Estadística Matemática

Fernando Tusell[1]

19 de septiembre de 2007

[1]Bastantes errores menos en esta versión son consecuencia de los comentarios recibidos de Araceli Garín, Vicente Núñez y de Mario S. de Juan y Pedro A. Gómez (curso 1.999-2.000). Todavía faltan muchos temas del programa por desarrollar, y otros están a medio escribir, tienen errores u obscuridades. Correcciones y comentarios son bienvenidos.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Elementos de Teoría de la Decisión.

1.1. Qué es un procedimiento estadístico.

Nos enfrentamos a una colección $\Theta = \{\theta_i, i \in I\}$ de posibles *estados de la naturaleza*, o simplemente *estados*[1]. No podemos observar directamente cuál es el θ_i que prevalece.

Nos enfrentamos también a un conjunto de decisiones que podemos tomar, o *espacio de decisión* $D = \{d_j, j \in J\}$. Existe, por fin, una *función de pérdida* $L\colon \Theta \times D \longrightarrow R$ completamente especificada, proporcionando las pérdidas asociadas a cada par (θ_i, d_j); $L(\theta_i, d_j)$ es la pérdida derivada de tomar la decisión d_j cuando el estado de la naturaleza es θ_i. Obviamente, si θ_i fuera observable, no tendríamos ningún problema en seleccionar en cada caso la decisión d_j óptima, que minimiza L.

Asociada a cada estado θ_i suponemos una distribución $F_{X|\theta}(x\,|\theta)$ generando una cierta variable aleatoria observable, X. Esta variable aleatoria toma valores en un conjunto S. Podemos muestrear la población $F_{X|\theta}(x\,|\theta)$ y obtener valores de X mediante la realización de un *experimento*. Los valores que observemos son toda la evidencia de que disponemos para conjeturar cuál es el estado de la naturaleza vigente, y en consecuencia la decisión óptima.

De un modo informal, un *procedimiento estadístico* es una regla para escoger una decisión d_j a la vista del valor x que toma X (o quizá del conjunto de valores $\boldsymbol{x}$ que toman n observaciones de X, en el caso de que nos sea posible disponer

[1] El conjunto de índices I es finito o infinito; ni siquiera ha de ser numerable, como pondrán de manifiesto los ejemplos a continuación.

de más de una). Más precisamente, un procedimiento estadístico es una aplicación $\delta\colon S \longrightarrow D$, que al resultado de cada experimento hace corresponder una decisión[2].

Aunque aparentemente muy abstracto, el marco anterior engloba de forma general lo que habitualmente estamos acostumbrados a llamar procedimientos estadísticos, como ponen de manifiesto los siguientes ejemplos.

Ejemplo 1.1 Consideremos el caso en que nos enfrentamos a una población de sujetos caracterizados por sufrir o no una enfermedad. Deseamos estimar por punto la proporción de los afectados, θ, con ayuda de una muestra de sujetos de tamaño n. El conjunto de posibles estados de la naturaleza sería $\Theta = \{\theta\colon \theta \in R, 0 \leq \theta \leq 1\}$, y el espacio de decisión sería $D = \{d\colon d \in R, 0 \leq d \leq 1\}$. Diferentes criterios de estimación podrían además contemplarse como reflejo de la utilización de diferentes funciones de pérdida. Por ejemplo, la estimación mínimo cuadrática se originaría como consecuencia de emplear una función de pérdida cuadrática, $L(\theta, \hat{\theta})$; otras posibilidades serían una pérdida "valor absoluto", $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$, o "cero-uno",

$$L(\theta, \hat{\theta}) = \begin{cases} 0 & \text{si } |\theta - \hat{\theta}| < b, \\ c & \text{en otro caso.} \end{cases}$$

Ejemplo 1.2 Si en el Ejemplo 1.1 deseáramos realizar estimación por intervalo en lugar de por punto, podríamos considerar como espacio de decisión el formado por todos los intervalos (θ_1, θ_2). La decisión consistiría en escoger uno de tales intervalos.

En este caso, sin embargo, no es nada obvio cuál haya de ser la pérdida a emplear. Podríamos pensar, a imagen del ejemplo anterior, en emplear una pérdida que fuera nula si el intervalo realmente contiene al parámetro, y mayor que cero, quizá constante, en caso contrario. Es decir,

$$L(\theta, d = (\hat{\theta}_1, \hat{\theta}_2)) = \begin{cases} 0 & \text{si } \theta \in (\hat{\theta}_1, \hat{\theta}_2), \\ c & \text{en otro caso.} \end{cases}$$

Pero ello no tiene mucho sentido: haría óptimos intervalos como $(-\infty, \infty)$. La pérdida parece que debiera tomar en cuenta la amplitud del intervalo construido. Véase Meeden y Varderman (1985).

Ejemplo 1.3 Supongamos que debemos aceptar o rechazar un lote de piezas, dependiendo de la fracción de defectuosas que contenga. En este caso, Θ sería el intervalo $[0, 1]$ (cada estado correspondería a una fracción defectiva). El espacio de decisión será: $D = \{d_1 = \text{Aceptar}, d_2 = \text{Rechazar}\}$. El experimento consistiría en tomar una o varias piezas, cada una de las cuales proporcionaría un valor de X: $X = 1$ (pieza defectuosa) o $X = 0$ (pieza correcta). El procedimiento estadístico sería entonces la regla que genera una

[2] En el caso de que el experimento consista en tomar n observaciones de X, tendríamos $\delta\colon S^n \longrightarrow D$, en que $S^n = \underbrace{S \times \ldots \times S}_{n \text{ veces}}$. Cada resultado muestral es un punto de S^n. Llamamos a S^n (ó S) *espacio muestral*.

decisión a partir del o los valores de X observados. La función de pérdida podría, al menos en principio, especificarse con facilidad. $L(\theta, d_1)$ sería el coste de aceptar una remesa con proporción defectiva θ (coincidiría quizá con el precio de las piezas en malas condiciones que hay que desechar). $L(\theta, d_2)$ sería el coste de rechazar una remesa con proporción defectiva θ (quizá el coste de los portes, o una indemnización al proveedor, si el verdadero θ estaba dentro de lo estipulado en las condiciones del pedido).

Ejemplo 1.4 El diagnóstico médico proporciona otro ejemplo de problema de decisión con función de pérdida, en general, fuertemente asimétrica. En un problema de esta naturaleza, el espacio de estados de la naturaleza es:

$$\Theta = \{\theta_1 = \text{ Paciente enfermo}, \theta_2 = \text{ Paciente sano}\}.$$

El espacio de decisiones incluye también dos: declarar al paciente sano (d_1), o enfermo (d_2). El experimento, típicamente, consiste en hacer algún tipo de análisis clínico. La función de pérdida —difícil o imposible de especificar en unidades monetarias— probablemente daría mucha mayor importancia a diagnosticar como sano a un paciente enfermo (con riesgo de agravamiento) que a diagnosticar como enfermo a uno sano (sin más trascendencia quizá que el susto o la inconveniencia de un tratamiento inadecuado).

En general, como se desprende de los ejemplos anteriores, los problemas de contraste de hipótesis o estimación de parámetros pueden ser descritos como problemas de decisión. La Teoría de la Decisión suministra un marco adecuado para plantearlos y resolverlos.

