,\theta)$ *satisface:*

$$\begin{aligned}\frac{d}{d\theta}\mathbb{E}_\theta\Big[\frac{\partial}{\partial\theta}\log p(X\,|\,\theta)\Big] &= \int_{\mathcal{X}}\frac{\partial}{\partial\theta}\Big[\Big(\frac{\partial}{\partial\theta}\log p(x\,|\,\theta)\Big)p(x\,|\,\theta)\Big]\,dx \\ &= \int_{\mathcal{X}}\frac{\partial^2}{\partial\theta^2}p(x\,|\,\theta)\,dx\end{aligned}$$

en donde $\mathcal{X} := Ran\,X$*, entonces se cumple:*

$$\mathbb{E}_\theta\Big[\Big(\frac{\partial}{\partial\theta}\log p(X\,|\,\theta)\Big)^2\Big] = -\mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\theta^2}\log p(X\,|\,\theta)\Big]$$

Por lo anterior es común simplemente definir la *información de Fisher* del modelo paramétrico $p(x\,|\,\theta)$ como:

$$I(\theta) := -\mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\theta^2}\log p(X\,|\,\theta)\Big]$$

Con lo anterior, el Corolario 3.3 puede expresarse como:

$$\mathbb{V}_\theta\big(W(\mathbf{X})\big) \geq \frac{\left[\frac{d}{d\theta}\mathbb{E}_\theta\big(W(\mathbf{X})\big)\right]^2}{nI(\theta)}$$

Para el caso de muestras aleatorias, $W(\mathbf{X})$ es lo que se utiliza en estadística frecuentista para estimar θ o alguna función de θ. Recuérdese que bajo el enfoque frecuentista la única fuente de información sobre θ es la muestra aleatoria. En caso de que $\mathbb{E}_\theta[W(\mathbf{X})] = \theta$ (que en estadística frecuentista se dice en tal caso que $W(\mathbf{X})$ es un estimador *insesgado* de θ), la varianza de dicho estimador satisface:

$$\mathbb{V}_\theta\big(W(\mathbf{X})\big) \geq \frac{1}{nI(\theta)}$$

Si se tienen varios estimadores de θ cuya esperanza es justamente θ se prefieren aquellos que tengan menor varianza (i.e. los que sean más informativos), y el mejor será aquél o aquéllos cuya varianza coincida con la cota inferior de Crámer-Rao ya que por dicho teorema es la mínima. Notemos pues que si $I(\theta)$ es "grande" entonces hay posibilidades de obtener un estimador con menor varianza y en tal caso decimos que el modelo paramétrico $p(x \,|\, \theta)$ es muy informativo. Si $I(\theta)$ es pequeño entonces la mínima varianza posible de un estimador de θ será grande y en tal caso diremos que el modelo paramétrico es poco informativo. Como $I(\theta)$ depende justamente de θ entonces la varianza de los estimadores (frecuentistas) de θ dependerá del supuesto "verdadero" valor de θ.

La pregunta natural que surge es cuáles familias paramétricas satisfacen las condiciones del teorema de Crámer-Rao (conocidas usualmente como *condiciones de regularidad*) así como la condición del Lema 3.4. Afortunadamente varias de las conocidas, entre las que se encuentran las pertenecientes a la familia exponencial como son la normal, gamma, beta, lognormal, binomial, Poisson, binomial negativa, entre otras. Distribuciones cuyo rango de la variable aleatoria depende del parámetro no satisfacen dichas condiciones, como es el caso de la uniforme.

Ejemplo 4. Sea $p(x \,|\, \theta) = \text{Binomial}(x \,|\, m, \theta)$ con m fija y $\theta \in \Theta =]\,0, 1\,[$. Es inmediato verificar que:

$$I(\theta) = \frac{m}{\theta(1-\theta)}$$

Observemos que $I(\theta) \to \infty$ conforme $\theta \to 0$ o bien $\theta \to 1$ y que alcanza un mínimo en $\theta = \frac{1}{2}$. Esto quiere decir que para valores de θ cercanos a 0 o a 1 un estimador (muestral) $W(\mathbf{X})$ de mínima varianza se vuelve más informativo, y menos informativo conforme θ se aproxime a $\frac{1}{2}$. $\diamond$

Todo lo anterior fue para el caso en que el espacio paramétrico sea unidimensional, pero para el caso multidimensional se tiene un resultado análogo (ver Lehmann (1998)), sólo que en este caso θ es un vector de parámetros y obtenemos una *matriz de información de Fisher* $\mathbf{I}(\theta) = \|I_{ij}(\theta)\|$ cuyas entradas están dadas por:

$$I_{ij}(\theta) = -\mathbb{E}\Big[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(X \,|\, \theta)\Big]$$

3.5. Definición. Para un modelo paramétrico $p(x \,|\, \theta)$ la *distribución a priori no informativa de Jeffreys* para θ está dada por:

$$p(\theta) \propto \sqrt{I(\theta)}\ , \quad \theta \in \Theta \subset \mathbb{R}$$

En el caso multidimensional se tiene:

$$p(\theta) \propto \sqrt{\det \mathbf{I}(\theta)}$$

En cualquier caso la denotaremos $\pi(\theta)$.

La idea es favorecer los valores de θ para los cuales $I(\theta)$, o en su caso $\det \mathbf{I}(\theta)$, es grande, lo que resta influencia a la distribución a priori dando mayor peso a la información muestral. La raíz cuadrada aparece para que resulte invariante bajo transformaciones uno-a-uno. La anterior *Regla de Jeffreys* tiene la desventaja de que en ocasiones produce *distribuciones impropias* (i.e. no integran a 1) lo cual no es del todo grave si realmente lo que se quiere es trabajar con una distribución a posteriori $\pi(\theta \,|\, \mathbf{x})$ que describa la incertidumbre sobre θ basándose únicamente en la información muestral $\mathbf{x} = (x_1, \ldots, x_n)$, por lo que bastará que se cumpla la condición:

$$\int_\Theta p(\mathbf{x} \,|\, \tilde{\theta})\, \pi(\tilde{\theta})\, d\tilde{\theta} < \infty$$

para que la distribución a posteriori sea en efecto una distribución de probabilidad sobre θ :

$$\pi(\theta \,|\, \mathbf{x}) = \frac{p(\mathbf{x} \,|\, \theta)\, \pi(\theta)}{\int_\Theta p(\mathbf{x} \,|\, \tilde{\theta})\, \pi(\tilde{\theta})\, d\tilde{\theta}}$$

3.6. Lema. *La distribución a priori no informativa de Jeffreys* $\pi(\theta) \propto \sqrt{I(\theta)}$ *es invariante ante transformaciones uno-a-uno, esto es, si* $\varphi = \varphi(\theta)$ *es una transformación uno-a-uno de* θ *entonces la distribución a priori de* φ *es* $p(\varphi) \propto \sqrt{I(\varphi)}$.

Demostración. Sea $\varphi = \varphi(\theta)$ una transformación uno-a-uno de θ. Entonces:

$$\frac{\partial \log p(X \mid \theta)}{\partial \varphi} = \frac{\partial \log p(X \mid \varphi(\theta))}{\partial \theta} \frac{\partial \theta}{\partial \varphi}$$

en donde $\theta = \theta(\varphi)$ es la inversa de la transformación φ. Para obtener la información de Fisher de φ calculamos:

$$\frac{\partial^2 \log p(X \mid \varphi)}{\partial \varphi^2} = \frac{\partial \log p(X \mid \varphi(\theta))}{\partial \theta} \frac{\partial^2 \theta}{\partial \varphi^2} + \frac{\partial^2 \log p(X \mid \varphi(\theta))}{\partial \theta^2} \Big(\frac{\partial \theta}{\partial \varphi} \Big)^2$$

Multiplicando ambos miembros por -1 y calculando esperanza respecto a $p(x \mid \theta)$:

$$I(\varphi) \;=\; -\mathbb{E}\left[\frac{\partial \log p(X \mid \theta)}{\partial \theta} \right] \frac{\partial^2 \theta}{\partial \varphi^2} + I(\theta) \Big(\frac{\partial \theta}{\partial \varphi} \Big)^2$$

pero tenemos que:

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial \log p(X \mid \theta)}{\partial \theta} \right] &= \mathbb{E}\left[\frac{\frac{\partial}{\partial \theta} p(X \mid \theta)}{p(X \mid \theta)} \right] \\
&= \int_{-\infty}^{\infty} \frac{\frac{\partial}{\partial \theta} p(x \mid \theta)}{p(x \mid \theta)} p(x \mid \theta)\, dx \\
&= \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} p(x \mid \theta)\, dx \\
&= \frac{d}{d\theta} \int_{-\infty}^{\infty} p(x \mid \theta)\, dx \quad \text{(por las condiciones de regularidad)} \\
&= \frac{d}{d\theta}(1) \;=\; 0
\end{aligned}$$

por lo que:

$$I(\varphi) = I(\theta) \Big(\frac{\partial \theta}{\partial \varphi} \Big)^2$$

esto es:

$$\sqrt{I(\varphi)} = \sqrt{I(\theta)}\, |\partial \theta / \partial \varphi|$$

pero $|\partial\theta/\partial\varphi|$ es el valor absoluto del jacobiano de la transformación inversa por lo que si $\pi(\theta) \propto \sqrt{I(\theta)}$ entonces:

$$p(\varphi) \propto \sqrt{I(\theta(\varphi))}\,|\partial\theta/\partial\varphi| = \sqrt{I(\varphi)}$$

y por lo tanto la distribución a priori de Jeffreys es invariante ante transformaciones uno-a-uno. □

3.7. Corolario. *El mismo resultado es válido para el caso en que θ tiene un espacio paramétrico multidimensional. (Ver Lehmann (1998)).*

Ejemplo 5. Utilizando el resultado del Ejemplo 4 con $m = 1$ obtenemos la distribución a priori de Jeffreys para la familia paramétrica Bernoulli:

$$\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$$

esto es, el *kernel* de $\pi(\theta)$ corresponde a una distribución beta por lo que en este caso $\pi(\theta) = \text{Beta}(\theta \,|\, \frac{1}{2}, \frac{1}{2})$ y como la distribución beta es conjugada de la familia Bernoulli, del Ejercicio 1 del Capítulo 2 tenemos que la distribución a posteriori de θ es $\pi(\theta \,|\, (x_1, \ldots, x_n)) = \text{Beta}(\theta \,|\, \frac{1}{2} + r, \frac{1}{2} + n - r)$ donde $r := \sum x_j$. ◇

El siguiente es un ejemplo en el que la distribución a priori de Jeffreys es impropia; sin embargo, esto no es del todo relevante ya que lo que se busca es que las inferencias se apoyen exclusivamente en la información muestral que se obtenga por lo que lo importante será que las distribuciones a posteriori sean *propias* (i.e. que integren a 1):

Ejemplo 6. Consideremos la familia paramétrica Poisson$(x \,|\, \lambda)$, en donde $\lambda \in \mathbb{R}^+$. Es inmediato verificar que la información de Fisher está dada por:

$$I(\lambda) = \frac{1}{\lambda}$$

y por lo tanto la distribución a priori de Jeffreys es $\pi(\lambda) \propto \lambda^{-1/2}$ la cual resulta ser impropia. Sin embargo, la distribución a posteriori de λ :

$$\pi(\lambda \,|\, \mathbf{x}) \propto p(\mathbf{x} \,|\, \lambda)\,\pi(\theta) \propto e^{-n\lambda}\lambda^{\sum x_j - 1/2}$$

Lo anterior es el kernel de la distribución gamma por lo que $\pi(\lambda \,|\, (x_1, \ldots, x_n)) = \text{Gamma}(\lambda \,|\, \sum x_j + 1/2, n)$. ◇

Además de la Regla de Jeffreys existen otras propuestas para la construcción de distribuciones no informativas, entre las que destacan las *distribuciones de referencia* de Bernardo (1979), Bernardo y Smith (1994).

§ EJERCICIOS

1. Verifique las familias conjugadas del Cuadro 3.1.

2. Un problema que interesa en riesgo de crédito es la posibilidad de que una empresa que emitió títulos de deuda (pagarés, bonos, etc.) para financiarse, incumpla en el pago de dicha obligación al vencimiento de dichos títulos. En primer término, existe incertidumbre en cuanto ha si será o no solvente para regresar el dinero que obtuvo en préstamo en la fecha fijada. En segundo término y en caso de que incurra en incumplimiento, existe también incertidumbre en cuanto al porcentaje de incumplimiento, esto es, puede ocurrir que no pueda cumplir al 100 % con el reembolso pero quizás pueda hacer un pago parcial del $c\,\%$ de la obligación y en tal caso diremos que el incumplimiento fue del $(100 - c)\,\%$, con $0 \leq c \leq 100$. Por lo general cada empresa tiene un perfil particular y no es comparable con otras, y de hecho ni con su propio historial crediticio ya que las condiciones del pasado para una misma empresa suelen ser muy diferentes a las del presente, por lo que se pretende modelar el porcentaje de incumplimiento únicamente con base en la información a priori que proporcione un analista o grupo de analistas de crédito. Supongamos que se cuenta con la siguiente información a priori: los analistas de crédito estiman que la probabilidad de incumplimiento se ubica entre 5 y 15 % y que en caso de que se de el incumplimiento el porcentaje de incumplimiento se ubica entre 60 y 100 %. Proponga los modelos adecuados, obtenga la distribución predictiva a priori del porcentaje de incumplimiento y calcule el porcentaje esperado de incumplimiento.

3. Supongamos que la llegada de autos a la caseta de cobro de una autopista los días viernes de 5 a 8 p.m. se puede modelar mediante la familia paramétrica Poisson. Hacemos dos preguntas al encargado de la caseta: ¿Como cuántos autos llegan en promedio por minuto a la caseta? A lo cual nos responde que 5. Tomando en cuenta que el dato

anterior es una apreciación subjetiva ¿Cuál cree usted que sería en el mayor de los casos el número promedio de autos por minuto? A lo cual nos responde que 12.

a) Utilizando una distribución conjugada especifique la distribución a priori del parámetro con base en la información que se tiene. Calcule el valor esperado del parámetro así como la probabilidad de que dicho parámetro sea mayor a 8.

b) Supongamos ahora que procedemos a tomar una muestra aleatoria y obtenemos $\mathbf{x} = (679, 703, 748, 739, 693)$. Obtenga la distribución a posteriori del parámetro y calcule el valor esperado del parámetro así como la probabilidad de que dicho parámetro sea mayor a 8. Compare con el inciso a). Grafique en una misma hoja la distribución a priori, la distribución a posteriori con el primer dato, con los primeros dos y así sucesivamente hasta la a posteriori con los cinco datos.

c) Utilizando la Regla de Jeffreys y la información del inciso anterior obtenga la distribución a posteriori del parámetro y calcule el valor esperado del parámetro así como la probabilidad de que dicho parámetro sea mayor a 8. Compare con el inciso b). Grafique lo análogo al inciso anterior. ¿Qué se puede concluir sobre la información a priori proveniente del encargado de la caseta?

4. Utilizando la Regla de Jeffreys obtenga las distribuciones a posteriori de las siguientes distribuciones uniparamétricas univariadas:

a) Geométrica

b) Exponencial

c) Normal (con precisión conocida)

d) Normal (con media conocida)

Capítulo 4

Elementos de la teoría de la decisión

Revisaremos algunos resultados de la teoría de la decisión que son útiles para hacer inferencias pero no daremos aquí ni la construcción axiomática ni la mayoría de las demostraciones de los resultados que al respecto se utilizarán. Para detalles sobre esto se recomienda ampliamemte el libro de Bernardo y Smith (1994).

Uno de los principales objetivos de la teoría de la decisión es el desarrollo de procesos lógicos para la toma de decisiones bajo condiciones de incertidumbre. La idea es plantear los problemas de inferencia estadística como problemas de decisión, y aprovechar por tanto los resultados que ya se tienen respecto a esto último.

4.1. Representación formal

4.1. Definición. Un *problema de decisión* está definido conjuntamente por los elementos $(\mathcal{E}, \mathcal{C}, \mathcal{A}, \preceq)$ en donde:

1. $\mathcal{E}$ es un álgebra de eventos relevantes que denotaremos mediante E_j ,

2. $\mathcal{A}$ es un conjunto de opciones o acciones potenciales, cuyos elementos denotaremos mediante a_i ,

3. $\mathcal{C}$ es el conjunto de consecuencias posibles y mediante c_{ij} denotaremos la consecuencia de haber elegido la acción $a_i \in \mathcal{A}$ bajo la ocurrencia de el evento $E_j \in \mathcal{E}$,

4. $\preceq$ es una relación (binaria) de *preferencia* para algunos de los elementos de $\mathcal{A}$.

Ejemplo 7. Supongamos que nos enfrentamos al trivial problema de decidir si salimos a la calle con o sin paraguas. Como conjunto de acciones posibles tenemos entonces que $\mathcal{A} := \{a_1, a_2\}$ en donde a_1 puede representar la acción de llevar paraguas y a_2 la acción de no llevarlo. Son muchos los eventos que al respecto podríamos considerar, pero por el momento aceptemos la idea intuitiva de que esencialmente tenemos dos eventos relevantes (para lo que se pretende decidir) : llueve (E_1) o no llueve (E_2). Con lo anterior podemos entonces determinar el conjunto de consecuencias posibles:

$$\mathcal{C} = \{c_{ij} = (a, E) \,:\, a \in \mathcal{A}\,,\, E \in \{E_1, E_2\}\}$$

Así por ejemplo c_{22} es la consecuencia de haber decidido no llevar paraguas y que efectivamente no haya llovido; c_{21} es la consecuencia de haber decido no llevar paraguas y que haya llovido. Intuitivamente podríamos decir que nos gusta más la consecuencia c_{22} que la c_{21} (esto en la mayoría de los casos quizás, porque hay quienes disfutan mojarse). $\diamond$

Resolver un problema de decisión significa determinar $\preceq$, esto es, definir un criterio para decidir qué acciones son preferibles a otras. Hay que notar que $\mathcal{A}$ representa la parte del problema de decisión que controlamos, que está en nuestras manos en un momento dado. $\mathcal{E}$ representa la parte que no controlamos pues se refiere a eventos cuya ocurrencia no depende de nosotros.