1.2. Riesgo y riesgo de Bayes.

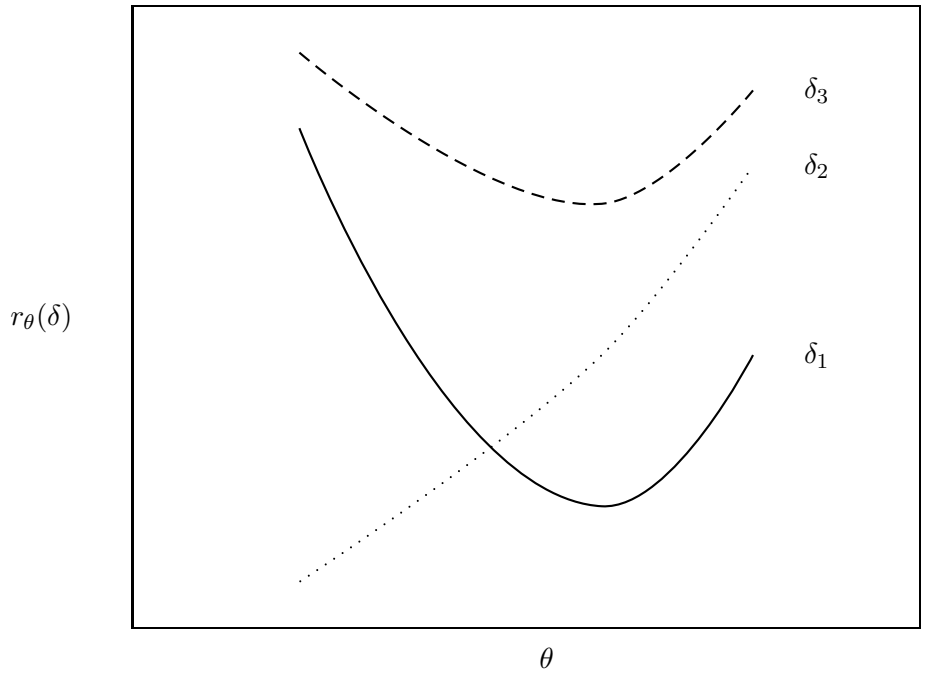
Queremos escoger nuestros procedimientos estadísticos de modo que proporcionen pérdidas reducidas. Observemos que si empleamos el procedimiento δ tomaremos la decisión $\delta(\boldsymbol{X})$, que es aleatoria: la aleatoriedad de la información muestral que utilizamos se transmite a la decisión que adoptamos y en consecuencia a la pérdida $L(\theta_i, \delta(\boldsymbol{X}))$ en que incurrimos. Tiene por ello sentido hablar del valor medio de dicha pérdida.

Definición 1.1 *Denominamos* riesgo $r_\theta(\delta)$ *al valor medio de la pérdida:*

$$r_\theta(\delta) = E_\theta L(\theta, \delta(\boldsymbol{X})) \tag{1.1}$$

El subíndice del operador de valor medio indica la distribución con respecto a la cuál se toma dicho valor medio (recuérdese que cada estado de la naturaleza θ genera X con una distribución $F_{X|\theta}(x\,|\theta)$ en general diferente). Obsérvese que se trata de una función de θ; el riesgo puede variar dependiendo del estado de la naturaleza ante el que estemos. Parece sensato considerar $r_\theta(\delta)$ para juzgar un procedimiento estadístico, pues proporciona, para cada θ, una medida promedio de la pérdida derivada de su empleo.

Figura 1.1: Procedimientos no comparables (δ_1 y δ_2) e inadmisible (δ_3)



Definición 1.2 *Sean dos procedimientos estadísticos δ_1 y δ_2. Se dice que δ_1 es* mejor *que δ_2 si $r_\theta(\delta_1) \leq r_\theta(\delta_2)\ \forall\theta \in \Theta$, con $r_\theta(\delta_1) < r_\theta(\delta_2)$ para algún θ. Análogamente, se dice que δ_1 es* equivalente *a δ_2 si $r_\theta(\delta_1) = r_\theta(\delta_2), \forall\theta$. Se dice que ambos procedimientos no son* comparables *si no son equivalentes, y ninguno de ellos mejora al otro.*

Definición 1.3 *Si un procedimiento δ_1 es mejor que otro δ_2 decimos de éste último que es* inadmisible. *Si, por el contrario, δ no puede ser mejorado por ningún otro, decimos que es* admisible.

La Figura 1.1 muestra las funciones de riesgo de tres procedimientos estadísticos. En ella, δ_1 y δ_2 no son comparables. δ_3 es inadmisible: resulta mejorado por δ_1 y por δ_2. El Ejemplo 1.5 presenta dos procedimientos, uno de ellos inadmisible al ser mejorado por el otro. Nótese que la admisibilidad o inadmisibilidad de un procedimiento depende de la función de pérdida considerada. Un procedimiento inadmisible con respecto a una función de pérdida, puede no serlo respecto de otra.

Ejemplo 1.5 Supongamos una situación como la descrita en el Ejemplo 1.3, y admitamos que la función de pérdida es cuadrática:

$$L(\theta, \delta) = (\delta - \theta)^2$$

Podemos tomar una muestra aleatoria simple formada por tres observaciones X_i, $i = 1, 2, 3$, en que $X_i = 1$ si la i-ésima pieza es defectuosa y $X_i = 0$

en caso contrario. Entonces, $X_i \sim \text{Binaria}(\theta)$. Consideremos los siguientes dos procedimientos estadísticos:

$$\delta_1(\boldsymbol{X}) = \frac{X_1 + X_2 + X_3}{3} \tag{1.2}$$

$$\delta_2(\boldsymbol{X}) = \frac{X_1 + X_3}{2} \tag{1.3}$$

Entonces:

$$r_\theta(\delta_1) = E_\theta\left[L(\theta, \delta_1(\boldsymbol{X}))\right] = \frac{\theta(1-\theta)}{3} \tag{1.4}$$

$$r_\theta(\delta_2) = E_\theta\left[L(\theta, \delta_2(\boldsymbol{X}))\right] = \frac{\theta(1-\theta)}{2} \tag{1.5}$$

y es claro que, para cualquier valor de θ, $r_\theta(\delta_1) < r_\theta(\delta_2)$. Por tanto, el primer procedimiento siempre sería preferible al segundo.

Podría pensarse que el objetivo debe ser la búsqueda de un procedimiento mejor que cualquier otro. Tal búsqueda sería infructuosa, como el siguiente ejemplo pone de manifiesto.

Ejemplo 1.6 En la situación descrita en el Ejemplo 1.3 (continuado en el Ejemplo 1.5) consideremos los dos siguientes procedimientos para estimar θ:

$$\delta_1(\boldsymbol{X}) = \frac{X_1 + X_2 + X_3}{3} \tag{1.6}$$

$$\delta_2(\boldsymbol{X}) = 0{,}60 \tag{1.7}$$

cuyos riesgos respectivos son:

$$r_\theta(\delta_1) = \frac{\theta(1-\theta)}{3} \tag{1.8}$$

$$r_\theta(\delta_2) = E_\theta(0{,}60 - \theta)^2 = (0{,}60 - \theta)^2 \tag{1.9}$$

Es claro que δ_2 es un procedimiento poco sensato: para nada hace uso de la información muestral. Sin embargo, cuando $\theta \simeq 0{,}6$ da excelente resultado. Siendo δ_2 un procedimiento con el que difícilmente podemos sentirnos satisfechos, es el óptimo para un cierto estado θ.

El Ejemplo 1.6 pone de manifiesto que en general no existe un procedimiento *siempre* mejor que cualquier otro[3].

[3]Naturalmente, frente al Ejemplo 1.6 nuestra reacción sería: "Si prescindimos de considerar procedimientos que sólo excepcionalmente son muy buenos, y nos limitamos a procedimientos de buen funcionamiento para cualquier θ, quizá sí haya uno mejor que todos los demás". En alguna medida, esta conjetura es cierta: si limitamos nuestra atención a clases de procedimientos y de funciones de pérdida restringidas (por ejemplo, a los procedimientos insesgados y a las funciones de pérdida convexas), puede en ocasiones encontrarse un procedimiento superior a los restantes. Estudiaremos por el momento el criterio de Bayes, para retomar esta cuestión más adelante.

Siendo cierto en general que para dos procedimientos δ_1 y δ_2 se verifica $r_\theta(\delta_1) < r_\theta(\delta_2)$ para algunos valores de θ y $r_\theta(\delta_1) > r_\theta(\delta_2)$ para otros, podríamos intentar compararlos mediante un promedio ponderado de los riesgos para diferentes valores de θ.

Supongamos que los estados de la naturaleza θ se generan de acuerdo con una cierta distribución[4], cuya función de cuantía[5] es $\xi(\theta)$. Sería razonable comparar los dos procedimientos mediante sus "riesgos promedio" respectivos:

$$R_\xi(\delta_1) = E_\xi[r_\theta(\delta_1)] = \sum_{\theta\in\Theta} \xi(\theta) r_\theta(\delta_1) \tag{1.10}$$

$$R_\xi(\delta_2) = E_\xi[r_\theta(\delta_2)] = \sum_{\theta\in\Theta} \xi(\theta) r_\theta(\delta_2) \tag{1.11}$$

Definición 1.4 *Llamamos* riesgo de Bayes *del procedimiento δ relativo a la distribución definida por $\xi(\theta)$ a*

$$R_\xi(\delta) = E_\xi[r_\theta(\delta)] = \sum_{\theta\in\Theta} \xi(\theta) r_\theta(\delta) \tag{1.12}$$

El *criterio de Bayes* para la selección de procedimientos consiste en, dada una cierta $\xi(\theta)$, tomar aquél (o aquéllos) con mínimo riesgo de Bayes. Tal (o tales) procedimientos se denominan *Bayes relativos a $\xi(\theta)$*. El criterio de Bayes resulta intuitivamente atractivo y no es objeto de controversia si hay un modo objetivo e inambiguo de especificar $\xi(\theta)$. Es objeto de controversia, en cambio, si $\xi(\theta)$ solo refleja creencias *a priori.*

Una posibilidad atractiva cuando no se tiene información *a priori* consistiría en adoptar como $\xi(\theta)$ una función de densidad que reflejara "ignorancia absoluta". Pero no está claro qué forma debería tener, como muestra el Ejemplo 1.7 a continuación.

> **Ejemplo 1.7** Supongamos que deseamos estimar, como en el Ejemplo 1.3, la proporción θ de piezas defectuosas en un lote. Una propuesta frecuente para describir "completa ignorancia" *a priori* acerca del valor de θ consiste en tomar una densidad $\xi(\theta)$ uniforme en el intervalo $\Theta = [0, 1]$. Pero esta propuesta no puede ser tomada muy en serio. Piénsese que la parametrización del problema es algo completamente arbitrario: igual que estimamos

[4] Hay diferentes formas de entender esto. Puede imaginarse que, efectivamente, hay un mecanismo que aleatoriza los estados de la naturaleza: "Dios jugando a los dados", parafraseando la célebre afirmación de Einstein. Puede pensarse tambien en esta distribución como recogiendo las creencias *a priori* del analista, que pueden reflejar experiencia acumulada o ser puramente subjetivas (tal como sucede en ocasiones en Estadística Bayesiana).

[5] En lo que resta de esta Sección y en las dos que la siguen daremos por supuesto, por comodidad notacional, que la distribución de θ es discreta con función de cuantía (o probabilidad) $\xi(\theta)$. El caso en que la distribución de θ es continua, requiere solo cambiar los sumatorios de las expresiones como (1.10)-(1.11) por integrales, y la función de cuantía por una función de densidad. (El formalismo de la integral de Stieltjes permitiría recoger en una sola expresión todos los casos.)

θ, proporción de piezas defectuosas sobre el total, podríamos desear estimar $\gamma = \frac{\theta}{1-\theta}$ (razón de piezas defectuosas a piezas correctas). Si la completa ignorancia sobre un parámetro se describe mediante una densidad *a priori* uniforme, debiéramos ahora utilizar una densidad $\xi(\gamma)$ uniforme. Pero los resultados a que llegamos son diferentes: puede comprobarse con facilidad (véase el problema 1.1, p. 20) que $\xi(\theta)$ uniforme en $\Theta = [0, 1]$ implica una densidad

$$\xi(\gamma) \quad = \quad \frac{1}{(1+\gamma)^2} \tag{1.13}$$

para $(0 \leq \gamma < \infty)$. Análogamente, una densidad uniforme[6] para γ implica una densidad no uniforme para θ. ¡Si la propuesta fuera adecuada, el no saber nada acerca de θ supondría saber algo acerca de γ, y viceversa!

Hay otras opciones de distribución *a priori* no informativa. Examinaremos una en la Observación 5.3, pág, 63.

1.3. Cómputo de procedimientos de Bayes.

De la definición de $R_\xi(\delta)$ en la Sección 1.2 se deduce que:

$$\begin{aligned} R_\xi(\delta) &= \sum_{\theta\in\Theta} \xi(\theta) r_\theta(\delta) \\ &= \sum_{\theta\in\Theta} \xi(\theta) \sum_{\boldsymbol{x}} L(\theta, \delta(\boldsymbol{x})) f_{\boldsymbol{X}\,|\theta}(\boldsymbol{x}\,|\theta) \\ &= \sum_{\boldsymbol{x}} \underbrace{\left[\sum_{\theta\in\Theta} L(\theta, \delta(\boldsymbol{x}))\xi(\theta) f_{\boldsymbol{X}\,|\theta}(\boldsymbol{x}\,|\theta)\right]}_{\stackrel{\text{def}}{=} h_\xi(\boldsymbol{x}, \delta(\boldsymbol{x}))} \end{aligned} \tag{1.14}$$

Para minimizar el riesgo, tenemos que minimizar $h_\xi(\boldsymbol{x}, \delta(\boldsymbol{x}))$ en (1.14) para cada $\boldsymbol{x}$. Pueden ocurrir dos cosas:

- Que para cada $\boldsymbol{x}$ haya *una única* decisión $d = \delta(\boldsymbol{x})$ en D minimizando $h_\xi(\boldsymbol{x}, d))$. En este caso, hay un único procedimiento de Bayes relativo a $\xi(\theta)$.
- Que haya más de una decisión minimizando $h_\xi(\boldsymbol{x}, d)$ para algún $\boldsymbol{x}$. En este caso, hay más de un procedimiento de Bayes relativo a $\xi(\theta)$.