En la Definición 4.1 se definió a $\mathcal{E}$ como un *álgebra*. En un problema de decisión tenemos involucrado un fenómeno o experimento aleatorio para la parte que no controlamos. En probabilidad, se denota por Ω el *espacio muestral*, esto es, el conjunto de resultados posibles del fenómeno o experimento aleatorio. Los distintos eventos relacionados a este experimento se pueden identificar como subconjuntos de Ω. Se define un *espacio de eventos* como un conjunto de eventos asociados a un experimento o fenómeno aleatorio, esto es, un espacio de eventos es, en principio, un conjunto de subconjuntos de Ω. Sea $\mathcal{E}$ justamente ese espacio de eventos. Es cuestión de analizar algunos ejemplos sencillos de probabilidad para motivar algunas propiedades que debe tener dicho espacio de eventos:

4.2. Definición. Sea Ω un conjunto arbitrario y sea $\mathcal{E}$ un conjunto de subconjuntos de Ω. Se dice que $\mathcal{E}$ es un *álgebra* si:

1. $\Omega \in \mathcal{E}$,
2. Si $E \in \mathcal{E}$ entonces $E^c \in \mathcal{E}$,
3. Si $E_1, \ldots, E_n \in \mathcal{E}$ entonces $\bigcup_{j=1}^{n} E_j \in \mathcal{E}$.

Es fácil verificar que consecuencia de lo anterior es que el conjunto vacío $\varnothing \in \mathcal{E}$ y que si $E_1, \ldots, E_n \in \mathcal{E}$ entonces $\bigcap_{j=1}^{n} E_j \in \mathcal{E}$. Pero en probabilidad esto no es suficiente, se pide además que la unión (infinito) numerable de eventos también esté en el espacio de eventos, en cuyo caso se le denomina *σ-álgebra*. Sin embargo, más adelante mencionaremos el por qué se asigna a $\mathcal{E}$ una estructura de álgebra y no de σ-álgebra.

En un problema de decisión no trabajamos directamente con todo $\mathcal{E}$ sino con algunos de sus elementos que llamamos *eventos relevantes*, relevantes respecto a lo que se quiere decidir. Por el momento estableceremos que dicho conjunto de eventos relevantes sea finito. También pediremos que sea una partición de Ω. Igualmente pediremos que $\mathcal{A}$ sea finito.

Mencionamos ya que el conjunto de eventos relevantes es la parte que no controlamos del problema de decisión, esto es, tenemos *incertidumbre* respecto a cuál de los eventos relevantes ocurrirá, pero esto no nos impide estudiar científicamente el fenómeno o experimento aleatorio asociado a ellos e intentar reunir información que nos de idea acerca de la posibilidad de ocurrencia de cada uno de los eventos. Al respecto Lindley (2000) menciona que "un enfoque científico [sobre este problema] implica la medición de la incertidumbre ya que, citando a Kelvin, es sólo asociando números a cualquier concepto científico como puede ser adecuadamente entendido. La razón de medir no es sólo para ser más precisos respecto a la noción de que tenemos más incertidumbre acerca de lo que sucederá mañana en el mercado de valores en comparación con que salga el sol, sino también para poder combinar incertidumbres". Lindley (2000) argumenta el por qué la incertidumbre debe ser medida con probabilidad y Bernardo y Smith (1994) hacen una fundamentación rigurosa de ello y de cómo tomar decisiones bajo condiciones de incertidumbre. Esto último se traduce en poder determinar una relación de preferencia $\preceq$ sobre $\mathcal{A}$ y escoger la acción óptima.

Al hablar de una relación (binaria) de preferencia $\preceq$ no estamos suponiendo que cualquier par de acciones $(a_1, a_2) \in \mathcal{A} \times \mathcal{A}$ está necesariamente relacionado mediante $\preceq$. En caso de que dicha relación sea aplicable, mediante $a_1 \preceq a_2$ entenderemos que, bajo algún criterio que se defina, a_1 no es más preferible que a_2.

4.3. Definición. La relación de preferencia $\preceq$ induce las siguientes relaciones binarias para elementos $a_1, a_2 \in \mathcal{A}$:

1. $a_1 \sim a_2$ si y sólo si $a_1 \preceq a_2$ y $a_2 \preceq a_1$ (indiferencia),
2. $a_1 \prec a_2$ si y sólo si $a_1 \preceq a_2$ pero no se cumple que $a_2 \preceq a_1$ (preferencia estricta),
3. $a_1 \succeq a_2$ si y sólo si $a_2 \preceq a_1$,
4. $a_1 \succ a_2$ si y sólo si $a_2 \prec a_1$.

Y así como resulta necesario cuantificar la incertidumbre de algún modo, que en nuestro caso será por medio de probabilidad, también es necesario cuantificar las consecuencias. En el Ejemplo 7 resultó (quizás) intuitivamente claro que la consecuencia c_{22} es preferible a la consecuencia c_{21} pero si nos preguntamos lo mismo respecto a c_{12} y c_{21} posiblemente no resulte tan contundente la respuesta, o al menos no con la intensidad del otro caso. Nótese que de inicio evitamos escribir, por ejemplo, que $c_{22} \succ c_{21}$ porque hemos definido las relaciones de preferencia para elementos de $\mathcal{A}$ y no de $\mathcal{C}$. Claro que podríamos definir relaciones de preferencia análogas para $\mathcal{C}$ y tener cuidado en utilizar una simbología diferente, lo cual no sería práctico, así que utilizaremos los mismos símbolos pero conscientes de que las relaciones de preferencia de $\mathcal{A}$ son distintas a las de $\mathcal{C}$.

4.4. Definición. Entenderemos por *espacio de estados*, y lo denotaremos Θ, a una partición de Ω en eventos relevantes.

Como la incertidumbre sobre los eventos relevantes la mediremos con probabilidad, si se tiene una medida de probabilidad $\mathbf{P} : \mathcal{E} \to [\,0, 1\,]$ entonces tenemos una función de probabilidad $\mathbb{P} : \Theta \to [\,0, 1\,]$. Nótese además que $\mathcal{C} = \mathcal{A} \times \Theta$.

4.5. Definición. Una *función de utilidad* es una función $u : \mathcal{C} \to \mathbb{R}$.

El poder cuantificar de algún modo las distintas consecuencias nos provee de un criterio inmediato para determinar las relaciones de preferencia en $\mathcal{C}$:

4.6. Definición. $c_{ij} \preceq c_{kl}$ si y sólo si $u(c_{ij}) \leq u(c_{kl})$.

La definición anteior induce las siguientes relaciones binarias:

$$c_{ij} \sim c_{kl} \Leftrightarrow c_{ij} \preceq c_{kl} \text{ y } c_{kl} \preceq c_{ij} \,,$$

$$c_{ij} \prec c_{kl} \Leftrightarrow c_{ij} \preceq c_{kl} \text{ pero no se cumple que } c_{kl} \preceq c_{ij} \,.$$

4.2. Solución de un problema de decisión

Dijimos ya que resolver un problema de decisión $(\mathcal{E}, \mathcal{C}, \mathcal{A}, \preceq)$ consiste en determinar $\preceq$, es decir, definir una relación de preferencia entre los elementos de $\mathcal{A}$ y escoger la acción óptima (la más preferible). Existen diversas formas de hacerlo, pero aquí trataremos exclusivamente la forma que nos interesa para hacer inferencias desde el enfoque bayesiano y para ello requerimos tener identificado lo siguiente:

- El espacio de estados Θ,
- una función de probabilidad $\mathbb{P}$ sobre los elementos de Θ,
- una función de utilidad u sobre $\mathcal{C}$.

La función de probabilidad $\mathbb{P} : \Theta \to [\,0, 1\,]$ puede ser *a priori* o *a posteriori*, en el sentido en que se trató en el Capítulo 2. El cómo establecer o construir una función de utilidad dependerá de cada problema particular. Para mayor detalle acerca de algunas formas generales de funciones de utilidad nuevamente insistimos en consultar el libro de Bernardo y Smith (1994). Aquí nos limitaremos a ilustrar lo anterior mediante el siguiente:

Ejemplo 8. Retomando el Ejemplo 1 en la parte referente a la apuesta, podemos plantearlo como un problema de decisión. El fenómeno aleatorio involucrado es el lanzamiento de una moneda por lo que su espacio muestral es $\Omega = \{\text{águila}, \text{sol}\}$ y su álgebra de eventos es $\mathcal{E} = \{\Omega, \varnothing, \{\text{águila}\}, \{\text{sol}\}\}$. El espacio de estados $\Theta = \{E_1, E_2\}$ donde $E_1 := \{\text{águila}\}$ y $E_2 := \{\text{sol}\}$. Nótese que Θ es partición de Ω. El conjunto de acciones es $\mathcal{A} = \{a_1, a_2\}$ donde a_1 representa la acción de apostar en favor de que salga águila y a_2 en favor de

sol. Si el esquema de apuesta consiste en que quien apueste a favor de águila arriesgue b pesos y quien lo haga en favor de sol arriesgue a pesos entonces la función de utilidad queda como sigue:

$$u(c_{11}) = a \qquad u(c_{12}) = -b \qquad u(c_{21}) = -a \qquad u(c_{22}) = b$$

De acuerdo al Ejemplo 1 nótese que $E_1 \equiv \{X = 1\}$ y $E_2 \equiv \{X = 0\}$ por lo que:

$$\mathbb{P}(E_1) = \mathbb{P}(X = 1) \quad \text{y} \quad \mathbb{P}(E_2) = \mathbb{P}(X = 0)$$

para lo cual podemos utilizar la distribución predictiva a priori o a posteriori, según sea el caso, es decir, $\mathbb{P}(X = x) = p(x)$ o bien $\mathbb{P}(X = x) = p(x \,|\, \mathbf{x})$, y con esto queda definida una función de probabilidad $\mathbb{P}$ sobre el espacio de estados (relevantes) Θ. De manera tabular podemos resumir lo anterior como sigue:

$\mathbb{P}(E_j)$	$\mathbb{P}(E_1)$	$\mathbb{P}(E_2)$
$u(a_i, E_j)$	E_1	E_2
a_1	a	$-b$
a_2	$-a$	b

En el Ejemplo 1 se obtuvo la relación que debe existir entre los montos de apuesta a y b pesos para tener una *apuesta justa* y dicha relación se obtuvo a partir de la ecuación:

$$a\mathbb{P}(E_1) - b\mathbb{P}(E_2) = 0$$

El tener una apuesta o juego justo implica que seamos indiferentes respecto a apostar en favor de cualquiera de las opciones disponibles, que en términos de este problema de decisión lo escribimos como $a_1 \sim a_2$. Pero eso es tan sólo un caso particular. De forma más general podemos definir las variables aleatorias:

$$U_1 := \; a\mathbf{1}_{E_1} - b\mathbf{1}_{E_2}$$
$$U_2 := -a\mathbf{1}_{E_1} + b\mathbf{1}_{E_2}$$

esto es, U_1 representa la ganancia/pérdida que obtendrá quien decida tomar la acción a_1 y U_2 lo análogo para la acción a_2. Calculando sus esperanzas:

$$\mathbb{E}(U_1) := \; a\mathbb{P}(E_1) - b\mathbb{P}(E_2)$$

$$\mathbb{E}(U_2) := -a\mathbb{P}(E_1) + b\mathbb{P}(E_2)$$

Entonces para tener un juego justo se requiere que $\mathbb{E}(U_1) = 0$ y que $\mathbb{E}(U_2) = 0$, que de hecho tienen la misma solución, por lo que tendremos que $a_1 \sim a_2$ si $\mathbb{E}(U_1) = \mathbb{E}(U_2)$. Si por el contrario ocurriera que $\mathbb{E}(U_1) > \mathbb{E}(U_2)$ entonces si nos dan a escoger preferimos la acción a_1, esto es, $a_1 \succ a_2$. $\mathbb{E}(U_i)$ es lo que se conoce como la *utilidad esperada de la acción* a_i y es justamente lo que nos servirá como criterio para definir una relación de preferencia $\preceq$ sobre $\mathcal{A}$ y poder así elegir la acción óptima. $\diamond$

4.7. Definición. En un problema de decisión $(\mathcal{E}, \mathcal{C}, \mathcal{A}, \preceq)$ con espacio de estados (relevantes) $\Theta = \{E_1, \ldots, E_m\} \subset \mathcal{E}$, función de probabilidad $\mathbb{P}$ sobre Θ y función de utilidad u sobre $\mathcal{C}$, la *utilidad esperada de la acción* $a_i \in \mathcal{A} = \{a_1, \ldots, a_k\}$ se denota $\overline{u}(a_i)$ y se define como:

$$\overline{u}(a_i) := \sum_{j=1}^{m} u(a_i, E_j)\, \mathbb{P}(E_j) \qquad i = 1, \ldots, k$$

4.8. Definición. (Criterio general de decisión). En un problema de decisión como el de la Definición 4.7, la relación de preferencia $\preceq$ sobre $\mathcal{A}$ queda definida por:

$$a_1 \preceq a_2 \Leftrightarrow \overline{u}(a_1) \leq \overline{u}(a_2)$$

Estrictamente hablando, lo anterior no es una definición sino una proposición que se demuestra después de una rigurosa axiomatización de lo que hemos visto hasta el momento como lo desarrollan Bernardo y Smith (1994), pero como aquí nos limitamos a dar la motivación intuitiva para establecer dicho criterio, no quedó más remedio que definirlo así.

A partir de la Definición 4.8 es inmediato que:

$$a_1 \sim a_2 \Leftrightarrow \overline{u}(a_1) = \overline{u}(a_2)$$

$$a_1 \prec a_2 \Leftrightarrow \overline{u}(a_1) < \overline{u}(a_2)$$

Finalmente, el criterio que utilizaremos para elegir la acción óptima de $\mathcal{A}$, misma que denotaremos a_*, será aquélla que satisfaga:

$$\overline{u}(a_*) = \max_i \overline{u}(a_i)$$

Puede ocurrir que a_* no sea única. En tal caso hablaríamos entonces de el *conjunto de acciones óptimas* $\mathcal{A}^* \subset \mathcal{A}$ y en tal caso diremos que somos

indiferentes ante llevar acabo cualquiera de las acciones de $\mathcal{A}^*$. Es en este punto donde podemos retomar el por qué pedimos que tanto $\mathcal{A}$ como $\mathcal{E}$ sean finitos, pues de ser así, los resultados que desarrollan Bernardo y Smith (1994) garantizan que a_* existe, de otro modo puede o no ocurrir así. Necesitamos un par de definiciones más para ilustrarlo.

4.9. Definición. Una acción a_{i_1} está *dominada* por otra acción a_{i_2} si para todo j tenemos que $u(a_{i_1}, E_j) \leq u(a_{i_2}, E_j)$ y además existe un j_0 tal que $u(a_{i_1}, E_{j_0}) < u(a_{i_2}, E_{j_0})$.

4.10. Definición. Una acción es *admisible* si no existe otra acción que la domine. Una acción es *inadmisible* si existe al menos otra que la domine.

Lo anterior nos dice en pocas palabras que es (quizás) posible depurar el espacio de acciones, esto es, habría que eliminar de $\mathcal{A}$ las acciones inadmisibles, llamadas así porque, independientemente del evento que ocurra, siempre existe una mejor opción.

Ejemplo 9. Supongamos que una operadora de fondos de inversión nos ofrece cuatro tipos diferentes de sociedades de inversión, esto es, cuatro diferentes estrategias para invertir nuestro dinero. Por simplicidad supongamos que los cuatro portafolios de inversión de dichas sociedades invierten en dos opciones: acciones de la empresa ABC que cotiza en bolsa y en títulos que pagan un rendimiento fijo de 6 %. Lo que distingue a cada portafolios es el porcentaje destinado a una y otra opción de inversión:

portafolios	% en ABC	% a tasa fija
agresivo	80	20
moderado	50	50
conservador	20	80
sin riesgo	0	100

De acuerdo a lo anterior, el rendimiento de cada portafolios se desconoce (a excepción del último) ya que depende del rendimiento que tenga la empresa ABC y éste resulta incierto; sin embargo, podemos modelar nuestra incertidumbre respecto al rendimiento de ABC consultando a un analista financiero y pidiéndole (por ejemplo) nos de los escenarios posibles acerca del rendimiento que ABC tendrá de acuerdo a la información que él maneja.

Por simplicidad supongamos que nos plantea los siguientes escenarios con sus respectivas probabilidades:

Escenario	Rendimiento	Probabilidad
pesimista	$-5\,\%$	0.20
realista	$+15\,\%$	0.60
optimista	$+25\,\%$	0.20

Las probabilidades asignadas constituyen lo que ya hemos definido como *probabilidad a priori*. De acuerdo a lo anterior podemos pensar en una variable aleatoria X que represente el rendimiento de ABC y por tanto $\operatorname{Ran} X = \{-5\,\%, +15\,\%, +25\,\%\}$ con las probabilidades arriba señaladas.