En todos los casos, si definimos

$$H_\xi(\boldsymbol{x}) = \min_{d\in D} h_\xi(\boldsymbol{x}, d), \tag{1.15}$$

[6] Obsérvese que no procede hablar de una densidad uniforme sobre un intervalo de longitud infinita, como es el dominio de variación de γ. El problema se soluciona escribiendo $\xi(\gamma) \propto k$ y sustituyendo los signos $=$ por signos $\propto$. Se dice que se está ante una distribución *a priori difusa*. Se suele también denominar a $\xi(\gamma)$ densidad *a priori impropia*.

el riesgo de Bayes es $R_\xi(\delta) = \sum_{\boldsymbol{x}} H_\xi(\boldsymbol{x})$. El Ejemplo 1.8, aunque artificialmente simple, ilustra algunos de los conceptos introducidos.

Ejemplo 1.8 Supongamos que, dependiendo quizá de la climatología, un paraje puede adoptar uno de dos estados, θ_1 y θ_2. En el estado θ_1, el paraje produce sólo setas comestibles, mientras que en el estado θ_2 produce sólo setas tóxicas, indistinguibles a los ojos de un profano de las primeras. Las probabilidades respectivas de ambos estados son $\xi(\theta_1) = 0{,}90$ y $\xi(\theta_2) = 0{,}10$.

Para adquirir mayor información sobre el carácter de una seta recogida, podemos preguntar a un experto, que sin embargo no es infalible. En cada uno de los dos estados proporciona una respuesta X cuyos posibles valores son $X = C$ (declara la seta comestible) ó $X = T$ (declara la seta tóxica). La distribución de X para cada uno de los dos posibles estados aparece en la Tabla 1.1.

Cuadro 1.1: Función de cuantía $f_{X|\theta}(x\,|\theta)$

Respuesta X experto	$\theta_\mathbf{1}$ **(seta comestible)**	$\theta_\mathbf{2}$ **(seta tóxica)**
$X = C$	0.950	0.005
$X = T$	0.050	0.995

Hay dos posibles decisiones: d_1 = “Tirar la seta”, y d_2 = “Comer la seta”. Suponemos que las pérdidas asociadas a cada decisión en cada uno de los estados posibles son las recogidas en la Tabla 1.2.

Cuadro 1.2: Función de pérdida $L(\theta_i, d_j)$

Decisión adoptada	$\theta_\mathbf{1}$ **(seta comestible)**	$\theta_\mathbf{2}$ **(seta tóxica)**
d_1 (tirar)	100	0
d_2 (comer)	-10	1000

Consideramos tres posibles procedimientos estadísticos, que consisten en preguntar al experto y, obtenido un valor de X, actuar del modo que se indica en la Tabla 1.3.

Con la información anterior, es fácil calcular los riesgos respectivos de los tres procedimientos considerados:

$$\begin{aligned}
r_{\theta_1}(\delta_1) &= L(\theta_1, d_1)\text{Prob}\{\delta_1(X) = d_1|\theta_1\} + L(\theta_1, d_2)\text{Prob}\{\delta_1(X) = d_2|\theta_1\} \\
&= 100 \times 0 + (-10) \times 1 = -10 \\
r_{\theta_2}(\delta_1) &= L(\theta_2, d_1)\text{Prob}\{\delta_1(X) = d_1|\theta_2\} + L(\theta_2, d_2)\text{Prob}\{\delta_1(X) = d_2|\theta_2\} \\
&= 0 \times 0 + 1000 \times 1 = 1000 \\
r_{\theta_1}(\delta_2) &= L(\theta_1, d_1)\text{Prob}\{\delta_2(X) = d_1|\theta_1\} + L(\theta_1, d_2)\text{Prob}\{\delta_2(X) = d_2|\theta_1\} \\
&= 100 \times 0{,}05 + (-10) \times 0{,}95 = -4{,}5 \\
r_{\theta_2}(\delta_2) &= L(\theta_2, d_1)\text{Prob}\{\delta_2(X) = d_1|\theta_2\} + L(\theta_2, d_2)\text{Prob}\{\delta_2(X) = d_2|\theta_2\} \\
&= 0 \times 0{,}995 + 1000 \times 0{,}005 = 5 \\
r_{\theta_1}(\delta_3) &= L(\theta_1, d_1)\text{Prob}\{\delta_3(X) = d_1|\theta_1\} + L(\theta_1, d_2)\text{Prob}\{\delta_3(X) = d_2|\theta_1\} \\
&= 100 \times 1 + (-10) \times 0 = 100 \\
r_{\theta_2}(\delta_3) &= L(\theta_2, d_1)\text{Prob}\{\delta_3(X) = d_1|\theta_2\} + L(\theta_2, d_2)\text{Prob}\{\delta_3(X) = d_2|\theta_2\} \\
&= 0 \times 1 + 1000 \times 0 = 0
\end{aligned}$$

Cuadro 1.3: Procedimientos $\delta_i(X)$ considerados

Procedimiento	**Descripción**
$\delta_1(X)$	Sea cual fuere X, comer la seta (d_2).
$\delta_2(X)$	Si $X = C$, comer la seta (d_2). En caso contrario, tirar la seta.
$\delta_3(X)$	Sea cual fuere X, tirar la seta (d_1).

La Tabla 1.4 recoge los riesgos calculados. Puede observarse que ningún procedimiento es mejor a ninguno de los restantes.

Los respectivos riesgos de Bayes relativos a la distribución *a priori* especificada por $\xi(\theta)$ se calculan también fácilmente:

$$\begin{aligned}
R_\xi(\delta_1) &= r_{\theta_1}(\delta_1)\xi(\theta_1) + r_{\theta_2}(\delta_1)\xi(\theta_2) = 0{,}90 \times (-10) + 0{,}10 \times 1000 = 91 \\
R_\xi(\delta_2) &= r_{\theta_1}(\delta_2)\xi(\theta_1) + r_{\theta_2}(\delta_2)\xi(\theta_2) = 0{,}90 \times (-4{,}5) + 0{,}10 \times 5 = -3{,}55 \\
R_\xi(\delta_3) &= r_{\theta_1}(\delta_3)\xi(\theta_1) + r_{\theta_2}(\delta_3)\xi(\theta_2) = 0{,}90 \times 100 + 0{,}10 \times 0 = 90
\end{aligned}$$

El criterio de Bayes llevaría en este caso a seleccionar $\delta_2(X)$. El procedimiento seleccionado depende de la distribución *a priori* considerada. Si en lugar de la indicada hubiéramos tenido: $\xi(\theta_1) = 0{,}001$, $\xi(\theta_2) = 0{,}999$ (es decir, casi seguridad de que la seta procede de un paraje que sólo produce tóxicas), es fácil comprobar que el procedimiento escogido por el criterio de

Cuadro 1.4: Funciones de riesgo $r_{\theta_i}(\delta_j)$

Procedimiento $\delta_{\mathbf{j}}(\mathbf{X})$	$\theta_{\mathbf{1}}$ **(seta comestible)**	$\theta_{\mathbf{2}}$ **(seta tóxica)**
$\delta_1(X)$	-10	1000
$\delta_2(X)$	-4.5	5
$\delta_3(X)$	100	0

Bayes sería $\delta_3(X)$ (tirar la seta, incluso aunque el dictamen del experto sea que es comestible). Sucede que nuestras creeencias *a priori* son tan fuertes, que no basta la evidencia aportada por el experimento para hacernos cambiar de opinión.