El problema de decisión consiste justamente en elegir de entre los cuatro portafolios el óptimo de acuerdo a la información con que se cuenta. Sea el conjunto de acciones $\mathcal{A} := \{a_1, a_2, a_3, a_4\}$ y el espacio de estados (relevantes) $\Theta := \{E_1, E_2, E_3\}$ de modo que:

$$\begin{aligned} a_1 &\equiv \text{ invertir en el portafolios agresivo} \\ a_2 &\equiv \text{ invertir en el portafolios moderado} \\ a_3 &\equiv \text{ invertir en el portafolios conservador} \\ a_4 &\equiv \text{ invertir en el portafolios sin riesgo} \end{aligned}$$

$$\begin{aligned} E_1 &\equiv \{X = -5\,\%\} \\ E_2 &\equiv \{X = +15\,\%\} \\ E_3 &\equiv \{X = +25\,\%\} \end{aligned}$$

Para poder resolver este problema de decisión sólo nos falta especificar una función de utilidad para las distintas consecuencias. Por el momento consideremos una función de utilidad igual al rendimiento que puede obtener cada portafolios bajo cada escenario:

$$u(a_i, E_j) := \alpha_i x_j + (1 - \alpha_i) r$$

en donde α_i representa el porcentaje de inversión bajo la acción a_i y x_j representa el rendimiento de ABC bajo el escenario E_j y r la tasa fija del

6 % en este caso. Utilizando el criterio de la utilidad esperada máxima:

$\mathbb{P}(E)$	0.20	0.60	0.20	
$u(a, E)$	E_1	E_2	E_3	$\overline{u}(a_i)$
a_1	$-2.8\,\%$	$13.2\,\%$	$21.2\,\%$	$11.6\,\%$
a_2	$0.5\,\%$	$10.5\,\%$	$15.5\,\%$	$9.5\,\%$
a_3	$3.8\,\%$	$7.8\,\%$	$9.8\,\%$	$7.4\,\%$
a_4	$6.0\,\%$	$6.0\,\%$	$6.0\,\%$	$6.0\,\%$

Claramente la acción óptima es $a_* = a_1$. Cabe aclarar que a_1 es la acción óptima si realmente la función de utilidad propuesta corresponde a nuestras preocupaciones como inversionistas. De hecho, veremos rápidamente que, salvo casos extremos, la función de utilidad propuesta no es una buena elección para este problema específico. Suponiendo que tenemos la libertad de elegir libremente los porcentajes de inversión en las opciones ya mencionadas tendremos entonces que el conjunto de acciones es $\mathcal{A} = [\,0, 100\,\%\,]$, esto es, existe una infinidad no numerable de porcentajes distintos que podemos asignar a cada opción inversión (con la condición de que sumen $100\,\%$) y en tal caso $a \in \mathcal{A}$ representa la acción de invertir $a\,\%$ en ABC y el resto a tasa fija por lo que:

$$\overline{u}(a) = a\mathbb{E}(X) + (1-a)r \quad , a \in \mathcal{A}$$

y como $\mathbb{E}(X) = 13\,\%$ entonces reescribimos:

$$\overline{u}(a) = (7\,\%)a + 6\,\%$$

es decir, $\overline{u}(a)$ es la ecuación de una recta con pendiente positiva y alcanza su máximo en $a = 100\,\%$ por lo que la acción óptima sería en este caso $a_* = 100\,\%$ con una utilidad (rendimiento esperado en este caso) $\overline{u}(a_*) = 13\,\%$. Entonces, con la función de utilidad propuesta, la acción óptima es tomar el mayor riesgo posible: invertir el $100\,\%$ en ABC. Aquí es donde un inversionista mínimamente informado protestaría con semejante decisión. ¿Qué sucede? Pues que normalmente un inversionista considera, al menos, dos aspectos: rendimiento y *riesgo* de la inversión. Normalmente un inversionista busca altos rendimientos pero con el menor riesgo posible y de hecho la decisión de inversión bajo esta doble consideración implica balancear entre el rendimiento que quiere el inversionista y el riesgo que está dispuesto a tomar porque las inversiones de poco riesgo van acompañadas de rendimientos moderados y las inversiones que tienen la posibilidad de otorgar altos rendimientos van

acompañadas de mayor riesgo. Así que no quiere decir esto que esté mal la teoría, simplemente que hay que tener cuidado con la elección de una función de utilidad que refleje todo aquello que nos preocupe o interese sea tomado en cuenta. La función de utilidad que se propuso únicamente toma en cuenta rendimientos mas no riesgo.

Construiremos pues una función de utilidad (entre muchas que se podrían definir) que de algún modo refleje preocupación tanto en rendimientos altos como en controlar la cantidad de riesgo que se toma. Sea $u_{ij} := u(a_i, E_j)$ como se definió anteriormente y sea:

$$u_{i\bullet} := \frac{1}{m}\sum_{j=1}^{m} u_{ij}$$

Definimos la siguiente función de utilidad w :

$$\begin{aligned} w(a_i, E_j) &:= u_{ij} - A(u_{i\bullet} - u_{ij})^2 \\ &= u_{ij} - \alpha_i^2 A(x_\bullet - x_j)^2 \end{aligned}$$

en donde $x_\bullet := \frac{1}{m}\sum_{j=1}^{m} x_j$ y en donde $A \geq 0$ es lo que en ocasiones se denomina un *coeficiente de aversión al riesgo*. Nótese que si $A = 0$ entonces $w(a_i, E_j) = u(a_i, E_j)$ por lo que nos ocuparemos sólo del caso en que $A > 0$. La utilidad esperada para cada acción $a_i \in \mathcal{A}$ es:

$$\begin{aligned} \overline{w}(a_i) &= \overline{u}(a_i) - \alpha_i^2 A \sum_{j=1}^{m} (x_\bullet - x_j)^2 \, \mathbb{P}(E_j) \\ &= \alpha_i \big[\mathbb{E}(X) - r\big] + r - \alpha_i^2 A \big[\mathbb{V}(X) + (x_\bullet - \mathbb{E}(X))^2\big] \end{aligned}$$

Nótese cómo ahora la fórmula general para la utilidad esperada de una acción está tanto en términos de el valor esperado del rendimiento de ABC, esto es $\mathbb{E}(X)$, como del riesgo o dispersión de dicho rendimiento, es decir, $\mathbb{V}(X)$ en este caso. Y de hecho, es inmediato a partir de lo anterior obtener la fórmula general de la utilidad esperada por acción para el caso en que $a \in \mathcal{A} = [\,0, 100\,\%\,]$:

$$\overline{w}(a) = a\big[\mathbb{E}(X) - r\big] + r - a^2 A\big[\mathbb{V}(X) + (x_\bullet - \mathbb{E}(X))^2\big]$$

y para encontrar el valor de a que maximiza $\overline{w}(a)$ resolvemos:

$$\overline{w}\,'(a) = \mathbb{E}(X) - r - 2aA\big[\mathbb{V}(X) + (x_\bullet - \mathbb{E}(X))^2\big] = 0$$

y como $\overline{w}''(a) < 0$ entonces el valor de a que maximiza $\overline{w}(a)$, esto es el porcentaje óptimo de inversión en ABC (invirtiendo el resto a tasa fija):

$$a_* = \frac{\mathbb{E}(X) - r}{2A\big[\mathbb{V}(X) + (x_\bullet - \mathbb{E}(X))^2\big]}$$

Para analizar el resultado anterior definamos $\rho_X := \mathbb{E}(X) - r$ y $\delta_X := \mathbb{V}(X) + (x_\bullet - \mathbb{E}(X))^2$, y entonces:

$$a_* = \frac{\rho_X}{2A\delta_X}$$

En la expresión anterior ρ_X representa el rendimiento esperado de ABC por encima de lo que se obtiene a tasa fija r. Normalmente tendremos que $\rho_X > 0$, esto quiere decir que por lo general si vamos a considerar invertir en una opción con riesgo pedimos que al menos su rendimiento esperado sea mayor al de una opción sin riesgo. Por otro lado, δ_X representa una medida de riesgo y tiene dos componentes: la varianza del rendimiento de ABC así como una consideración respecto a asimetría en las probabilidades asignadas a los escenarios ya que $x_\bullet = \mathbb{E}(X)$ si, por ejemplo, X tiene distribución uniforme discreta. Y es aquí donde se ve que la elección óptima está considerando tanto rendimiento como riesgo ya que a mayor rendimiento de ABC se tendrá un mayor valor de a_* y a mayor riesgo (varianza y asimetría) de ABC se tendrá un menor valor de a_*. Con la solución óptima anterior se obtiene la siguiente utilidad óptima:

$$\overline{w}(a_*) = \frac{\rho_X^2}{4A\delta_X} + r$$

Aquí es interesante notar que aún cuando $\rho_X < 0$ se tendrá una utilidad (rendimiento esperado) por encima de la tasa fija r pero el que $\rho_X < 0$ implica que $a_* < 0$, y a primera vista parece un sin sentido un porcentaje negativo de inversión, pero quienes tengan un poco de conocimientos bursátiles saben que esto corresponde a lo que se conoce como *ventas en corto*, concepto que no discutiremos aquí pero baste mencionarlo para que nos quedemos tranquilos de que estamos obteniendo resultados coherentes. También puede ocurrir que $a_* > 100\,\%$ en cuyo caso tendremos un porcentaje negativo de inversión a tasa fija, lo cual también es posible pues esto quiere decir que, además del dinero propio, habría que pedir prestado más dinero (a dicha tasa fija del $6\,\%$) para invertirlo también en ABC. Con los datos específicos de este ejemplo obtenemos:

$$a_* = \frac{357.95\,\%}{A} \qquad\qquad \overline{w}(a_*) = \frac{12.53\,\%}{A} + 6\,\%$$

Para distintos valores de A tenemos las siguientes acciones óptimas:

A	a_*
$[\,0, 3.57\,]$	a_1 (pidiendo prestado)
3.58	a_1 (con inversión al $100\,\%$ en ABC)
$[\,3.58, 5.48\,]$	a_1
$[\,5.49, 5.52\,]$	$a_1 \sim a_2$
$[\,5.53, 10.20\,]$	a_2
$[\,10.21, 10.24\,]$	$a_2 \sim a_3$
$[\,10.25,\ 35.6\,]$	a_3
$[\,35.7, 35.9\,]$	$a_3 \sim a_4$
$[\,36, 62.2\,]$	a_4
$[\,62.3, 115.3\,]$	a_4 (con a_1 inadmisible)
$[\,115.4, 403.8\,]$	a_4 (con a_1, a_2 inadmisibles)
$[\,403.9, \infty\,[$	a_4 (con a_1, a_2, a_3 inadmisibles)

¿Qué valor de A se debe usar? Esto dependerá de qué tanto riesgo se esté dispuesto a tomar. Nótese que si no existe preocupación alguna por el riesgo ($A = 0$) entonces $w(a_i, E_j) = u(a_i, E_j)$. En cambio una gran preocupación por el riesgo se refleja en valores grandes para A. El cómo traducir el nivel de aversión al riesgo de un determinado inversionista en un valor para A es materia ya de un estudio más profundo que, como elegantemente dicen muchos libros, *is beyond the scope of this book*. Pero para no dejarlo así, una manera simplista de averiguar el coeficiente de aversión al riesgo A de un determinado inversionista sería, por ejemplo, preguntándole como qué rendimiento anda buscando. Supongamos que busca $1.5\,\%$ por arriba de la tasa fija del $6\,\%$, entonces:

$$\overline{u}(a_*) = a_*[\mathbb{E}(X) - r] + r = r + 1.5\,\%$$

despejando a_* e igualándola con la fórmula de la acción óptima obtenemos $A = 16.7$ de donde obtenemos $a_* = a_3$. Más aún, si en lugar de los cuatro portafolios que se mencionaron existe libertad para decidir los porcentajes como se quiera, la inversión óptima sería en este caso invertir $21.43\,\%$ en ABC y el resto a tasa fija. $\diamond$

Volviendo a la restricción de que $\mathcal{A}$ y Θ sean finitos, si alguno de ellos no lo fuera, el problema de encontrar la acción óptima puede aún así tener solución (como en el ejemplo anterior con $\mathcal{A} := [\,0, 100\,\%\,]$) o bien no tenerla, como se ilustra en el Ejercicio 3, al final de este capítulo.

4.3. Problemas de decisión secuencial

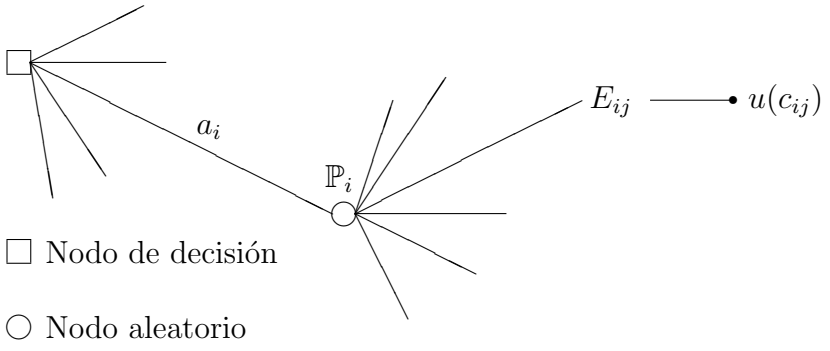
Hasta el momento hemos hablado de problemas de decisión en donde el espacio de estados o eventos relevantes Θ es el mismo bajo cualquier acción $a_i \in \mathcal{A}$, pero esto no tiene que ser necesariamente así, bien puede ocurrir que dependiendo de la acción que se tome se tenga un conjunto de eventos relevantes diferente, es decir, bajo una acción $a_i \in \mathcal{A}$ se tiene un conjunto particular de m_i eventos o estados relevantes $\Theta_i := \{E_{i1}, E_{i2}, \ldots, E_{im_i}\}$ dando lugar a los conjuntos de consecuencias $\mathcal{C}_i := \{c_{i1}, c_{i2}, \ldots, c_{im_i}\}$. Y al igual que en la sección anterior, si se tiene una función de utilidad definida para el conjunto de consecuencias $\bigcup \mathcal{C}_i$ y funciones de probabilidad $\mathbb{P}_i$ para los espacios de estados Θ_i entonces nuevamente la acción óptima será aquélla $a_* \in \mathcal{A}$ que satisfaga:

$$\overline{u}(a_*) = \max_i \overline{u}(a_i)$$

en donde:

$$\overline{u}(a_i) := \sum_{j=1}^{m_i} u(a_i, E_{ij})\, \mathbb{P}_i(E_{ij})$$

De manera esquemática:



Aunque ésta es una forma más general de un problema de decisión que el inicialmente presentado sigue siendo un *problema de decisión simple o de una sola etapa* en el sentido de que de que se toma una sola decisión, pero es posible tener un *problema de decisión secuencial* que es una concatenación de problemas de decisión simples, en donde algunas o todas las consecuencias consisten en tener que resolver nuevos problemas de decisión.

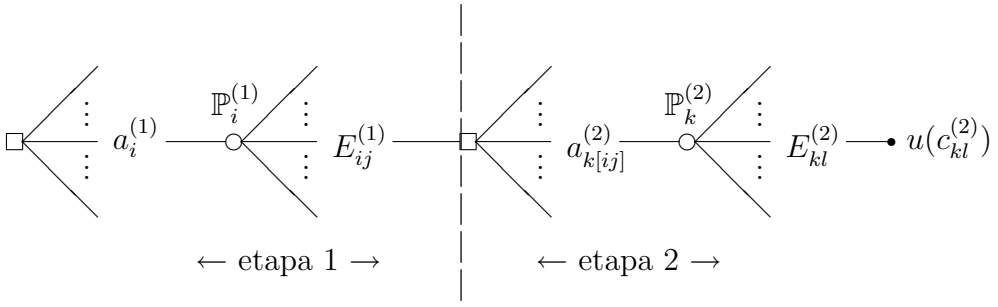
En un problema de decisión secuencial la decisión óptima en la primera etapa depende de las elecciones óptimas en las etapas subsecuentes. En el caso general de un problema de decisión con n etapas, la solución puede obtenerse de la siguiente manera:

1. Se resuelve primero la n-ésima etapa (la última) maximizando las utilidades esperadas apropiadas,

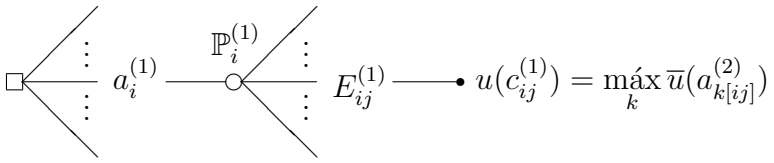
2. se resuelve la $(n-1)$-ésima etapa maximizando las correspondientes utilidades esperadas condicionalmente es las elecciones óptimas de la n-ésima etapa,

3. se continúa de esta manera siempre trabajando hacia atrás hasta que se obtenga la elección óptima en la primera etapa.