De la expresión (1.14) dedujimos que el procedimiento óptimo de acuerdo con el criterio de Bayes minimiza

$$h_\xi(\boldsymbol{x}, \delta(\boldsymbol{x})) = \sum_{\theta\in\Theta} L(\theta, \delta(\boldsymbol{x}))\xi(\theta) f_{\boldsymbol{X}\,|\theta}(\boldsymbol{x}\,|\theta) \tag{1.16}$$

para cada valor de $\boldsymbol{x}$. Como

$$\xi(\theta) f_{\boldsymbol{X}\,|\theta}(\boldsymbol{x}\,|\theta) = f_{\boldsymbol{X}}(\boldsymbol{x}, \theta) = f_{\theta|\boldsymbol{X}}(\theta\,|\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}), \tag{1.17}$$

tenemos que el procedimiento (o los procedimientos) Bayes relativos a la distribución *a priori* $\xi(\theta)$ minimizan

$$h_\xi(\boldsymbol{x}, \delta(\boldsymbol{x})) = f_{\boldsymbol{X}}(\boldsymbol{x}) \sum_{\theta\in\Theta} L(\theta, \delta(\boldsymbol{x})) f_{\theta|\boldsymbol{X}}(\theta\,|\boldsymbol{x})$$

para cada $\boldsymbol{x}$ y, por tanto, también para cada $\boldsymbol{x}$, minimizan

$$\sum_{\theta\in\Theta} L(\theta, \delta(\boldsymbol{x})) f_{\theta|\boldsymbol{X}}(\theta\,|\boldsymbol{x}). \tag{1.18}$$

En ausencia de experimento, escogeríamos un procedimiento δ que minimizara el riesgo de Bayes *a priori* , es decir:

$$\sum_{\theta\in\Theta} L(\theta, \delta)\xi(\theta). \tag{1.19}$$

La comparación de las expresiones (1.18) y (1.19) muestra que el método de elección de un procedimiento es siempre el mismo, con la sola variación de que en un

caso se emplea la distribución *a priori* sobre los estados de la naturaleza y en otro la distribución *a posteriori* conocido el resultado del experimento. Este resultado sólo influye alterando la distribución con respecto a la cual se calcula la pérdida media, que de ser $\xi(\theta)$ pasa a ser $f_{\theta|\boldsymbol{X}}(\theta\,|\boldsymbol{x})$. En el enfoque de la inferencia aportado por la Teoría de la Decisión, la información muestral interviene modificando la distribución *a priori* del analista y transformándola en una distribución *a posteriori*; la forma de operar con cada una de ambas distribuciones para seleccionar un procedimiento estadístico es sin embargo siempre la misma.

1.4. Procedimientos de Bayes con función de pérdida cuadrática.

Cuando la función de pérdida es cuadrática o, de modo un poco más general, de la forma

$$L(\theta, d) = w(\theta)\,[d-\theta]^2$$

siendo $w(\theta)$ una función no negativa cualquiera, entonces el procedimiento de Bayes relativo a una cierta distribución *a priori* $\xi(\theta)$ es particularmente fácil de obtener, como muestra el siguiente teorema.

Teorema 1.1 *Sea $L(\theta,d) = w(\theta)\,[d-\theta]^2$ y $w(\theta)$ una función no negativa. El procedimiento de Bayes relativo a $\xi(\theta)$ es:*

$$\delta_\xi(\boldsymbol{x}) \;=\; \frac{\sum_\theta w(\theta)\theta f_{\theta|\boldsymbol{X}}(\theta\,|\boldsymbol{x})}{\sum_\theta w(\theta) f_{\theta|\boldsymbol{X}}(\theta\,|\boldsymbol{x})} = \frac{E_{\theta|\boldsymbol{x}}\,[w(\theta)\theta]}{E_{\theta|\boldsymbol{x}}\,[w(\theta)]}. \tag{1.20}$$

DEMOSTRACION:

Para cada $\boldsymbol{x}$, $\delta(\boldsymbol{x})$ ha de ser, de acuerdo con (1.18), tal que minimice:

$$\sum_\theta w(\theta)\,[\delta(\boldsymbol{x})-\theta]^2\, f_{\theta|\boldsymbol{X}}(\theta\,|\boldsymbol{x}). \tag{1.21}$$

Minimizando la expresión anterior respecto a $\delta(\boldsymbol{x})$ se llega inmediatamente a (1.20).

■

1.5. Familias conjugadas

El cómputo de procedimientos de Bayes se simplifica si $f_{\theta|\boldsymbol{X}}(\theta\,|\boldsymbol{x})$ puede obtenerse con facilidad. De (1.17) se deduce que:

$$f_{\theta|\boldsymbol{X}}(\theta\,|\boldsymbol{x}) \propto \xi(\theta) f_{\boldsymbol{X}\,|\theta}(\boldsymbol{x}\,|\theta) \tag{1.22}$$

En ocasiones, $\xi(\theta)$ y $f_{\boldsymbol{X}\,|\theta}(\boldsymbol{x}\,|\theta)$ son tales que $f_{\theta|\boldsymbol{X}}(\theta\,|\boldsymbol{x})$ pertenece a la misma familia que $\xi(\theta)$; se dice entonces que $\xi(\theta)$ y $f_{\boldsymbol{X}\,|\theta}(\boldsymbol{x}\,|\theta)$ pertenecen a *familias conjugadas*. El siguiente ejemplo muestra las ventajas que se derivan de ello.

Ejemplo 1.9 Tenemos una única observación X procedente de una binomial $b(\theta, n)$, cuyo parámetro θ se trata de estimar con pérdida cuadrática $L(\theta, \delta(X)) = (\delta(X) - \theta)^2$.

Si la distribución *a priori* de θ fuera una beta de parámetros r y s, es decir, si:

$$\xi(\theta) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1}$$

con $0 < \theta < 1$, tendríamos, de acuerdo con (1.22), que:

$$\begin{aligned} f_{\theta|X}(\theta\,|x) &\propto \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1}\binom{n}{x}\theta^x(1-\theta)^{n-x} \quad (1.23)\\ &\propto \theta^{r+x-1}(1-\theta)^{n+s-x-1} \quad (1.24)\end{aligned}$$

Se reconoce con facilidad en (1.24) una densidad beta de parámetros $(r+x)$ y $(n+s-x)$, falta sólo de la correspondiente constante de normalización: $f_{\theta|X}(\theta\,|x)$ por tanto pertenece a la misma familia que la $\xi(\theta)$ escogida.

De acuerdo con (1.20), $\delta(X)$ será el valor medio condicionado de la distribución *a posteriori* de θ. Tratándose de una beta, se tiene (ver por ej. Trocóniz (1987), p. 299):

$$\delta(X) = m = \frac{r+X}{n+s-X+r+X} = \frac{r+X}{n+r+s}$$

que puede reescribirse así:

$$\delta(X) = \left(\frac{n}{n+r+s}\right)\cdot\frac{X}{n} + \frac{r}{n+r+s} \quad (1.25)$$

Cuando $n \to \infty$, $\delta(X) \to X/n$ (número de "aciertos" entre n), como cabría esperar. Sin embargo, para n moderado la distribución *a priori* $\xi(\theta)$ es de gran importancia.

El emplear una distribución beta como $\xi(\theta)$ tiene la ventaja de producir una distribución *a posteriori* inmediatamente reconocible, y de la que podemos obtener el valor medio con facilidad. Si $\xi(\theta)$ hubiera sido otra, hubiera sido en general precisa una operación de integración, y el resultado no hubiera podido obtenerse de forma tan simple.

Ejemplo 1.10 (continuación) Para uso posterior nos interesará disponer de la función de riesgo del estimador obtenido en el ejemplo anterior.

$$\begin{aligned} r_\theta(\delta) &= E\left[(\delta(X)-\theta)^2|\theta\right]\\ &= \mathrm{Var}_\theta(\delta(X)) + \left[\mathrm{Sesgo}_\theta(\delta(X))\right]^2\\ &= \left(\frac{n}{n+r+s}\right)^2\frac{\theta(1-\theta)}{n} + \left(\frac{r+n\theta}{n+r+s}-\theta\right)^2\end{aligned}$$

Ejemplo 1.11 Supongamos que la distribución de X es $N(\theta, \sigma^2)$, y la distribución *a priori* sobre θ es $N(\mu, b^2)$. Tenemos entonces que:

$$f_{\boldsymbol{X}\,|\theta}(\boldsymbol{x}\,|\theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^n\left(\frac{x_i-\theta}{\sigma}\right)^2\right\} \quad (1.26)$$

mientras que por otra parte, la densidad $\xi(\theta)$ es:

$$\xi(\theta) = \frac{1}{b\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\theta-\mu}{b}\right)^2\right\} \tag{1.27}$$

Por consiguiente:

$$\begin{aligned} f_{\theta|\boldsymbol{X}}(\theta|\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) &= \frac{1}{b\sqrt{2\pi}} \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^n \left(\frac{x_i-\theta}{\sigma}\right)^2 + \left(\frac{\theta-\mu}{b}\right)^2\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{\theta^2(\sigma^2+nb^2) - 2\theta(\sigma^2\mu + nb^2\overline{x}) + (\sigma^2\mu^2 + b^2\sum x_i^2)}{\sigma^2 b^2}\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{\theta - \frac{\mu\sigma^2+nb^2\overline{x}}{\sigma^2+nb^2}}{\sqrt{\frac{b^2\sigma^2}{\sigma^2+nb^2}}}\right)^2\right\}, \end{aligned} \tag{1.28}$$

esta última expresión obtenida al completar el cuadrado de la precedente. Es fácil reconocer en ella una densidad normal para θ:

$$(\theta|\boldsymbol{X}=\boldsymbol{x}) \sim N\left(\frac{\mu\sigma^2 + nb^2\overline{x}}{\sigma^2+nb^2},\ \frac{b^2\sigma^2}{\sigma^2+nb^2}\right)$$

Observación 1.1 Con una muestra de n observaciones $X_i \sim N(\theta, \sigma^2)$, el estimador *ridge* de parámetro k de θ vendría dado por:

$$\hat{\theta} = \frac{n\overline{x}}{n+k};$$

podemos ver que dicha expresión es idéntica a

$$\frac{\mu\sigma^2 + nb^2\overline{x}}{\sigma^2+nb^2} \tag{1.29}$$

cuando hacemos $\mu = 0$ y $b^2 = \sigma^2/k$. Por tanto, el uso del estimador ridge de parámetro k en este caso equivale a la utilización implícita de una distribución *a priori* $N(0, \sigma^2/k)$. Valores de k muy pequeños en relación a σ^2 implican gran incertidumbre acerca de θ (y una estimación muy próxima a la obtenida por máxima verosimilitud o mínimos cuadrados ordinarios). Valores relativamente grandes de k (siempre en relación a σ^2) suponen gran convicción de que θ está en las cercanías de $\mu = 0$.

Hay otros muchos casos en que el empleo de una distribución *a priori* conveniente simplifica la obtención de la distribución *a posteriori*. La siguiente tabla muestra algunos de los más frecuentes.

La comodidad de manejo de las familias conjugadas no debe hacernos perder de vista, sin embargo, algo fundamental: que el fundamento de la utilización de una distribución *a priori* se pierde si ésta no describe bien el mecanismo que genera los estados de la naturaleza —o nuestras creencias acerca del particular, si adoptamos una visión bayesiana—.

Cuadro 1.5: Algunas distribuciones *a priori* conjugadas

Distribución de X	**Parámetro de interés**	***A priori* conjugada**
Binomial, $b(\theta, n)$	θ	$Beta(r, s)$
Poisson, $P(\theta)$	θ	$\gamma(a, b)$
Exponencial, $f_X(x) = \theta e^{-\theta x}$	θ	$\gamma(a, b)$
Normal, $N(\theta, \sigma_0^2)$	θ	Normal, $N(\mu, \tau^2)$

1.6. Procedimientos aleatorizados.

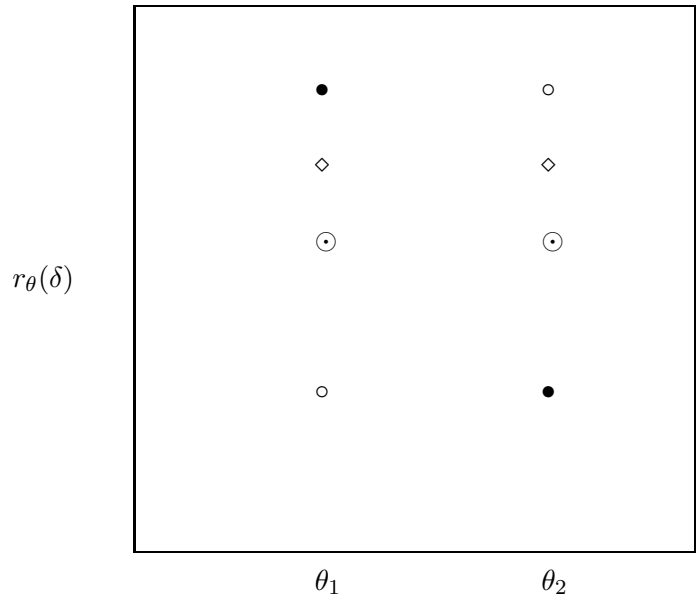
Se ha definido (Sección 1.1) procedimiento estadístico como una aplicación $\delta\colon S \longrightarrow D$. Ampliaremos ahora esta definición denominando *procedimiento estadístico aleatorizado* a una aplicación $\delta\colon S \longrightarrow \Pi(D)$, en que $\Pi(D)$ es el conjunto de distribuciones sobre D. En otras palabras, un procedimiento estadístico aleatorizado hace corresponder a cada resultado muestral una "lotería" en la que se puede obtener una de varias decisiones. De este modo, el mismo resultado X llevaría en ocasiones diferentes a tomar decisiones posiblemente diferentes.