En el caso de un problema de decisión secuencial de $n = 2$ etapas:



Resolviendo la última etapa:



en donde:

$$\overline{u}(a_{k[ij]}^{(2)}) = \sum_l u(c_{kl}^{(2)})\, \mathbb{P}_k^{(2)}(E_{kl}^{(2)})$$

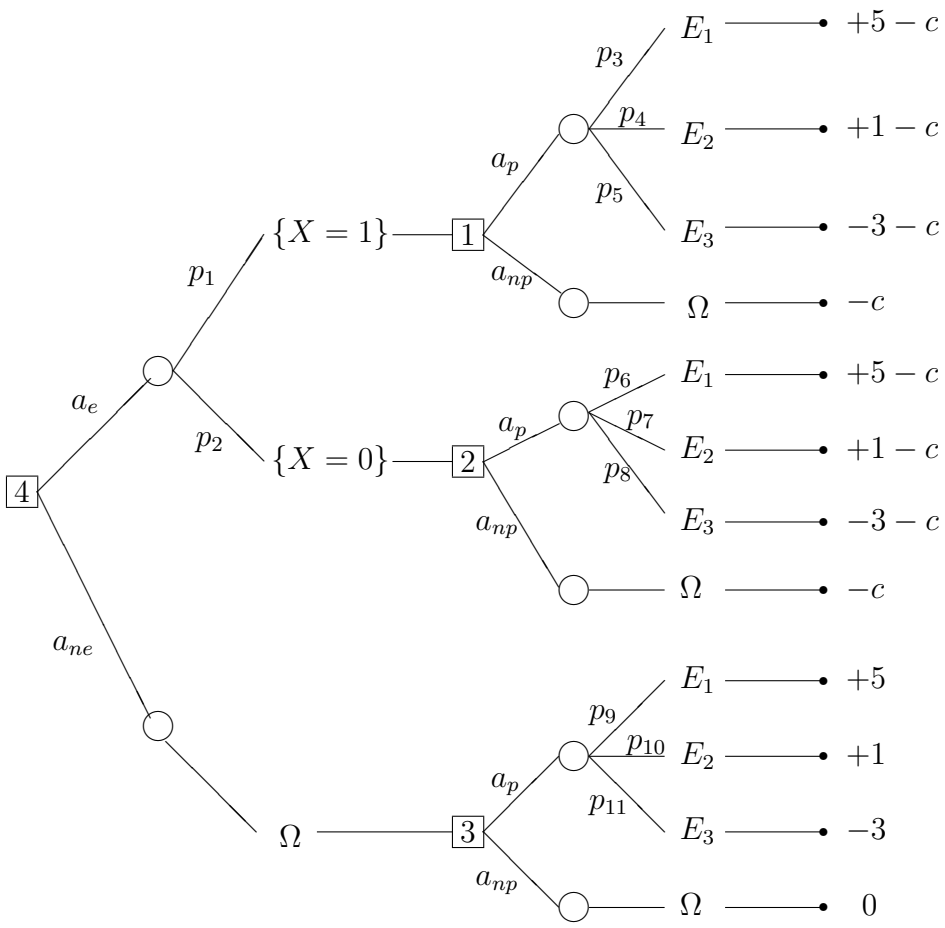
Ejemplo 10. Una empresa farmacéutica se plantea la posibilidad de lanzar al mercado un nuevo antigripal. Un despacho de actuarios le ofrece la realización de un estudio de mercado para reducir la incertidumbre sobre la proporción de médicos que lo recetarían. Sean los eventos:

$$\begin{aligned} E_1 &:= \text{una proporción alta de médicos lo recetarán} \\ E_2 &:= \text{una proporción moderada de médicos lo recetarán} \\ E_3 &:= \text{una proporción baja de médicos lo recetarán} \end{aligned}$$

A priori la compañía estima que $\mathbb{P}(E_1) = 0.2$, $\mathbb{P}(E_2) = 0.5$ y por tanto $\mathbb{P}(E_3) = 0.3$ y las ganancias que obtendría la empresa si lanza el producto bajo cada escenario serían +\$5, +\$1 y -\$3 millones de pesos, respectivamente. El estudio propuesto puede aconsejar la producción ($X = 1$) o desaconsejarla ($X = 0$) y las probabilidades de que el resultado del estudio sea aconsejar la producción dada la proporción de médicos que recetarían el antigripal son:

$$\mathbb{P}(X = 1 \,|\, E_1) = 0.9 \qquad \mathbb{P}(X = 1 \,|\, E_2) = 0.5 \qquad \mathbb{P}(X = 1 \,|\, E_3) = 0.2$$

¿Cuál es el precio máximo que la empresa farmacéutica debe pagar por el estudio? Sea c el costo de dicho estudio. Tenemos entonces:



El nodo de decisión 4 corresponde a la primera etapa y los nodos de decisión 1,2 y 3 corresponden a la segunda. En la primera etapa el conjunto de acciones es $\mathcal{A}^{(1)} := \{a_e, a_{ne}\}$ en donde:

$$a_e := \text{hacer el estudio}$$
$$a_{ne} := \text{no hacer el estudio}$$

En la segunda etapa los nodos de decisión tienen el mismo conjunto de

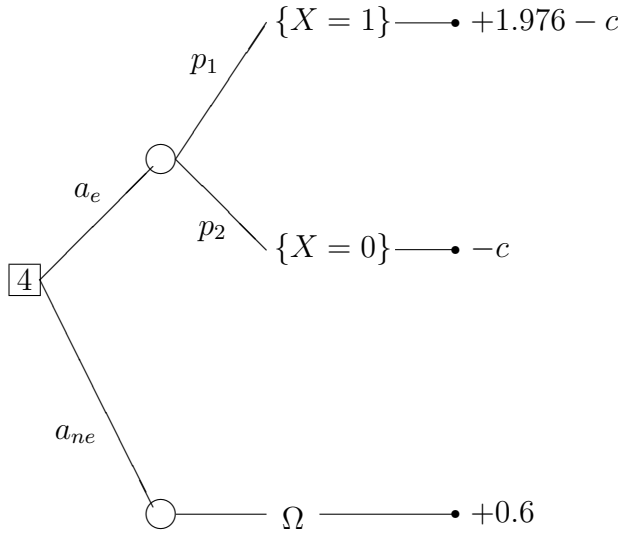
acciones $\mathcal{A}^{(2)} := \{a_p, a_{np}\}$ en donde:

$$\begin{aligned} a_p &:= \text{ producir el antigripal} \\ a_{np} &:= \text{ no producir el antigripal} \end{aligned}$$

La medida de probabilidad para el espacio de estados $\{E_1, E_2, E_3\}$ varía según el nodo aleatorio. Para el nodo aleatorio correspondiente al nodo de decisión 1 tenemos, utilizando la regla de Bayes:

$$\begin{aligned} p_3 &= \mathbb{P}(E_1 \,|\, X = 1) \\ &= \frac{\mathbb{P}(X = 1 \,|\, E_1)\mathbb{P}(E_1)}{\mathbb{P}(X = 1) \,|\, E_1)\mathbb{P}(E_1) + \mathbb{P}(X = 1) \,|\, E_1^c)\mathbb{P}(E_1^c)} \\ &= 0.367 \end{aligned}$$

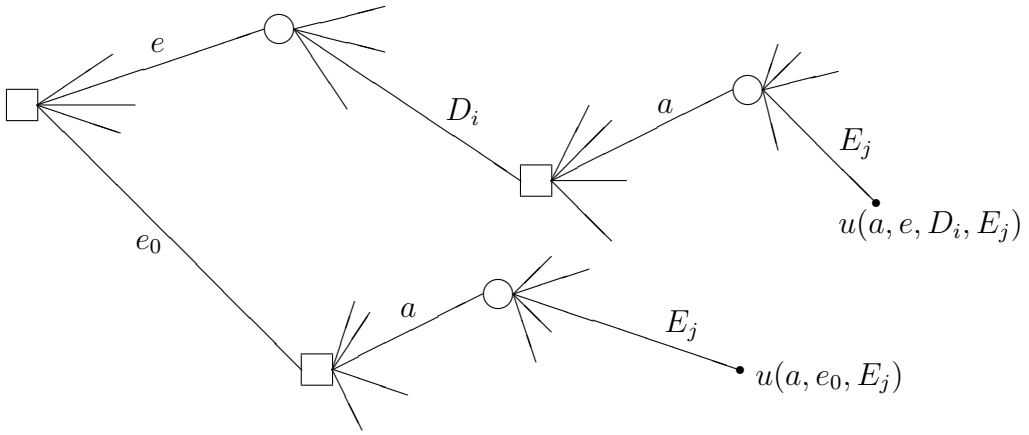
y de manera análoga $p_4 = 0.510, p_5 = 0.123, p_6 = 0.039, p_7 = 0.490, p_8 = 0.471, p_9 = 0.2, p_{10} = 0.5, p_{11} = 0.3, p_1 = 0.49, p_2 = 0.51$. Resolviendo los nodos de decisión de la segunda etapa obtenemos:



Resulta entonces preferible hacer el estudio a no hacerlo ($a_e \succ a_{ne}$) siempre y cuando se cumpla $\overline{u}(a_e) > \overline{u}(a_{ne})$ lo cual ocurre si y sólo si $c < 0.368$ así que para la empresa farmacéutica resulta conveniente pagar porque se haga el estudio siempre y cuando el costo de éste no exceda \$368,000. $\diamond$

Un problema de decisión secuencial que nos interesa de manera particular es aquél en que se tiene que decidir llevar a cabo un experimento de entre

varios posibles, y una vez escogido el experimento éste es utilizado en un problema de decisión subsecuente, y se desea escoger el experimento óptimo. Este problema particular se conoce como *diseño experimental*. Esquemáticamente:



Primero escogemos un experimento e y de acuerdo a los datos obtenidos D tomamos la acción a, después de la cual y ante la ocurrencia del evento E produce una consecuencia cuya utilidad denotaremos $u(a, e, D, E)$. Entre los posibles experimentos a escoger incluimos un *experimento nulo* e_0 que representa el caso en que decidamos irnos directo a las acciones posibles sin llevar a cabo experimento alguno.

Resolviendo primero los nodos de decisión de la segunda etapa tendremos que calcular la utilidad esperada de las acciones:

$$\overline{u}(a, e, D_i) = \sum_j u(a, e, D_i, E_j)\, \mathbb{P}(E_j \mid e, D_i, a)$$

Con lo anterior y para cada par (e, D_i) podemos escoger la acción óptima a seguir en cada caso, esto es, una acción a_i^* que maximice la expresión anterior. De este modo, la utilidad de la consecuencia (e, D_i) estará dada por:

$$u(e, D_i) = \overline{u}(a_i^*, e, D_i) = \max_a \overline{u}(a, e, D_i)$$

Ahora sólo queda resolver la primera etapa, es decir, determinar cuál es el experimento óptimo y para ello calculamos la utilidad esperada de cada experimento:

$$\overline{u}(e) = \sum_i \overline{u}(a_i^*, e, D_i)\, \mathbb{P}(D_i \mid e)$$

En el caso particular del experimento nulo tenemos:

$$\overline{u}(e_0) = \overline{u}(a_0^*, e_0) = \max_a \sum_j u(a, e_0, E_j)\mathbb{P}(E_j \mid e_0, a)$$

por lo que vale la pena llevar a cabo un experimento e siempre y cuando $\overline{u}(e) > \overline{u}(e_0)$:

4.11. Proposición. *La acción óptima es llevar a cabo el experimento e^* si $\overline{u}(e^*) > \overline{u}(e_0)$ y $\overline{u}(e^*) = \max_e \overline{u}(e)$; de lo contrario, la acción óptima es no realizar experimento alguno.*

Demostración. Es inmediata a partir de la Definición 4.8. □

Con lo anterior tenemos forma de definir un valor para la información adicional que se puede obtener en el contexto de un problema de decisión dado. Es posible calcular el valor esperado de la información que nos dan los datos como la esperanza (a posteriori) de la diferencia entre las utilidades que corresponden a las acciones óptimas después y antes de considerar los datos obtenidos:

4.12. Definición.

1. El *valor esperado de los datos D_i* proveniente de un experimento e está definido por:

$$v(e, D_i) := \sum_j \left[u(a_i^*, e, D_i, E_j) - u(a_0^*, e_0, E_j)\right] \mathbb{P}(E_j \mid e, D_i, a_i^*)$$

   donde a_i^* y a_0^* son las acciones óptimas dados los datos D_i y en ausencia de datos, respectivamente.

2. El *valor esperado de un experimento e* está dado por:

$$v(e) := \sum_i v(e, D_i)\, \mathbb{P}(D_i \mid e)$$

Y para tener una idea de qué tan grande es el valor $v(e)$ de un experimento e es posible calcularle una cota superior. Consideremos las acciones óptimas que estarían disponibles bajo *información perfecta*, esto es, suponiendo que

sabemos de antemano que el evento E_j va a ocurrir, y sea $a^*_{(j)}$ la acción óptima dado E_j, es decir tal que:

$$u(a^*_{(j)}, e_0, E_j) = \max_a u(a, e_0, E_j)$$

De este modo, dado E_j, la pérdida que se tiene por escoger cualquier otra acción a será:

$$u(a^*_{(j)}, e_0, E_j) - u(a, e_0, E_j)$$

Para $a = a^*_0$ (la acción óptima a priori) esta diferencia proporciona, condicional en E_j, el *valor de información perfecta* y, bajo ciertas condiciones, su valor esperado nos dará una cota superior para el incremento en utilidad que nos proporcionarían datos adicionales acerca de los eventos E_j :

4.13. Definición. La *pérdida de oportunidad* que se tiene si se toma la acción a y ocurre el evento E_j está dada por:

$$l(a, E_j) := \max_{a_i} u(a_i, e_0, E_j) - u(a, e_0, E_j)$$

y el *valor esperado de información perfecta* está dado por:

$$v^*(e_0) := \sum_j l(a^*_0, E_j)\, \mathbb{P}(E_j \,|\, a^*_0)$$

Es importante no perder de vista que las funciones $v(e, D_i)$ y $v(e)$ así como el número $v^*(e_0)$ dependen de las distribuciones (a priori) :

$$\{\mathbb{P}(E_j \,|\, a) \,:\, a \in \mathcal{A}\}$$

Existen situaciones en las que es posible separar la función de utilidad $u(a, e, D_i, E_j)$ en dos componentes: el *costo* de llevar a cabo el experimento e para obtener los datos D_i y la *utilidad* que se obtiene cuando se escoge la acción a y ocurre el evento E_j. Comúnmente el componente *utilidad* no depende de (e, D_i) por lo que, suponiendo aditividad de ambos componentes:

$$u(a, e, D_i, E_j) = u(a, e_0, E_j) - c(e, D_i)\ , \quad c(e, D_i) \geq 0$$

Más aún, las distribuciones de probabilidad sobre los eventos son, por lo general, independientes de las acciones. Bajo las condiciones anteriores es posible calcular una cota superior para el valor esperado de un experimento:

4.14. Proposición. *Si la función de utilidad es de la forma:*

$$u(a, e, D_i, E_j) = u(a, e_0, E_j) - c(e, D_i)\ , \quad c(e, D_i) \geq 0\,,$$

y las distribuciones de probabilidad son tales que

$$\mathbb{P}(E_j \,|\, e, D_i, a) = \mathbb{P}(E_j \,|\, e, D_i)\,, \quad \mathbb{P}(E_j \,|\, e_0, a) = \mathbb{P}(E_j \,|\, e_0)\,,$$

entonces, para cualquier experimento disponible e, se tiene que

$$v(e) \leq v^*(e_0) - \overline{c}(e)\,,$$

en donde

$$\overline{c}(e) := \sum_i c(e, D_i)\, \mathbb{P}(D_i \,|\, e)$$

es el costo esperado del experimento e.

Demostración. Utilizando las definiciones 4.12 y 4.13 podemos reexpresar $v(e)$ como:

$$\begin{aligned}
v(e) &= \sum_i \left[\sum_j \Big\{ u(a_i^*, e_0, E_j) - c(e, D_i) - u(a_0^*, e_0, E_j) \Big\} \mathbb{P}(E_j \,|\, e, D_i) \right] \mathbb{P}(D_i \,|\, e) \\
&= \sum_i \left[\max_a \sum_j \Big\{ u(a, e_0, E_j) - u(a_0^*, e_0, E_j) \Big\} \mathbb{P}(E_j \,|\, e, D_i) \right] \mathbb{P}(D_i \,|\, e) - \overline{c}(e) \\
&\leq \sum_i \sum_j \Big[\max_a u(a, e_0, E_j) - u(a_0^*, e_0, E_j) \Big] \mathbb{P}(E_j \cap D_i \,|\, e) - \overline{c}(e) \\
&\leq \Big\{ \sum_j l(a_0^*, E_j)\, \mathbb{P}(E_j \,|\, a_0^*) \Big\} \Big\{ \sum_i \mathbb{P}(D_i \,|\, E_j, e) \Big\} - \overline{c}(e) \\
&\leq \Big\{ \sum_j l(a_0^*, E_j)\, \mathbb{P}(E_j \,|\, a_0^*) \Big\} - \overline{c}(e) \\
&\leq v^*(e_0) - \overline{c}(e)
\end{aligned}$$

□

Ejemplo 11. Continuando con el Ejemplo 10 tendríamos como e el experimento de llevar a cabo el estudio y como e_0 el no llevarlo a cabo. Para evitar conflicto de notación lo que en el ejemplo anterior denotamos como c ahora lo denotaremos k. Entonces:

$$\overline{u}(e) = 0.968 - k\ , \quad \overline{u}(e_0) = 0.6$$

por lo que vale la pena llevar a cabo el experimento e siempre y cuando $\overline{u}(e) > \overline{u}(e_0)$, es decir, siempre que $k < 0.368$. Los datos D_i que se pueden obtener están en este caso representados por los eventos $D_1 := \{X = 1\}$ y $D_2 := \{X = 0\}$:

$$v(e, D_1) = -k \ , \quad v(e, D_2) = -k - 1.00776$$

y como $\overline{c}(e) = k$ entonces:

$$v(e) = -k - 1.00776 \leq 0.9 - k = v^*(e_0) - \overline{c}(e) \qquad \diamond$$

4.4. Inferencia e información

En las secciones previas hemos visto que para resolver un problema de decisión $(\mathcal{E}, \mathcal{C}, \mathcal{A}, \preceq)$ requerimos de una función de utilidad $u : \mathcal{C} \to \mathbb{R}$ y de una medida de probabilidad $\mathbb{P}$ sobre $\mathcal{E}$, y aplicamos el criterio de la utilidad esperada máxima sobre el conjunto de acciones $\mathcal{A}$. En esta sección nos concentraremos en algunos aspectos sobre la asignación de dicha medida de probabilidad.

La medida de probabilidad $\mathbb{P}$ se asigna de acuerdo al estado de información que en un momento dado tiene un individuo (o grupo de individuos). Dado un estado inicial de información M_0, las probabilidades que se asignen a los distintos eventos de $\mathcal{E}$ son probabilidades condicionales en dicha información inicial (o probabilidades a priori) y en estricto sentido debiéramos denotarlas $\mathbb{P}(\ \cdot \mid M_0)$ aunque por simplicidad se suele omitir el condicionamiento y lo denotamos simplemente $\mathbb{P}(\ \cdot\ )$. Después de ese momento inicial puede el individuo (o grupo de individuos) recibir información adicional (por ejemplo, resultados de un nuevo experimento, encuesta, etc.) misma que denotamos como un evento G. Esto nos lleva inmediatamente a *actualizar* la medida de probabilidad $\mathbb{P}(\ \cdot\ )$ a una medida de probabilidad $\mathbb{P}_1(\ \cdot\ ) := \mathbb{P}(\ \cdot \mid G)$. Normalmente dicha información adicional G corresponde a datos recolectados en relación con el fenómeno aleatorio de interés y en tal caso denotaremos dicha información como D y por tanto la medida de probabilidad inicial quedará actualizada a una medida de probabilidad $\mathbb{P}(\ \cdot \mid D)$, que equivale a lo que en el Capítulo 2 definimos como probabilidad a posteriori.

La utilidad esperada de cada acción (y por tanto la utilidad esperada máxima) depende tanto de la función de utilidad sobre $\mathcal{C}$ como de la medida

de probabilidad que se utilice y por ello analizaremos el conjunto de medidas de probabilidad que se pueden asignar y el problema de escoger una en particular como un problema de decisión.

4.15. Definición. Sea el espacio de estados (relevantes) $\Theta := \{E_j \,:\, j \in J\}$. Definimos como la *clase de distribuciones condicionales* sobre Θ:

$$\mathcal{Q} := \{\mathbf{q} \equiv (q_j, j \in J) \,:\, q_j \geq 0, \sum_{j \in J} q_j = 1\}\,.$$

Si supiéramos de antemano que el evento E_{j^*} va a ocurrir (información perfecta) la distribución de probabilidad "ideal" sobre el espacio de estados Θ sería $\mathbf{q}^* \equiv (q_j = \mathbf{1}_{\{j=j^*\}})$. No siempre es posible tener información perfecta ni tampoco garantía de que una determinada distribución de probabilidad $\mathbf{q}$ que elijamos sea la más adecuada. Supongamos por un momento que la distribución de probabilidad "correcta" sobre Θ la denotamos por:

$$\mathbf{p} \equiv \big(p_j = \mathbb{P}(E_j \,|\, D) \,:\, j \in J, p_j > 0, \sum_{j \in J} p_j = 1\big)$$

Nótese que en la definición de $\mathbf{p}$ tenemos la condición estricta $p_j > 0$. Esto es, pediremos que los elementos del espacio de eventos relevantes tengan medida de probabilidad distinta de cero.

Consideremos el problema de decisión $(\mathcal{E}, \mathcal{C}, \mathcal{A}, \preceq)$ en donde $\mathcal{A} := \mathcal{Q}$ y el espacio de estados relevantes es $\Theta := \{E_j \,:\, j \in J\}$ y por tanto el conjunto de consecuencias está dado por $\mathcal{C} = \mathcal{Q} \times \Theta$. Sólo nos falta definir una función de utilidad u sobre $\mathcal{C}$ que describa el "valor" $u(\mathbf{q}, E_j)$ de utilizar la distribución de probabilidad $\mathbf{q}$ bajo la ocurrencia del evento E_j. De cómo definir tal función de utilidad nos ocuparemos a continuación, y a este tipo particular de funciones se les conoce como *funciones de puntaje* ("score functions", en inglés).

4.16. Definición. Una *función de puntaje* para una familia de distribuciones de probabilidad $\mathcal{Q} := \{\mathbf{q} = (q_j \,:\, j \in J)\}$ definidas sobre una partición $\Theta := \{E_j, j \in J\}$ es una función $u : \mathcal{Q} \times \Theta \to \mathbb{R}$. Se dice que esta función es *suave* si es continuamente diferenciable como función de cada q_j.

Esta condición de suavidad resulta deseable en tanto que esperaríamos que cambios pequeños en algún q_j produzca cambios pequeños en el puntaje asignado por u.

Y nuevamente la elección óptima de entre los elementos de $\mathcal{Q}$ será aquélla $\mathbf{q}^*$ tal que:

$$\overline{u}(\mathbf{q}^*) = \max_{\mathbf{q}\in\mathcal{Q}} \overline{u}(\mathbf{q})$$

en donde

$$\overline{u}(\mathbf{q}) = \sum_{j\in J} u(\mathbf{q}, E_j)\, \mathbb{P}(E_j \mid D)$$

Una característica razonable que debe tener una función de puntaje u es que $\mathbf{q}^* = \mathbf{p}$ en donde, recordemos, $\mathbf{p} \equiv (p_j \,:\, j \in J)$ tal que $p_j := \mathbb{P}(E_j \mid D)$:

4.17. Definición. Una función de puntaje u es *propia* si para cada distribución de probabilidad $\mathbf{p} \equiv (p_j \,:\, j \in J)$ definida sobre una partición $\Theta := \{E_j \,:\, j \in J\}$ se cumple:

$$\sup_{\mathbf{q}\in\mathcal{Q}} \Big\{ \sum_{j\in J} u(\mathbf{q}, E_j) p_j \Big\} = \sum_{j\in J} u(\mathbf{p}, E_j) p_j$$

en donde el supremo se alcanza sólo si $\mathbf{q} = \mathbf{p}$.

Entre las aplicaciones que tienen las funciones de puntaje propias se encuentra el pago de premios justos a meteorólogos o analistas financieros por sus predicciones, en donde sus predicciones suelen ser asignaciones de probabilidades para distintos escenarios posibles, más que una predicción de qué escenario va a ocurrir exactamente. También está el caso de exámenes de opción múltiple. El método tradicional de evaluación de estos exámenes suele ser binario (acertó o no acertó); sin embargo, ante un acierto existe la posibilidad de que le haya atinado a pesar de desconcer la respuesta correcta, y también el que no haya acertado no implica que el conocimiento respecto a dicha pregunta sea completamente nulo. Utilizando funciones de puntaje se puede solicitar a quien responde el examen que, en vez de escoger una de las opciones, asigne una distribución de probabilidad para cada una de las opciones. Si está completamenrte seguro acerca de la respuesta puede asignar probabilidad 1 a la misma y cero al resto de las opciones; si está indeciso entre dos de las opciones puede asignar toda la masa de probabilidades en ellas para expresarlo. Esto nos da un panorama más amplio acerca de lo que sabe quien presenta un examen de opción múltiple. La función de puntaje se diseña de modo que quien no tenga idea de la respuesta le resulte más conveniente confesarlo vía la asignación de una distribución de probabilidad uniforme discreta que intentar simular que sabe. Un ejemplo de función de puntaje propia es el siguiente:

4.18. Definición. Una *función de puntaje cuadrática* para las distribuciones $\mathbf{q} \equiv (q_j : j \in J)$ definidas sobre una partición $\{E_j : j \in J\}$ es cualquier función de la forma

$$u(\mathbf{q}, E_j) = A\Big\{2q_j - \sum_{i\in J} q_i^2\Big\} + B_j \ , \quad A > 0$$

o alternativamente

$$u(\mathbf{q}, E_j) = A\Big\{1 - \sum_{i\in J}(q_i - \mathbf{1}_{\{i=j\}})^2 + B_j\Big\} \ , \quad A > 0$$

en donde $(q_i - \mathbf{1}_{\{i=j\}})^2$ representa una penalización. Es inmediato verificar que ambas expresiones son equivalentes.

4.19. Proposición. *Una función de puntaje cuadrática es propia.*

Demostración. Sea la función:

$$f(\mathbf{q}) \equiv f(q_1, \ldots, q_m) := \sum_{j\in J} u(\mathbf{q}, E_j)p_j$$

Como u es función de puntaje cuadrática entonces

$$f(\mathbf{q}) = \sum_{j\in J}\Big\{A\big(2q_j - \sum_{i\in J} q_i^2\big) + B_j\Big\}p_j \ , \quad A > 0$$

Calculando las parciales de f:

$$\begin{aligned}
\frac{\partial}{\partial q_k} f(\mathbf{q}) &= \frac{\partial}{\partial q_k}\Big[\Big\{A\big(2q_k - \sum_{i\in J} q_i^2\big) + B_k\Big\}p_k + \sum_{j\in J, j\neq k}\Big\{A\big(2q_j - \sum_{i\in J} q_i^2\big) + B_j\Big\}p_j\Big] \\
&= \big\{A(2 - 2q_k)\big\}p_k + \sum_{j\in J, j\neq k}\big\{A(-2q_k)\big\}p_j \\
&= 2A(1 - q_k)p_k - 2Aq_k \sum_{j\in J, j\neq k} p_j \\
&= 2Ap_k - 2Aq_kp_k - 2Aq_k \sum_{j\in J, j\neq k} p_j \\
&= 2Ap_k - 2Aq_k \sum_{j\in J} p_j \\
&= 2A(p_k - q_k)
\end{aligned}$$

Por lo que

$$\frac{\partial}{\partial q_k} f(\mathbf{q}) = 0 \Leftrightarrow p_k = q_k \, , \, k = 1, \ldots, m$$

Además

$$\frac{\partial^2}{\partial q_k^2} f(\mathbf{q}) = -2A < 0$$

lo que implica que $f(\mathbf{q})$ alcanza un máximo si y sólo si $\mathbf{q} = \mathbf{p}$ y por lo tanto la función de puntaje cuadrática es propia. □

Ejemplo 12. Mencionamos ya que una de las aplicaciones de las funciones de puntaje es en el caso de exámenes de opción múltiple. En la forma tradicional de evaluar las respuestas de este tipo de exámenes (correcta o incorrecta) no se puede evitar dos problemas: primero, la deshonestidad de un estudiante que sin saber la respuesta escoge una al azar y le atina; segundo, una respuesta incorrecta no implica ausencia total de conocimiento (por ejemplo, de entre 5 respuestas posibles quizás le quedaba claro que tres de ellas no eran la respuesta correcta). Las funciones de puntaje propias nos permiten forzar al estudiante a ser honesto así como reconocer grados parciales de conocimiento.

Una forma de resolver este problema es pedirle al estudiante que proporcione una distribución de probabilidad sobre las posibles respuestas que describa lo que piensa acerca de la respuesta correcta. Desde el punto de vista del estudiante, contestar la pregunta es un problema de decisión donde el conjunto de acciones es ahora la clase:

$$\mathcal{Q} := \{\mathbf{q} \equiv (q_1, \ldots, q_m) \, : \, q_j \geq 0, \sum q_j = 1\}$$

de distribuciones de probabilidad sobre el conjunto $\{E_1, \ldots, E_m\}$ de respuestas posibles. En este caso la utilidad esperada puede ser definida de acuerdo a la calificación que se espera obtener:

$$\overline{u}(\mathbf{q}) = \sum_{j=1}^{m} u(\mathbf{q}, E_j) \, p_j$$

en donde $u(\mathbf{q}, E_j)$ es la calificación otorgada a un estudiante que reporta la distribución $\mathbf{q}$ cuando la respuesta correcta es E_j y p_j es la probabilidad personal que el estudiante asigna al evento de que la respuesta correcta sea E_j. Notemos que, en principio, no hay razón para suponer que la distribución $\mathbf{q}$

reportada por el estudiante como su respuesta es la misma que la distribución $\mathbf{p}$ que describe en realidad su conocimiento (es decir, el estudiante puede ser también deshonesto al contestar bajo este esquema).

Por su parte, el maestro está interesado en garantizar la honestidad del estudiante y para ello escoge una función de utilidad propia de modo que la utilidad esperada del estudiante se maximice si y sólo si $\mathbf{q} = \mathbf{p}$, y con ello el estudiante se autoperjudica si es deshonesto. Como ya vimos, la función de utilidad cuadrática es propia, y para determinar las constantes A y B_j establecemos condiciones adicionales. Por ejemplo, supongamos que se decide otorgar un punto si la distribución $\mathbf{q}$ asigna probabilidad 1 a la respuesta correcta, y cero puntos a la distribución uniforme $q_j = \frac{1}{m}$ (la cual describe ausencia de conocimiento). Esto nos lleva al sistema de ecuaciones:

$$\begin{aligned} A + B_j &= 1 \\ \frac{A}{m} + B_j &= 0 \end{aligned}$$

de donde $A = \frac{m}{m-1}$ y $B_j = \frac{1}{m-1}$ y por lo tanto:

$$u(\mathbf{q}, E_j) = \frac{m}{m-1}\big(2q_j - \sum_{i=1}^{m} q_i^2\big) - \frac{1}{m-1}$$

Notemos además que esta función asigna valores ngativos a distribuciones que asignen probabilidades altas a respuestas incorrectas. En particular, un estudiante que asigne probabilidad 1 a una respuesta incorrecta tendrá un descuento de $\frac{m+1}{m-1}$ puntos en su calificación, lo que implica que le resulte mejor, en promedio, en caso de desconocer por completo la respuesta correcta, admitir honestamente su ignorancia que intentar atinarle. Y por otro lado, si tenemos un estudiante que a pesar de no estar seguro de la respuesta correcta le queda claro que $m-2$ de las opciones deben descartarse y asigna probabilidad $\frac{1}{2}$ a dos opciones, una de las cuales es la correcta, entonces obtiene al menos $\frac{m-2}{2(m-1)}$ puntos, cantidad menor a uno pero positiva en caso de que $m > 2$. $\diamond$

Para otro tipo de aplicaciones, un conjunto particularmente importante de funciones de puntaje son aquéllas que dependen únicamente del evento que finalmente ocurre, esto es, de la probabilidad que la distribución $\mathbf{q}$ haya asignado a dicho evento, y de ahí el adjetivo de *local*:

4.20. Definición. Una función de puntaje u es *local* si para cada $\mathbf{q} \in \mathcal{Q}$ definida sobre la partición $\{E_j : j \in J\}$ existen funciones $\{u_j(\,\cdot\,) : j \in J\}$ tales que $u(\mathbf{q}, E_j) = u_j(q_j)$.

Este caso particular de funciones de puntaje es importante ya que una vez que se observe el evento que ocurre es respecto a éste que se comparan las inferencias o predicciones acerca de lo que iba a ocurrir, y por ello resultan adecuadas para la inferencia estadística. Lo que sigue es caracterizar este tipo de funciones:

4.21. Proposición. *Si u es una función de puntaje suave, propia y local para una clase de distribuciones $\mathbf{q} \in \mathcal{Q}$ definidas sobre una partición $\{E_j : j \in J\}$ que contiene más de dos elementos, entonces tiene que ser de la forma $u(\mathbf{q}, E_j) = A \log q_j + B_j$ en donde $A > 0$ y donde $\{B_j : j \in J\}$ son constantes arbitrarias.*

Demostración. Como $u(\,\cdot\,)$ es local y propia, entonces para algunas $\{u_j(\,\cdot\,) : j \in J\}$ se tiene que

$$\sup_{\mathbf{q}} \sum_{j \in J} u(\mathbf{q}, E_j) p_j = \sup_{\mathbf{q}} u_j(q_j) p_j = \sum_{j \in J} u_j(p_j) p_j \,,$$

en donde $p_j > 0$, $\sum_j p_j = 1$, y el supremo se toma sobre la clase de distribuciones $\mathbf{q} \equiv (q_j : j \in J)$ tales que $q_j \geq 0$ y $\sum_j q_j = 1$.

Denotando $\mathbf{p} \equiv (p_1, p_2, \ldots)$ y $\mathbf{q} \equiv (q_1, q_2, \ldots)$ en donde

$$p_1 = 1 - \sum_{j>1} p_j \,, \quad q_1 = 1 - \sum_{j>1} q_j \,,$$

caracterizamos las funciones $\{u_j(\,\cdot\,) : j \in J\}$ buscando un punto extremo de:

$$F(q_2, q_3, \ldots) := \Big(1 - \sum_{j>1} p_j\Big) u_1\Big(1 - \sum_{j>1} q_j\Big) + \sum_{j>1} p_j u_j(q_j)$$

Para que F sea estacionaria en algún punto $(q_2, q_3, \ldots)$ es necesario (ver Jeffreys y Jeffreys (1946), p.315) que

$$\frac{\partial}{\partial \alpha} F(q_2 + \alpha\varepsilon_2, q_3 + \alpha\varepsilon_3, \ldots) \restriction_{\alpha=0} = 0$$

para cualquier $\varepsilon := (\varepsilon_2, \varepsilon_3, \ldots)$ tal que las ε_j son suficientemente pequeñas. Calculando dicha derivada obtenemos:

$$\sum_{j>1} \left\{ \Big(1 - \sum_{i>1} p_i\Big) u_1'\Big(1 - \sum_{j>1} q_j\Big) \right\} \varepsilon_j = 0$$

para ε_j suficientemente pequeñas y en donde u' es la derivada de u. Como u es propia entonces $(p_2, p_3, \ldots)$ debe ser un punto extremo de F y obtenemos así el sistema de ecuaciones:

$$p_1 u_1'(p_1) = p_j u_j'(p_j)\,, \quad j = 1, 2, \ldots$$

para todos los valores $p_2, p_3, \ldots$ lo que implica que, para una constante A:

$$p\, u_j'(p) = A\,, \quad 0 < p \leq 1\,, \quad j = 1, 2, \ldots$$

de donde $u_j(p) = A \log p + B_j$. La condición de que $A > 0$ es suficiente para garantizar que tal punto extremo es, en efecto, un máximo. □

4.22. Definición. Una *función de puntaje logarítmica* para distribuciones de probabilidad estrictamente positivas $\mathbf{q} \equiv (q_j : j \in J)$ definidas sobre una partición $\{E_j : j \in J\}$ es cualquier función de la forma:

$$u(\mathbf{q}, E_j) = A \log q_j + B_j\,, \quad A > 0\,.$$

A continuación veremos la aplicación de este tipo de funciones de puntaje para aproximar, por ejemplo, una distribución de probabilidad $\mathbf{p}$ por medio de otra $\mathbf{q}$, o bien medir de algún modo qué tanto se aproxima una a la otra:

4.23. Proposición. *Si nuestras preferencias están descritas por una función de puntaje logarítimica, la pérdida esperada de utilidad al usar una distribución de probabilidad* $\mathbf{q} \equiv (q_j : j \in J)$ *definida sobre una partición* $\{E_j : j \in J\}$ *en lugar de la distribución* $\mathbf{p} \equiv (p_j : j \in J)$ *que representa lo que realmente creemos, está dada por:*

$$\delta(\mathbf{q} \,|\, \mathbf{p}) = A \sum_{j \in J} p_j \log \frac{p_j}{q_j}\,, \quad A > 0\,.$$

Más aún, $\delta(\mathbf{q} \,|\, \mathbf{p}) \geq 0$ *con igualdad si y sólo si* $\mathbf{q} = \mathbf{p}$.