Esto es algo difícilmente asumible: ¿por qué habríamos de hacer depender nuestra decisión de una lotería? Dada la distribución *a priori* , y realizado el experimento, parece que no debiéramos recurrir a aleatorizar nuestra decisión. Hay dos formas de responder a esto. Una, que, como hace notar Kiefer (1983), tal forma de actuar no debiera ser motivo de escándalo. Al fin y al cabo, cuando se hace casi cualquier tipo de experimento se aleatoriza el diseño: la evidencia muestral depende así de una especie de "lotería" previa —la que nos ha llevado a escoger un diseño experimental en particular y no otro—. La segunda, y más importante para lo que sigue, es que la consideración de procedimientos aleatorizados permite obtener resultados interesantes, en particular completando la clase de los procedimientos de Bayes de modo que incluya algunos de interés. La Sección 1.8 aclarará esta cuestión.

Ejemplo 1.12 Tomemos el caso simple en que hay dos posibles estados de la naturaleza, θ_1 y θ_2. Consideraremos también tres procedimientos δ_1, δ_2 y δ_3, cuyas funciones de riesgo se representan gráficamente en la Figura 1.2

Puede comprobarse que ni δ_1 ni δ_2 (cuyos riesgos están representados en la figura por $\bullet$ y $\circ$ respectivamente) son mejores que δ_3; cada uno de ellos tiene menor riesgo en uno de los estados y mayor en el otro. Sin embargo, si

Figura 1.2: $\delta_4 = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2$ ($\odot$) es mejor que δ_3 ($\diamond$)



adoptamos la regla de aleatorizar entre δ_1 y δ_2 arrojando una moneda regular al aire, obtenemos un nuevo procedimiento (aleatorizado) δ_4, representado en la figura mediante $\odot$, que sí es mejor que δ_3. Su función de riesgo es $r_\theta(\delta_4) = \frac{1}{2}r_\theta(\delta_1) + \frac{1}{2}r_\theta(\delta_2)$.

1.7. Clases completas.

La siguiente definición introduce un concepto que necesitamos en lo que sigue.

Definición 1.5 *La clase C de procedimientos es* completa *si para cada procedimiento que no esté en C hay uno en C que es mejor. Si C es la clase más restringida de procedimientos que es completa, se dice que es* mínima completa.

Esta definición podría parafrasearse diciendo que una clase completa contiene la totalidad de procedimientos admisibles. Tenemos por otra parte la noción de clase *esencialmente completa*:

Definición 1.6 *La clase C de procedimientos es* esencialmente completa *si para cada procedimiento que no esté en C hay uno en C que es mejor o igual. Si C es la clase más restringida de procedimientos que es esencialmente completa, se dice que es* esencialmente mínima completa.

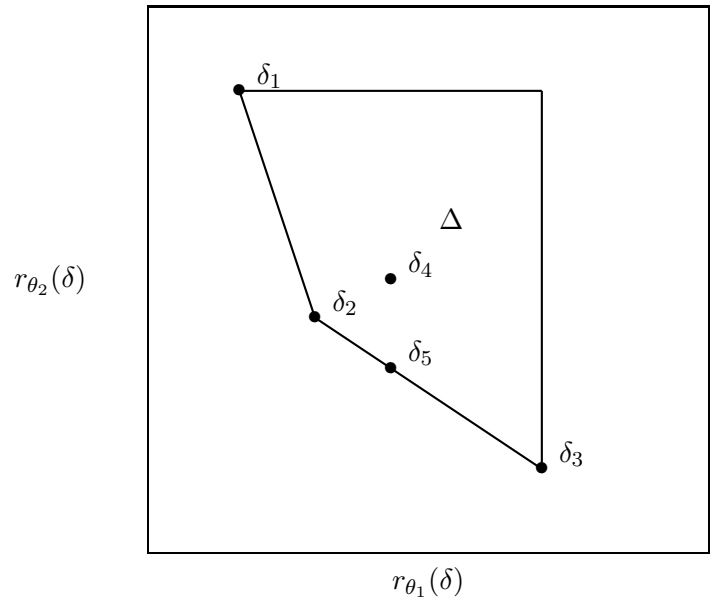
Bajo condiciones muy generales, de habitual cumplimiento en la práctica, la única clase mínima completa coincide con la clase de todos los procedimientos

admisibles. Una clase esencialmente mínima completa contiene *un* representante de cada grupo de procedimientos admisibles equivalentes (ver Kiefer (1983), p. 54).

1.8. Representación gráfica de procedimientos estadísticos.

Hemos representado gráficamente funciones de riesgo. Construiremos ahora gráficos en que cada punto representa un procedimiento, y cada eje un estado de la naturaleza. Por simplicidad, consideraremos sólo el caso en que $\Theta = \{\theta_1, \theta_2\}$. En la Figura 1.3, el procedimiento δ_1 tiene riesgos $r_{\theta_1}(\delta_1) = 1$, y $r_{\theta_2}(\delta_1) = 6$. Análogamente, δ_2 tiene riesgos $r_{\theta_1}(\delta_2) = 2$, y $r_{\theta_2}(\delta_2) = 3$. Obsérvese que un procedimiento δ_4 que consistiera en aleatorizar entre δ_1 y δ_3 con probabilidades respectivas π_1 y π_2 tendría función de riesgo $r_\theta(\delta_4) = \pi_1 r_\theta(\delta_1) + \pi_2 r_\theta(\delta_3)$, combinación lineal convexa de las de δ_1 y δ_3, y podríamos representarlo como un punto del segmento que une los puntos correspondientes a δ_1 y δ_3.

Figura 1.3: El contorno rayado en grueso incluye los procedimientos en la clase completa minimal. δ_4 es inadmisible (resulta mejorado, por ejemplo, por el procedimiento aleatorizado δ_5, cuyo riesgo es el mismo cuando $\theta = \theta_1$ e inferior cuando $\theta = \theta_2$)



Si consideramos procedimientos aleatorizados, *toda combinación lineal convexa de procedimientos puede verse como otro posible procedimiento*. Ello hace ver que el conjunto de posibles procedimientos es, cuando lo representamos como en la Figura 1.3, un conjunto convexo.