Demostración. Utilizando la Definición 4.22 tenemos que la utilidad esperada de usar la distribución $\mathbf{q}$ cuando $\mathbf{p}$ es la correcta es:

$$\overline{u} = \sum_{j\in J} \big(A \log p_j + b_j\big) p_j$$

por lo que

$$\begin{aligned} \delta(\mathbf{q}\,|\,\mathbf{p}) &= \overline{u}(\mathbf{p}) - \overline{u}(\mathbf{q}) \\ &= \sum_{j\in J} \Big\{ \big(A \log p_j + B_j\big) - \big(A \log q_j + B_j\big) \Big\} p_j \\ &= A \sum_{j\in J} p_j \log \frac{p_j}{q_j} \end{aligned}$$

Como la función de puntaje es logarítmica y por tanto propia entonces por la Proposición 4.21 tenemos que $\overline{u}(\mathbf{p}) \geq \overline{u}(\mathbf{q})$ con igualdad si y sólo si $\mathbf{p} = \mathbf{q}$. O bien, utilizando el hecho de que $1 + x \leq e^x$ de donde $x \leq e^{x-1}$ y si $x > 0$ entonces $\log x \leq x - 1$ con igualdad si $x = 1$:

$$-\delta(\mathbf{q}\,|\,\mathbf{p}) = \sum_{j\in J} p_j \log \frac{q_j}{p_j} \leq \sum_{j\in J} p_j \Big(\frac{q_j}{p_j} - 1\Big) = \sum_j q_j - \sum_j p_j = 1 - 1 = 0$$

con igualdad si y sólo si $q_j = p_j$ para todo j. □

Una aplicación inmediata de lo anterior es en el caso de que se desee aproximar una distribución por medio de otra:

4.24. Definición. La *discrepancia logarítmica* entre una distribución de probabilidad estrictamente positiva $\mathbf{p} \equiv (p_j : j \in J)$ sobre una partición $\{E_j : j \in J\}$ y una aproximación $\hat{\mathbf{p}} \equiv (\hat{p}_j : j \in J)$ está definida por:

$$\delta(\hat{\mathbf{p}}\,|\,\mathbf{p}) := \sum_{j\in J} p_j \log \frac{p_j}{\hat{p}_j}\,.$$

Es inmediato notar que la discrepancia logarítmica no es una métrica, comenzando por que no es simétrica. Se puede hacer simétrica mediante:

$$\kappa\{\mathbf{q}, \mathbf{p}\} := \delta(\mathbf{q}\,|\,\mathbf{p}) + \delta(\mathbf{p}\,|\,\mathbf{q})$$

pero aún así no es métrica ya que no cumple la desigualdad del triángulo.

Ejemplo 13. Conocido resultado de probabilidad es que, bajo ciertas condiciones, se puede obtener una buena aproximación del modelo binomial por medio del modelo Poisson:

$$p_j = \binom{n}{j}\theta^j(1-\theta)^{n-j}\mathbf{1}_{\{0,1,\ldots,n\}}(j)$$
$$\hat{p}_j = \exp(-n\theta)\frac{(n\theta)^j}{j!}\mathbf{1}_{\{0,1,\ldots\}}(j)$$

de donde

$$\delta(\hat{\mathbf{p}}\,|\,\mathbf{p}) = \sum_{k=2}^{n}\log k + n\big[(1-\theta)\log(1-\theta) + \theta(1-\log n)\big] - \varphi(n,\theta)$$

en donde

$$\begin{aligned}\varphi(n,\theta) &:= \mathbb{E}_{\mathbf{p}}\Big\{\log\big[(n-X)\,!\big]\Big\}, \quad X\sim\mathbf{p}\\ &= \theta^{\,n}\sum_{k=2}^{n}\log(k\,!)\binom{n}{k}\Big(\frac{1}{\theta}-1\Big)^k\end{aligned}$$

Tenemos que $\delta\to 0$ conforme $n\to\infty$ y/o $\theta\to 0$, es decir, la aproximación es buena para valores grandes de n y/o valores de θ cercanos a cero. $\diamond$

Y en general, sea $\mathbf{p}$ una distribución de probabilidad y sea $\mathcal{Q}$ una familia de distribuciones de probabilidad para aproximar $\mathbf{p}$. Bajo discrepancia logarítmica, la *mejor aproximación* de $\mathbf{p}$ será aquella $\mathbf{q}^*\in\mathcal{Q}$ tal que:

$$\delta(\mathbf{q}^*\,|\,\mathbf{p}) = \min_{\{\,\mathbf{q}\in\mathcal{Q}\}}\delta(\mathbf{q}\,|\,\mathbf{p})\,.$$

En el Capítulo 2 y en el presente hemos visto que la probabilidad a priori $\mathbb{P}(\,\cdot\,)$ sobre un espacio de estados relevantes se *actualiza* con la recolección de datos D vía la regla de Bayes a una medida de probabilidad $\mathbb{P}(\,\cdot\,|\,D)$. También hemos visto que en un contexto de inferencia estadística, por ejemplo, la función de puntaje logarítmica resulta adecuada. Esto permitirá calcular la utilidad esperada de recolectar los datos D:

4.25. Proposición. *Si nuestras preferencias están descritas en términos de una función de puntaje logarítmica sobre la clase de distribuciones de probabilidad definidas en una partición* $\{E_j : j \in J\}$*, entonces el incremento esperado en utilidad proveniente de los datos* D*, cuando la distribución de probabilidad a priori* $\{\mathbb{P}(E_j) : j \in J\}$ *es estrictamente positiva, está dado por*

$$A \sum_{j \in J} \mathbb{P}(E_j \,|\, D) \log \frac{\mathbb{P}(E_j \,|\, D)}{\mathbb{P}(E_j)}$$

en donde $A > 0$ *es arbitraria y* $\{\mathbb{P}(E_j \,|\, D) : j \in J\}$ *es la probabilidad a posteriori dados los datos* D*. Más aún, este incremento esperado en utilidad es no negativo, y es cero si y sólo si* $\mathbb{P}(E_j \,|\, D) = \mathbb{P}(E_j)$ *para todo* j *.*

Demostración. Por la Definición 4.22 tenemos que la utilidad de reportar las distribuciones de probabilidad $\mathbb{P}(\,\cdot\,)$ y $\mathbb{P}(\,\cdot \mid D)$, bajo el supuesto de que ocurra el evento E_j, están dadas por $A \log \mathbb{P}(E_j) + B_j$ y $A \log \mathbb{P}(E_j \,|\, D) + B_j$, respectivamente. De este modo, el incremento esperado en utilidad proveniente de los datos D está dado por

$$\begin{aligned} \sum_{j \in J} \Big\{ \big(A \log \mathbb{P}(E_j \,|\, D) + B_j \big) - \big(A \log \mathbb{P}(E_j) + B_j \big) \Big\} \mathbb{P}(E_j \,|\, D) \\ = \; A \sum_{j \in J} \mathbb{P}(E_j \,|\, D) \log \frac{\mathbb{P}(E_j \,|\, D)}{\mathbb{P}(E_j)} \,, \end{aligned}$$

cantidad que, por la Proposición 4.23, es no negativa, y cero si y sólo si $\mathbb{P}(E_j \,|\, D) = \mathbb{P}(E_j)$ para todo j . □

Lo anterior motiva la siguiente definición:

4.26. Definición. La *cantidad de información de los datos* acerca de una partición $\{E_j : j \in J\}$ proveniente de los datos D cuando la distribución a priori sobre dicha partición es $\mathbf{p}_0 := \{\mathbb{P}(E_j) : j \in J\}$ se define como:

$$I(D \,|\, \mathbf{p}_0) := \sum_{j \in J} \mathbb{P}(E_j \,|\, D) \log \frac{\mathbb{P}(E_j \,|\, D)}{\mathbb{P}(E_j)}$$

en donde $\{\mathbb{P}(E_j \,|\, D) : j \in J\}$ es la distribución de probabilidad condicional dados los datos D .

Equivalentemente y de acuerdo a la Definición 4.24 la cantidad de información de los datos D es $\delta(\mathbf{p}_0 \,|\, \mathbf{p}_D)$, esto es la discrepancia logarítmica considerando a $\mathbf{p}_0$ como una aproximación de la distribución de probabilidad $\mathbf{p}_D := \big(\mathbb{P}(E_j \,|\, D) \,:\, j \in J\big)$.

Lo anterior nos permite calcular la cantidad esperada de información de un experimento e antes de que los resultados del mismo sean conocidos:

4.27. Definición. La *información esperada de un experimento e* sobre una partición $\{E_j \,:\, j \in J\}$ con distribución a priori $\mathbf{p}_0 := \{\mathbb{P}(E_j) \,:\, j \in J\}$ está dada por:

$$I(e \,|\, \mathbf{p}_0) := \sum_i I(D_i \,|\, \mathbf{p}_0)\, \mathbb{P}(D_i)$$

en donde los posibles resultados del experimento e, denotados por $\{D_i\}$, ocurren con probabilidades $\{\mathbb{P}(D_i)\}$.

Ejemplo 14. Retomando el Ejemplo 1, tenemos como espacio de estados relevantes $\{E_0, E_1\}$ en donde E_0 representa el evento de que salga sol y E_1 el evento de que salga águila. Suponiendo que la moneda fue escogida al azar (i.e. $\alpha = 1$) y de acuerdo a lo obtenido en el Ejemplo 1 tenemos que

$$\begin{aligned}
\mathbb{P}(E_j) &= \frac{1}{2} + \frac{\alpha}{4}(-1)^{j+1}\mathbf{1}_{\{0,1\}}(j) \\
\mathbb{P}(E_j \,|\, D) &= \frac{3^j\nu + 2}{4(\nu+1)}\mathbf{1}_{\{0,1\}}(j) = \frac{1}{4}\Big(3^j + \frac{1}{\nu+1}(-1)^j\Big)\mathbf{1}_{\{0,1\}}(j)
\end{aligned}$$

en donde

$$\nu = \frac{3^{\sum_{k=1}^n x_k}}{2^n}$$

y con lo anterior obtenemos

$$\begin{aligned}
I(D \,|\, \mathbf{p}_0) = \frac{1}{4}\Big\{ &\Big(1 + \frac{1}{\nu+1}\Big)\Big[\log 2 - \log 3 + \log\big(1 + \frac{1}{\nu+1}\big)\Big] \\
&+ \Big(3 - \frac{1}{\nu+1}\Big)\Big[\log 2 - \log 5 + \log\big(3 - \frac{1}{\nu+1}\big)\Big]\Big\}
\end{aligned}$$

En el Ejemplo 1 se vio que si el tamaño de muestra $n \to \infty$ entonces ocurre una de dos cosas: $\nu \to \infty$ (lo cual implicaría que $\theta = \frac{3}{4}$) o bien $\nu \to 0$ (lo cual implicaría que $\theta = \frac{1}{2}$). En el primer caso obtenemos $\lim_{n\to\infty} I(D \,|\, \mathbf{p}_0)$

es aproximadamente igual a 0.03537489 y en el segundo caso aproximadamente igual a 0.03226926. Resulta muy ilustrativo analizar gráficamente el comportamiento de $I(D \mid \mathbf{p}_0)$ para distintos valores de n y compararlo con el comportamiento de $p(x = 1 \mid \mathbf{x})$, bajo ambos escenarios de comportamiento asintótico de ν. (Sugerencia: simule muestras de tamaño $n \geq 100$). Ante la pregunta de por qué el supremo de $I(D \mid \mathbf{p}_0)$ es mayor cuando $\theta = \frac{3}{4}$ que con $\theta = \frac{1}{2}$, en este caso podemos contestar momentánea y parcialmente que la información de Fisher alcanza su mínimo justamente en $\theta = \frac{1}{2}$ (ver Ejemplo 4, Capítulo 3). Por último, para calcular la información esperada de este experimento $I(e \mid \mathbf{p}_0)$ sólo falta calcular $\mathbb{P}(D_i)$, que viene a ser en este caso la distribución predictiva a priori conjunta de n observaciones:

$$\begin{aligned} p(x_1, \ldots, x_n) &= \int_\Theta p(x_1, \ldots, x_n \mid \theta)\, p(\theta)\, d\theta \\ &= \frac{\nu + 1}{2^{n+1}} \end{aligned}$$

No procederemos en este caso a calcular explícitamente $I(e \mid \mathbf{p}_0)$ porque sólo estamos considerando un sólo tipo de experimento, y tendría sentido en tanto se tuvieran distintos experimentos a realizar para la obtención de datos, se calculan sus informaciones esperadas respectivas y se selecciona aquél que tenga la información esperada mayor. $\diamond$

4.5. Acciones y utilidades generalizadas

Para la aplicación de los resultados de teoría de la decisión a la inferencia estadística resulta necesario considerar conjuntos de acciones y espacios de estados infinito no numerables (como puede ser un subconjunto de $\mathbb{R}$, por ejemplo) así como permitir que $\mathcal{E}$ sea un σ-álgebra. Para ello son necesarias una serie de justificaciones formales que no analizaremos a detalle (para ello ver Bernardo y Smith (1994)).

Consideremos un problema de decisión $(\mathcal{E}, \mathcal{C}, \mathcal{A}, \preceq)$ en donde $\mathcal{A}$ y el espacio de estados Θ son infinito no numerables, $\mathcal{E}$ un σ-álgebra. Tenemos entonces que el conjunto de consecuencias $\mathcal{C}$ es también infinito no numerable. Sea la función de utilidad $u : \mathcal{C} \to \mathbb{R}$. Sea $p(\theta)$ una función de densidad de probabilidades sobre Θ. Tenemos entonces que la acción óptima será aquélla

$a^* \in \mathcal{A}$ tal que:

$$\overline{u}(a^*) = \max_{a \in \mathcal{A}} \overline{u}(a)\,,$$

en donde

$$\overline{u}(a) := \int_\Theta u(a,\theta)\, p(\theta)\, d\theta\,.$$

Cabe aclarar que $p(\theta)$ es una distribución de probabilidad condicional en la información que se tiene en un momento dado, esto es, puede tratarse de una distribución a priori o a posteriori, por ejemplo. Retomaremos los conceptos de la sección anterior pero para el caso que ahora nos ocupa.

4.28. Definición. Sea Θ un espacio de estados. Definimos como la *clase de distribuciones condicionales* sobre Θ:

$$\mathcal{Q} := \{\, q(\theta) \,:\, q(\theta) \geq 0, \int_\Theta q(\theta)\, d\theta = 1\}$$

4.29. Definición. Una *función de puntaje* para una familia de distribuciones de probabilidad $\mathcal{Q}$ definidas sobre un espacio de estados Θ es una función $u : \mathcal{Q} \times \Theta \to \mathbb{R}$. Se dice que esta función es *suave* si es continuamente diferenciable como función de θ.

4.30. Definición. Una función de puntaje u es *propia* si para cada distribución de probabilidad $p(\theta)$ definida sobre Θ se cumple:

$$\sup_{q \in \mathcal{Q}} \int_\Theta u(q,\theta)\, p(\theta)\, d\theta = \int_\Theta u(p,\theta)\, p(\theta)\, d\theta\,,$$

en donde el supremo se alcanza si y sólo si $q = p$ casi seguramente, esto es, excepto posiblemente sobre conjuntos de medida cero.

4.31. Definición. Una *función de puntaje cuadrática* para la familia de distribuciones de probabilidad $\mathcal{Q}$ definidas sobre Θ es cualquier función de la forma

$$u(q,\theta) = A\Big\{2\, q(\theta) - \int_\Theta q^2(\tilde{\theta})\, d\tilde{\theta}\Big\} + B(\theta)\,, \quad A > 0\,,$$

en donde la función $B(\,\cdot\,)$ es cualquiera mientras se garantice la existencia de

$$\overline{u}(q) = \int_\Theta u(q,\theta)\, p(\theta)\, d\theta\,.$$

4.32. Proposición. *Una función de puntaje cuadrática es propia.*

Demostración. Escogemos $q \in \mathcal{Q}$ de modo que se maximice

$$\begin{aligned}\overline{u}(q) &= \int_\Theta u(q,\theta)\, p(\theta)\, d\theta \\ &= \int_\Theta \Big[A\big\{2q(\theta) - \int_\Theta q^2(\tilde\theta)\, d\theta\big\} + B(\theta) \Big]\, p(\theta)\, d\theta \\ &= \int_\Theta \Big[2Ap(\theta)q(\theta) - Ap(\theta) \int_\Theta q^2(\tilde\theta)\, d\tilde\theta + B(\theta)p(\theta) \Big]\, d\theta\,,\end{aligned}$$

pero maximizar la expresión anterior (respecto a $q \in \mathcal{Q}$) equivale a maximizar

$$-\int_\Theta \big[\, p(\theta) - q(\theta) \,\big]^2\, d\theta$$

ya que, como p y B son fijas, entonces

$$\begin{aligned}-\int_\Theta \big[\, p(\theta) - q(\theta) \,\big]^2\, d\theta &= -\int_\Theta \big[\, p^2(\theta) - 2p(\theta)q(\theta) + q^2(\theta) \,\big]\, d\theta \\ &= \frac{1}{A}\int_\Theta \big[\, 2Ap(\theta)q(\theta) - Ap^2(\theta) \,\big]\, d\theta - \int_\Theta p(\theta) \int_\Theta q^2(\tilde\theta)\, d\tilde\theta\, d\theta \\ &= \frac{1}{A}\int_\Theta \big[\, 2Ap(\theta)q(\theta) - Ap(\theta) \int_\Theta q^2(\tilde\theta)\, d\tilde\theta - Ap^2(\theta) \,\big]\, d\theta\end{aligned}$$

y por lo tanto $\overline{u}(q)$ se maximiza sobre $\mathcal{Q}$ siempre y cuando $q = p$ casi seguramente. □

4.33. Definición. Una función de puntaje u es *local* si para cada $q \in \mathcal{Q}$ existen funciones $\{u_\theta : \theta \in \Theta\}$ tales que $u(q,\theta) = u_\theta\big(q(\theta)\big)$.

Análogamente al caso discreto, caracterizaremos las funciones de puntaje local:

4.34. Proposición. *Si $u : \mathcal{Q} \times \Theta \to \mathbb{R}$ es una función de puntaje local, suave y propia, entonces debe ser de la forma*

$$u(q,\theta) = A \log q(\theta) + B(\theta)\,,$$

en donde $A > 0$ es una constante arbitraria y $B(\,\cdot\,)$ es cualquier función siempre y cuando se garantice la existencia de $\overline{u}(q)$.

Demostración. Maximizamos respecto a $q \in \mathcal{Q}$ la utilidad esperada

$$\overline{u}(q) = \int_\Theta u(q,\theta)\, p(\theta)\, d\theta$$

sujeto a la condición $\int_\Theta q(\theta)\, d\theta = 1$. Como u es local, lo anterior se reduce a encontrar un punto extremo de

$$F(q) := \int_\Theta u_\theta\big(q(\theta)\big)\, p(\theta)\, d\theta - A\Big[\int_\Theta q(\theta)\, d\theta - 1\Big]\,.$$

Para que F sea estacionaria en alguna $q \in \mathcal{Q}$ es necesario que

$$\frac{\partial}{\partial\alpha} F\big(q(\theta) + \alpha\tau(\theta)\big) \restriction_{\alpha=0} = 0$$

para cualquier función $\tau : \Theta \to \mathbb{R}$ con norma suficientemente pequeña (ver Jeffreys y Jeffreys (1946), Capítulo 10). Esta condición se reduce a la ecuación diferencial

$$D\, u_\theta\big(q(\theta)\big)\, p(\theta) - A = 0\,,$$

en donde $D\, u_\theta$ denota la primera derivada de u_θ. Como u_θ es propia entonces el máximo de $F(q)$ debe alcanzarse cuando $q = p$ por lo que una función de puntaje local, suave y propia debe satisfacer la ecuación diferencial

$$D\, u_\theta\big(p(\theta)\big)\, p(\theta) - A = 0\,,$$

de donde se obtiene que $u_\theta\big(p(\theta)\big) = A \log p(\theta) + B(\theta)$. □

4.35. Definición. Una *función de puntaje logarítmica* para las distribuciones de probabilidad $q \in \mathcal{Q}$ definidas sobre Θ es una función $u : \mathcal{Q} \times \Theta \to \mathbb{R}$ de la forma

$$u(q,\theta) = A \log q(\theta) + B(\theta)\,,$$

En donde $A > 0$ es una constante arbitraria y $B(\,\cdot\,)$ es cualquier función que garantice la existencia de $\overline{u}(q)$ para todo $q \in \mathcal{Q}$.

4.36. Proposición. *Si nuestras preferencias están descritas por una función de puntaje logarítmica, la pérdida esperada de utilidad al usar una densidad de probabilidades q en vez de p está dada por:*

$$\delta(q \,|\, p) = A \int_\Theta p(\theta) \log \frac{p(\theta)}{q(\theta)}\, d\theta\,.$$

Más aún, $\delta(q \,|\, p) \geq 0$ con igualdad si y sólo si $q = p$ casi seguramente.

Demostración. Análoga a la del caso discreto. □

4.37. Definición. La *discrepancia logarítmica* de una densidad de probabilidades p estrictamente positiva sobre Θ respecto a una aproximación $\hat{p}$ está definida por:

$$\delta(p \,|\, \hat{p}) := A \int_{\Theta} p(\theta) \log \frac{p(\theta)}{\hat{p}(\theta)} \, d\theta \,.$$

Ejemplo 15. Utilizando discrepancia logarítmica, la mejor aproximación normal $N(x \,|\, \mu, \lambda)$ para cualquier variable aleatoria absolutamente continua X que toma valores en todo $\mathbb{R}$ con función de densidad f_X y con primeros dos momentos finitos tales que $\mathbb{E}(X) = m$ y $\mathbb{V}(X) = \tau^{-1}$ es aquélla que utiliza $\mu = m$ y $\lambda = \tau$. (Recuerde que λ es la *precisión*, que es el inverso de la varianza). Los detalles se dejan como ejercicio.

4.38. Proposición. *Si nuestras preferencias están descritas por una función de puntaje logarítmica para la clase de densidades de probabilidad $p(\theta \,|\, \mathbf{x})$ definidas sobre Θ, entonces el incremento esperado en utilidad proveniente de los datos $\mathbf{x}$, cuando la densidad de probabilidades a priori es $p(\theta)$, está dado por*

$$A \int_{\Theta} p(\theta \,|\, \mathbf{x}) \log \frac{p(\theta \,|\, \mathbf{x})}{p(\theta)} \, d\theta \,,$$

en donde $p(\theta \,|\, \mathbf{x})$ es la densidad a posteriori de θ dado $\mathbf{x}$, cantidad que resulta ser no negativa, y cero si y sólo si $p(\theta \,|\, \mathbf{x}) = p(\theta)$.

Demostración. Análoga a la del caso discreto. □

4.39. Definición. La *cantidad de información de los datos* acerca de Θ proveniente de los datos $\mathbf{x}$ cuando la distribución a priori es $p(\theta)$ se define como:

$$I(\mathbf{x} \,|\, p(\theta)) := \int_{\Theta} p(\theta \,|\, \mathbf{x}) \log \frac{p(\theta \,|\, \mathbf{x})}{p(\theta)} \, d\theta \,,$$

en donde $p(\theta \,|\, \mathbf{x})$ es la distribución a posteriori correspondiente. Es decir,

$$I(\mathbf{x} \,|\, p(\theta)) := \delta(p(\theta) \,|\, p(\theta \,|\, \mathbf{x}))$$

4.40. Definición. La *información esperada de un experimento* e acerca de Θ cuando la distribución a priori es $p(\theta)$ está definida por:

$$I(e \,|\, p(\theta)) := \int_{\mathcal{X}} I(\mathbf{x} \,|\, p(\theta))\, p(\mathbf{x} \,|\, e)\, d\mathbf{x}$$

en donde $p(\mathbf{x} \,|\, e)$ es la distribución de probabilidad sobre el conjunto $\mathcal{X}$ de los resultados posibles del experimento.

§ EJERCICIOS

1. Una compañía debe decidir si acepta o rechaza un lote de artículos (considere estas acciones como a_1 y a_2, respectivamente). Los lotes pueden ser de tres tipos: E_1 (muy bueno), E_2 (aceptable) y E_3 (malo). La función de utilidad se presenta en la siguiente tabla:

$u(a_i, \theta_j)$	E_1	E_2	E_3
a_1	3	2	0
a_2	0	1	3

   La compañía supone que los eventos E_1, E_2 y E_3 son equiprobables.

   a) Describe la estructura del problema de decisión.

   b) Determina las acciones admisibles.

   c) Resuelve el problema de decisión utilizando el criterio de la utilidad esperada máxima (Definición 4.8).

   d) Determina todas las distribuciones de probabilidad sobre el espacio de estados tales que se obtiene la misma solución del inciso anterior.

2. Un alumno tiene que presentar examen final de un curso y le queda poco tiempo para estudiar. Supongamos que $\mathcal{A} := \{a_1, a_2, a_3\}$ donde:

   $a_1 :=$ Estudiar con detalle la primera parte del curso y nada de la segunda
   $a_2 :=$ Estudiar con detalle la segunda parte y nada de la primera
   $a_3 :=$ Estudiar con poco detalle todo el curso

   Y supongamos que el espacio de estados es $\Theta := \{E_1, E_2, E_3\}$ donde:

   $E_1 :=$ El examen está más cargado hacia la primera parte
   $E_2 :=$ El examen está más cargado hacia la segunda parte
   $E_3 :=$ El examen está equilibrado

   Aunque no se se tiene una función de probabilidad sobre Θ supongamos que resulta razonable suponer que $\mathbb{P}(E_2) > \mathbb{P}(E_1)$ y que $\mathbb{P}(E_2) > \mathbb{P}(E_3)$.

Definimos ahora una función de utilidad sobre el conjunto de consecuencias $\mathcal{C}$ que de hecho será la calificación que podrá el estudiante obtener:

$u(a_i, E_j)$	E_1	E_2	E_3
a_1	9	2	5
a_2	2	9	5
a_3	6	6	7

Verifique que de acuerdo a las restricciones del problema la acción a_1 nunca tiene la posibilidad de ser la acción óptima y determine los conjuntos de valores de $\mathbb{P}(E_1), \mathbb{P}(E_2)$ y $\mathbb{P}(E_3)$ para los cuales $a_2 \succ a_3$, $a_2 \prec a_3$ y $a_2 \sim a_3$.

3. Considere un problema de decisión en donde el conjunto de acciones $\mathcal{A}$ y el espacio de estados Θ tienen un número infinito numerable de elementos. Supongamos que $\mathcal{A} := \{a_0, a_1, a_2, \ldots\}$ y $\Theta := \{E_1, E_2, \ldots\}$ y que la función de utilidad está dada por la siguiente tabla:

$u(a_i, E_j)$	E_1	E_2	E_3	E_4	E_5	$\ldots$
a_0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\ldots$
a_1	1	0	0	0	0	$\ldots$
a_2	1	1	0	0	0	$\ldots$
a_3	1	1	1	0	0	$\ldots$
a_4	1	1	1	1	0	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Demuestre que la única acción admisible es a_0 y que a_0 no satisface el criterio de la utilidad esperada máxima, sea cual sea la distribución de probabilidad sobre Θ.

4. Considere el siguiente problema de decisión. En un juego, se tiene un conjunto de 9 cartas que contiene: 2 Ases, 3 Reyes y 4 Sotas. Al jugador, quien paga \$150 por el derecho de jugar, se le entrega una carta al azar y, una vez con esta carta en su poder, puede optar por pedir otra carta o bien pasar. Si decide pasar, simplemente pierde su pago inicial. Si, por el contrario, pide otra carta, las recompensas se pagan de acuerdo

a la siguiente tabla:

Cartas	Recompensa
2 Ases o 2 Reyes	+$2000
2 Sotas o 1 As y 1 Sota	+$1000
otra combinación	−$1000

a) Describa la estructura del problema y obtenga la decisión óptima para un jugador que ya pagó su derecho de juego...

a.1) ... si resuelve decidir sin mirar la primera carta;

a.2) ... si resuelve decidir sólo después de observar la primera carta;

b) ¿Es preferible mirar la primera carta antes de decidir si se pide una segunda carta o resulta indiferente?

5. Verifique los resultados del Ejemplo 13.

6. Del Ejemplo 14 simule y grafique $I(D \,|\, \mathbf{p}_0)$ para distintos valores de n y compare su comportamiento con el de $p(x = 1 \,|\, \mathbf{x})$ bajo los dos escenarios posibles.

7. Demuestre que, utilizando discrepancia logarítmica, la mejor aproximación normal $N(x \,|\, \mu, \lambda)$ para cualquier variable aleatoria absolutamente continua X que toma valores en todo $\mathbb{R}$ con función de densidad f_X y con primeros dos momentos finitos tales que $\mathbb{E}(X) = m$ y $\mathbb{V}(X) = \tau^{-1}$ es aquélla que utiliza $\mu = m$ y $\lambda = \tau$. (Recuerde que λ es la *precisión*, que es el inverso de la varianza.)

Capítulo 5

Inferencia estadística paramétrica bayesiana

El objetivo de haber revisado en el capítulo anterior algunos conceptos y resultados importantes de teoría de la decisión es justamente resolver problemas de inferencia estadística como problemas de decisión. En cada caso supondremos que se tiene un fenómeno aleatorio de interés que se modela mediante un vector (o variable) aleatorio X cuya distribución de probabilidad pertenece a una familia paramétrica $\mathcal{P} := \{p(x \,|\, \theta) \,:\, \theta \in \Theta\}$ y que se cuenta con una distribución de probabilidad sobre el espacio paramétrico Θ, ya sea a priori o a posteriori, denotadas $p(\theta)$ o $p(\theta \,|\, \mathbf{x})$, respectivamente. Utilizaremos $p(\theta)$ en el entendido de que puede tratarse de cualquiera de las dos anteriores, salvo especificación en contrario. De igual modo utilizaremos indistintamente la distribuciones predictiva a priori $p(x)$ y la distribución predictiva a posteriori $p(x \,|\, \mathbf{x})$.

5.1. Estimación puntual

El problema de la estimación puntual se plantea como un problema de decisión $(\mathcal{E}, \mathcal{C}, \mathcal{A}, \preceq)$ en donde el espacio de estados es justamente el espacio paramétrico Θ y el conjunto de acciones es también $\mathcal{A} = \Theta$, en el sentido de que habremos de tomar la acción de escoger un valor particular para θ. Para evitar confusión, a los elementos de $\mathcal{A}$ los denotaremos mediante $\hat{\theta}$. La función de utilidad será entonces una función $u : \Theta \times \Theta \to \mathbb{R}$.

5.1. Definición. La *estimación puntual* de θ respecto a la función de utilidad $u(\hat{\theta}, \theta)$ y a una distribución de probabilidad $p(\theta)$ sobre Θ es la acción óptima $\hat{\theta}^* \in \mathcal{A} = \Theta$ tal que

$$\overline{u}(\hat{\theta}^*) = \max_{\hat{\theta} \in \Theta} \overline{u}(\hat{\theta}) \,,$$

en donde

$$\overline{u}(\hat{\theta}) = \int_{\Theta} u(\hat{\theta}, \theta)\, p(\theta)\, d\theta \;=: \mathbb{E}_{\theta}\big[\, u(\hat{\theta}, \theta) \,\big] \,.$$

Ejemplo 16. Supongamos que $\Theta \subset \mathbb{R}$ y que escogemos la función de utilidad cuadrática $u(\hat{\theta}, \theta) := -(\hat{\theta} - \theta)^2$. Entonces

$$\begin{aligned}
\overline{u}(\hat{\theta}) &= \mathbb{E}_{\theta}\big[\, u(\hat{\theta}, \theta) \,\big] \\
&= -\mathbb{E}_{\theta}\big[\, (\hat{\theta} - \theta)^2 \,\big] \\
&= -\mathbb{E}_{\theta}\big[\, (\hat{\theta} - \mathbb{E}_{\theta}(\theta) + \mathbb{E}_{\theta}(\theta) - \theta)^2 \,\big] \\
&= -\mathbb{E}_{\theta}\big[\, (\hat{\theta} - \mathbb{E}_{\theta}(\theta))^2 \,\big] - 2\,\mathbb{E}_{\theta}\big[\, (\hat{\theta} - \mathbb{E}_{\theta}(\theta))(\mathbb{E}_{\theta}(\theta) - \theta) \,\big] - \mathbb{E}_{\theta}\big[\, (\theta - \mathbb{E}_{\theta}(\theta))^2 \,\big]
\end{aligned}$$

en donde el segundo término es cero y el tercero es la varianza de θ por lo que

$$\overline{u}(\hat{\theta}) = -\Big\{ \big[\, \hat{\theta} - \mathbb{E}_{\theta}(\theta) \,\big]^2 + \mathbb{V}_{\theta}(\theta) \Big\} \,.$$

El estimador puntual de θ es $\hat{\theta}^* \in \mathcal{A} = \Theta$ tal que

$$\begin{aligned}
\overline{u}(\hat{\theta}^*) &= \max_{\hat{\theta} \in \Theta} \left(-\Big\{ \big[\, \hat{\theta} - \mathbb{E}_{\theta}(\theta) \,\big]^2 + \mathbb{V}_{\theta}(\theta) \Big\} \right) \\
&= \min_{\hat{\theta} \in \Theta} \Big\{ \big[\, \hat{\theta} - \mathbb{E}_{\theta}(\theta) \,\big]^2 + \mathbb{V}_{\theta}(\theta) \Big\}
\end{aligned}$$

de donde se obtiene que

$$\hat{\theta}^* = \mathbb{E}_{\theta}(\theta) = \int_{\Theta} \theta\, p(\theta)\, d\theta \,,$$

siempre y cuando dicha esperanza exista, por supuesto. $\diamond$

Generalizando el ejemplo anterior al caso en que $\Theta \subset \mathbb{R}^k$, si se tiene como función de utilidad la forma cuadrática

$$u(\hat{\theta}, \theta) = -(\hat{\theta} - \theta)^T H (\hat{\theta} - \theta)$$

entonces

$$\overline{u}(\hat{\theta}) = -\int_{\Theta} (\hat{\theta} - \theta)^T H (\hat{\theta} - \theta)\, p(\theta)\, d\theta\,.$$

Derivando $\overline{u}(\hat{\theta})$ respecto a $\hat{\theta}$ e igualando a cero obtenemos

$$-2H \int_{\Theta} (\hat{\theta} - \theta)\, p(\theta)\, d\theta = 0$$

de donde $H\hat{\theta} = H\,\mathbb{E}_{\theta}(\theta)$. Si H^{-1} existe entonces

$$\hat{\theta^*} = \mathbb{E}_{\theta}(\theta) = \int_{\Theta} \theta\, p(\theta)\, d\theta\,,$$

siempre y cuando dicha esperanza exista, por supuesto. Nótese que lo anterior, más que un ejemplo, es un resultado de carácter general.

El resultado es análogo si lo que se desea es estimar puntualmente una observación futura de X, misma que denotaremos $\hat{x}^*$. En este caso el espacio de estados es $RanX$ (el conjunto de todos los valores posibles de X) y por lo tanto la estimación puntual de una observación futura de X respecto a la función de utilidad $u(\hat{x}, x)$ y a una distribución predictiva $p(x)$ es la acción óptima $\hat{x}^* \in RanX$ tal que

$$\overline{u}(\hat{x}^*) = \max_{\hat{x} \in\ RanX} \overline{u}(\hat{x})\,,$$

en donde

$$\overline{u}(\hat{x}) = \int_{RanX} u(\hat{x}, x)\, p(x)\, dx \;=: \mathbb{E}_{p(x)}\big[\, u(\hat{x}, x)\,\big]\,.$$

Y nuevamente, si la función de utilidad es la de el ejemplo anterior entonces:

$$\hat{x}^* \;=\; \int_{RanX} x\, p(x)\, dx\,.$$

Ejemplo 17. Nos remitimos al Ejercicio 3 del Capítulo 3, pero aquí supondremos que de acuerdo al responsable de la caseta en el mayor de los casos el número promedio de autos por minuto es 8 (y no 12). De acuerdo a un procedimiento similar al que se utiliza para resolver dicho ejercicio se obtiene como distribución a priori para λ una distribución Gamma con hiperparámetros $\alpha = 9.108$ y $\beta = 0.01012$ y por tanto la distribución a posteriori de λ es Gamma también:

$$p(\lambda \,|\, \mathbf{x}) = Ga(\lambda \,|\, 9.108 + \sum x_j\,,\, 0.01012 + n)\,.$$

Suponiendo que como información muestral tenemos $\mathbf{x} = (679\,,\,703)$ entonces

$$p(\lambda \,|\, \mathbf{x}) = Ga(\lambda \,|\, 1391.108\,,\, 2.01012)\,.$$

Utilizando la función de utilidad del Ejemplo 16 obtenemos como estimación puntual de λ :

$$\hat{\lambda}^* = \frac{1391.108}{2.01012} = 692.05\,.$$

5.2. Contraste de hipótesis

Supongamos que se tienen m hipótesis acerca del parámetro (o vector de parámetros) θ :

$$H_1 : \theta \in \Theta_1 \quad , \quad H_2 : \theta \in \Theta_2 \quad , \quad \ldots \quad , \quad H_m : \theta \in \Theta_m\,,$$

en donde los conjuntos Θ_j son subconjuntos del espacio paramétrico Θ . Podemos suponer que los subconjuntos Θ_j son disjuntos. En caso de que no lo fueren, los redifinimos de modo que sí lo sean. Por ejemplo, si $\Theta_i \cap \Theta_k \neq \varnothing$ definimos un nuevo subconjunto $\Theta_j := \Theta_i \cap \Theta_k$ y redefinimos Θ_i como $\Theta_i \setminus \Theta_j$ y a Θ_k como $\Theta_k \setminus \Theta_j$, además de agregar la hipótesis $H_j : \theta \in \Theta_j$.

En caso de que $\cup_{j=1}^{m} \Theta_j \neq \Theta$ definimos el subconjunto $\Theta_{m+1} := \Theta \setminus \cup_{j=1}^{m} \Theta_j$ de modo que $\{\Theta_j : j = 1, \ldots, m+1\}$ sea partición de Θ , además de agregar la hipótesis $H_{m+1} : \theta \in \Theta_{m+1}$.

Con base en lo anterior, supongamos entonces que los subconjuntos $\Theta_1\,, \ldots\,, \Theta_m$ constituyen una partición del espacio paramétrico Θ . El contraste de las hipótesis $H_1\,, \ldots\,, H_m$ se plantea como un problema de decisión que consiste en escoger una de estas hipótesis. Definimos como espacio de estados

$$\Phi := \{H_1\,,\, \ldots\,,\, H_m\}\,,$$

y como medida de probabilidad $\mathbb{P}$ sobre Φ utilizamos la distribución a priori o a posteriori (según sea el caso) de θ ya que

$$\mathbb{P}(H_j) = \mathbb{P}(\theta \in \Theta_j) = \int_{\Theta_j} p(\theta \,|\, \mathbf{x})\, d\theta\,.$$

Consideremos el conjunto de acciones $\mathcal{A} := \{a_1, \ldots, a_m\}$ en donde a_j representa la acción de actuar como si la hipótesis H_j fuese a ocurrir. Si se define además una función de utilidad $u : \mathcal{A} \times \Phi \to \mathbb{R}$ podemos entonces resolver el problema de elegir una de las hipótesis como un problema de decisión simple: se elige aquella $a_* \in \mathcal{A}$ tal que

$$\overline{u}(a_*) = \max_{a_i \in \mathcal{A}} \overline{u}(a_i)$$

en donde

$$\overline{u}(a_i) = \sum_{j=1}^{m} u(a_i, H_j)\mathbb{P}(H_j)\,.$$

En caso de que se tengan hipótesis pero en relación a una observación futura de la variable o vector aleatorio $X \sim p(x \,|\, \theta)$ el procedimiento es análogo:

$$H_1 : X \in \mathcal{X}_1 \quad , \quad \ldots \quad , \quad H_m : X \in \mathcal{X}_m\,,$$

en donde $\mathcal{X} = Ran\, X$ y $\{\mathcal{X}_1, \ldots, \mathcal{X}_m\}$ es partición de $\mathcal{X}$. Como medida de probabilidad $\mathbb{P}$ sobre el espacio de estados Φ utilizamos la distribución predictiva a priori o a posteriori (según sea el caso) ya que

$$\mathbb{P}(H_j) = \mathbb{P}(X \in \mathcal{X}_j) = \int_{\mathcal{X}_j} p(x \,|\, \mathbf{x})\, d\theta\,.$$

Pero... ¿qué función de utilidad ocupar? Dependerá de las características de cada problema y de todo aquello que se desee sea tomado en consideración. Por ejemplo, si utilizamos una función de utilidad muy simple como

$$u(a_i, H_j) := \mathbf{1}_{\{i=j\}}$$

obtenemos

$$\overline{u}(a_i) = \sum_{j=1}^{m} \mathbf{1}_{\{i=j\}}\mathbb{P}(H_j) = \mathbb{P}(H_i) = \int_{\Theta_i} p(\theta \,|\, \mathbf{x})\, d\theta$$

y por lo tanto la solución óptima será aquella $a_* \in \mathcal{A}$ tal que

$$\overline{u}(a_*) = \max_{j=1,\ldots,m} \mathbb{P}(H_j)\,,$$

es decir, bajo esta función de utilidad la decisión óptima es escoger aquella hipótesis que tenga la probabilidad más alta de ocurrir. Dijimos que esta función de utilidad es demasiado simple porque sólo toma en consideración la probabilidad de cada hipótesis, sin tomar en cuenta aspectos de otra índole que pudiera ser de interés tomar también en consideración (por ejemplo, consideraciones de tipo económico) como se ilustrará en el siguiente:

Ejemplo 18. Continuando con el Ejemplo 17 supongamos, de manera simplificada, que de acuerdo a normas de la Secretaría de Comunicaciones y Transportes el número de cobradores que se deben de tener en dicha caseta va de acuerdo al número de autos que llegan, y que en el caso particular de los viernes de 5 a 8 p.m. resulta como sigue:

Núm. de autos	Núm. de cobradores
0 a 690	5
691 a 750	10
más de 750	15

El responsable de la caseta está en libertad de contratar con anticipación al número cobradores que considere pertinentes para cada viernes de acuerdo a sus expectativas de aforo vehicular. Pero si sus expectativas se ven rebasadas tendrá que contratar cobradores emergentes. Supongamos que un cobrador contratado con anticipación cuesta $300 pesos pero uno contratado de última hora (emergente) cuesta $700. De acuerdo a la información que se tiene (Ejemplo 17) el responsable de la caseta desea tomar una decisión óptima en cuanto al número de cobradores a contratar con anticipación (5, 10 o 15).

Lo anterior se puede plantear como un contraste de hipótesis:

$$H_1 : X \in \{0, 1, \ldots, 690\}\,,\ H_2 : X \in \{691, \ldots, 750\}\,,\ H_3 : X \in \{751, 752, \ldots\}\,,$$

en donde, recordemos, X representa el número de autos que llegan a la caseta. Con la información del Ejemplo 17 así como del Ejemplo 3 tenemos que la distribución predictiva a posteriori es Poisson-Gamma:

$$p(x \,|\, \mathbf{x}) = Pg(x \,|\, 1391.108, 2.01012, 1)$$

con lo que

$$\mathbb{P}(H_1) = 0.4849 \quad , \quad \mathbb{P}(H_2) = 0.4786 \quad , \quad \mathbb{P}(H_3) = 0.0365\,.$$

Sólo nos falta la función de utilidad, que en este caso está implícita en las condiciones mismas del problema:

$\mathbb{P}(H_j)$	0.4849	0.4786	0.0365	
$u(a_i, H_j)$	H_1	H_2	H_3	$\overline{u}(a_i)$
a_1	$-\$1500$	$-\$5000$	$-\$8500$	$-\$3,430.60$
a_2	$-\$3000$	$-\$3000$	$-\$6500$	$-\$3,127.75$
a_3	$-\$4500$	$-\$4500$	$-\$4500$	$-\$4,500.00$

Por ejemplo $u(a_2, H_3) = -\$6500$ porque en este caso la acción a_2 implica contratar 10 cobradores con anticipación y si el escenario que ocurre finalmente es H_3 entonces esto implica contratar 5 cobradores emergentes y por ello el desembolso total es de $10 \times \$300 + 5 \times \$700 = \$6500$.

De acuerdo a lo anterior tenemos que la solución óptima es a_2 por tener la máxima utilidad esperada. Nótese que a_2 implica actuar como si H_2 fuese a ocurrir, y H_2 no es precisamente la hipótesis que tiene la mayor probabilidad de cumplirse. Esto es porque en este caso la función de utilidad tomó en cuenta no sólo las probabilidades de los escenarios sino también la intensidad de sus consecuencias económicas.

5.3. Estimación por regiones

En ocasiones, la descripción de la información sobre θ (o bien sobre una observación futura de X) a través de $p(\theta \,|\, \mathbf{x})$ (o bien $p(x \,|\, \mathbf{x})$) no resulta accesible para cierto tipo de usuarios de la estadística, a quienes resulta preferible obtener regiones (subconjuntos) $C \subset \Theta$ (o bien $C \subset \mathcal{X} = Ran\, X$) que tengan una probabilidad dada de contener al valor correcto de θ (o de una observación futura de X). De la construcción de estas regiones nos ocupamos en esta sección.

5.2. Definición. Una región (o subconjunto) $C \subset \Theta$ tal que

$$\int_C p(\theta \,|\, \mathbf{x})\, d\theta = \alpha$$

en donde $0 \leq \alpha \leq 1$ es llamada *región de probabilidad* α para θ con respecto a $p(\theta \,|\, \mathbf{x})$.

Nótese que C no es necesariamente un intervalo. La solución para C en la ecuación $\int_C p(\theta \,|\, \mathbf{x})\, d\theta = \alpha$ no es única y por tanto podemos hablar del conjunto de soluciones

$$\mathcal{A} := \{C \subset \Theta : \int_C p(\theta \,|\, \mathbf{x})\, d\theta = \alpha\}\,,$$

lo cual implica la necesidad de definir un criterio adicional para elegir una región C adecuada. Esto se puede resolver como un problema de decisión en donde el conjunto $\mathcal{A}$ que acabamos de definir es el conjunto de acciones (es decir, cada acción es una de las distintas regiones que podemos elegir), el espacio de estados es el espacio paramétrico Θ cuya medida de probabilidad queda definida mediante $p(\theta \,|\, \mathbf{x})$. Sólo nos hace falta una función de utilidad que contenga ese criterio adicional, que puede ser, por ejemplo, el preferir regiones C que tengan el menor tamaño posible, mismo que denotaremos $\|C\|$, pero que contengan al valor correcto de θ :

$$u(C, \theta) = -k\|C\| + \mathbf{1}_C(\theta)\,, \quad k > 0\,.$$

Mediante esta función de utilidad obtenemos la utilidad esperada para cada $C \in \mathcal{A}$ mediante

$$\overline{u}(C) = \int_\Theta u(C, \theta) p(\theta \,|\, \mathbf{x})\, d\theta = -k\|C\| + \alpha\,,$$

de donde es claro entonces que la región óptima será aquélla $C^* \in \mathcal{A}$ tal que su tamaño $\|C^*\|$ sea mínimo. A tal C^* se le denomina *región de probabilidad α de máxima densidad*.

Ejemplo 19. Utilizaremos la información del Ejemplo 17. Aunque ya dijimos que las regiones que se pueden construir no son necesariamente intervalos, supongamos en este caso que deseamos construir un intervalo de probabilidad 0.95 de máxima densidad para λ. Representemos dicho intervalo mediante $[\,\lambda_1, \lambda_2\,]$. Entonces el problema consiste en encontrar los valores para λ_1 y λ_2 tal que $\mathbb{P}(\lambda \in [\,\lambda_1, \lambda_2\,]) = 0.95$ y que la longitud del intervalo $[\,\lambda_1, \lambda_2\,]$ sea mínima. Esto es

$$\begin{aligned} &\text{minimizar:} \qquad h(\lambda_1, \lambda_2) = \lambda_2 - \lambda_1 \\ &\text{sujeto a:} \quad \int_{\lambda_1}^{\lambda_2} Ga(\lambda \,|\, 1391.108\,,\, 2.01012)\, d\lambda = 0.95\,. \end{aligned}$$

Resolviendo numéricamente lo anterior obtenemos como solución óptima el intervalo $[\,655.88\,,\, 728.6\,]$.

§ EJERCICIOS

1. Obtenga el estimador puntual $\hat{\theta}^* \in \Theta \subset \mathbb{R}$ bajo las siguientes funciones de utilidad:

   a) $u(\hat{\theta}, \theta) = k|\hat{\theta} - \theta|$ con $k < 0$ una constante.

   b) $u(\hat{\theta}, \theta) = -\Big(\dfrac{\hat{\theta} - \theta}{\hat{\theta}}\Big)^2$.

2. Respecto al Ejemplo 2 supongamos que de los 150 expedientes revisados 17 resultan incompletos. Obtenga estimaciones puntuales de θ bajo las siguientes funciones de utilidad:

   a) $u(\hat{\theta}, \theta) = -(\hat{\theta} - \theta)^2$.

   b) $u(\hat{\theta}, \theta) = -|\hat{\theta} - \theta|$.

   c) $u(\hat{\theta}, \theta) = -\Big(\dfrac{\hat{\theta} - \theta}{\hat{\theta}}\Big)^2$.

3. Una máquina produce cierto componente que debe tener una longitud especificada. Sea la variable aletoria X igual al margen de error en dicha longitud y supongamos que se distribuye Normal con media cero y precisión $\lambda > 0$ desconocida. Suponga que no cuenta con información a priori y que obtiene una muestra

   $$\mathbf{x} = (0.033, 0.002, -0.019, 0.013, -0.008, -0.0211, 0.009, 0.021, -0.015)\,.$$

   Construya un intervalo de probabilidad 0.95 de máxima densidad para el margen de error en la longitud de dicha componente.

Bibliografía

Bernardo, J.M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society, Serie B* **41**, 113-147.

Bernardo, J.M. y Smith, A.F.M (1994). *Bayesian Theory*, Wiley.

Casella, G. y Berger, R.L. (1990). *Statistical Inference*, Wadsworth.

Gelman, Carlin, Stern y Rubin (1995). *Bayesian Data Analysis*, Chapman & Hall.

Gutiérrez Peña, E.A. (1995). *Bayesian Topics Relating to the Exponential Family*, tesis doctoral, University of London.

Gutiérrez Peña, E.A. (1998). Análisis bayesiano de modelos jerárquicos lineales. *Monografías IIMAS-UNAM* **7**, Núm. 16.

Jeffreys,H. y Jeffreys, B.S (1946/1972). *Methods of Mathematical Physics*, Cambridge University Press.

Jeffreys, H. (1961). *Theory of Probability*, Oxford University Press.

Lehmann, E.L. y Casella, G. (1998). *Theory of Point Estimation*, Springer.

Lindley, D.V. (2000). The Philosophy of Statistics. *The Statistician* **49**, 293-337.

Migon, H.S y Gammerman, D. (1999). *Statistical inference: an integrated approach*, Oxford University Press.

Press, S.J. (1989). *Bayesian Statistics: Principles, models, and applications*, Wiley.