Por otra parte, el riesgo de Bayes de un procedimiento δ_i cuando hay dos únicos estados viene dado por:

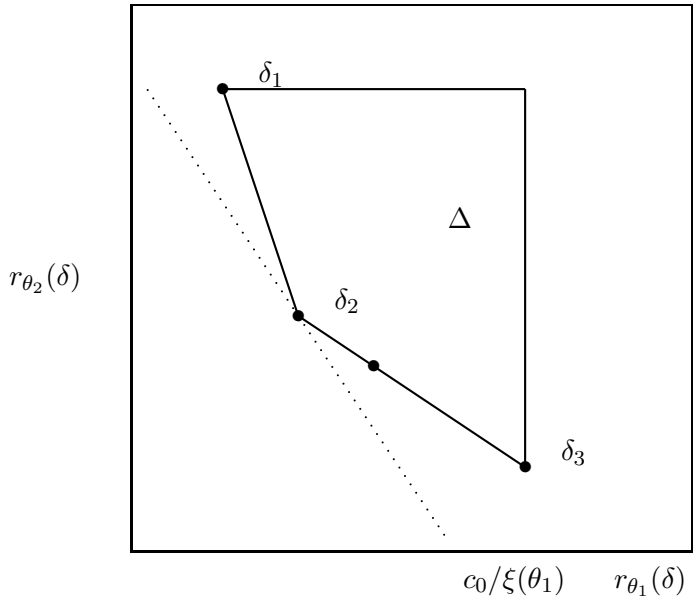
$$R_\xi(\delta_i) = \xi(\theta_1) r_{\theta_1}(\delta_i) + \xi(\theta_2) r_{\theta_2}(\delta_i)$$

y por lo tanto el lugar geométrico de los procedimientos con igual riesgo de Bayes c es la recta

$$\xi(\theta_1) r_{\theta_1}(\delta_i) + \xi(\theta_2) r_{\theta_2}(\delta_i) = c \tag{1.30}$$

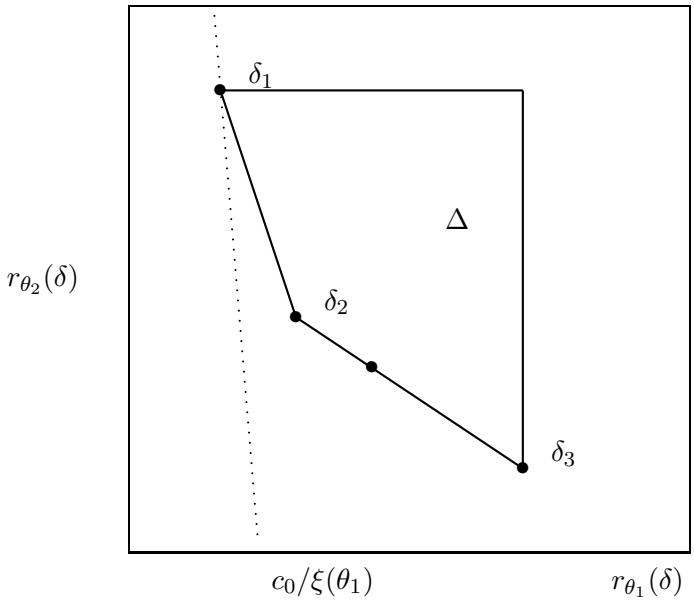
La Figura 1.4 muestra un conjunto de procedimientos Δ cuyo borde inferior es la clase minimal completa. Para diferentes valores de c, la ecuación (1.30) proporciona diferentes rectas paralelas, cuya pendiente depende de ξ, y tanto más cercanas al origen cuanto menor sea c. El procedimiento de Bayes relativo a $\xi(\theta)$ en el caso representado en dicha figura sería δ_2. Para cualquier c menor que c_0, la recta correspondiente no intersectaría Δ.

Figura 1.4: El procedimiento de Bayes relativo a $\xi(\theta)$ es δ_2, y el riesgo de Bayes c_0



Es fácil ver de modo intuitivo que para una diferente distribución *a priori* el procedimiento de Bayes sería diferente (como ilustra la Figura 1.5, en que el procedimiento de Bayes es δ_1).También es fácil ver que puede no haber un único procedimiento de Bayes; si la distribución *a priori* fuera tal que las rectas de riesgo Bayes constante tuvieran exactamente la misma pendiente que uno de los segmentos (δ_1, δ_2) ó (δ_2, δ_3), el contacto entre la recta de mínimo riesgo y el conjunto de procedimientos Δ se produciría en más de un punto.

Figura 1.5: El procedimiento de Bayes relativo a $\xi(\theta)$ es δ_1, y el riesgo de Bayes c_0



Finalmente, es de interés señalar que, mientras que el contorno dibujado en grueso representa la clase mínima completa, la formada por los procedimientos $\{\delta_1, \delta_2, \delta_3\}$ es esencialmente mínima completa.

1.9. Límites de sucesiones de procedimientos de Bayes

En ocasiones, un procedimiento no es de Bayes, pero es límite de una sucesión de procedimientos de Bayes. El siguiente ejemplo muestra esto con claridad.

Ejemplo 1.13 Consideremos el caso en que hemos de estimar con función de pérdida cuadrática el parámetro media de una población $N(\theta, \sigma^2)$, y la distribución *a priori* sobre θ es $\theta \sim N(\mu, b^2)$. En tal caso, hemos visto (Ejemplo 1.11) que la distribución *a posteriori* de θ es:

$$(\theta|\boldsymbol{X}) \sim N\left(\frac{\mu\sigma^2 + nb^2\overline{X}}{\sigma^2 + nb^2},\ \frac{b^2\sigma^2}{\sigma^2 + nb^2}\right)$$

y por consiguiente, de acuerdo con el Teorema 1.1:

$$\begin{aligned}\delta(\boldsymbol{X}) &= E[\theta|\boldsymbol{X} = \boldsymbol{x}] = \int \theta f_{\theta|\boldsymbol{X}}(\theta|\boldsymbol{x})d\theta \\ &= \frac{\overline{X}b^2 + \mu\sigma^2/n}{b^2 + \sigma^2/n} \\ &= \frac{\sigma^2/n}{b^2 + \sigma^2/n}\mu + \frac{b^2}{b^2 + \sigma^2/n}\overline{X}\end{aligned}$$

Cuando $n \to \infty$, $\delta(\boldsymbol{X}) \to \overline{X}$; la distribución *a priori* es reducida a la irrelevancia por el peso abrumador de la evidencia muestral. Se dice que $\overline{X}$ es límite de procedimientos de Bayes.

1.10. Interés de los procedimientos de Bayes.

Hay buen número de razones para interesarse por los procedimientos de Bayes. Idealmente, desearíamos restringir nuestra atención a los procedimientos admisibles —aquellos que no pueden ser mejorados por ningún otro—, o, aún mejor, a una subclase esencialmente completa y mínima de procedimientos admisibles. La clase de los procedimientos de Bayes y de sus límites es, en general, algo más amplia. Si D y Θ son finitos, la clase de procedimientos de Bayes es completa. Si Θ no es finito, se puede en general obtener una clase completa incluyendo también los procedimientos que son límite de procedimientos de Bayes. La clase de procedimientos de Bayes, quizás completada, es por ello un buen punto de partida.

Por otra parte, los procedimientos de Bayes pueden justificarse desde varios puntos de vista, desde el totalmente bayesiano hasta aquél que utiliza como distribución *a priori* una distribución derivada de la experiencia anterior.

Por último, podemos relajar de diversas maneras el requerimiento de que $\xi(\theta)$ (y $L(\theta, d)$) sean conocidas, y tratar de encontrar procedimientos que sean ventajosos en condiciones muy generales, o que sean de mínimo riesgo en las circunstancias más desfavorables. Esta última alternativa da lugar a los procedimientos minimax y se explora junto con la caracterización de procedimientos admisibles en el Capítulo 2.

CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

1.1 Compruébese que, como se dice en el Ejemplo 1.7, si $\xi(\theta)$ es uniforme en $\Theta = [0, 1]$ la densidad de $\gamma = \theta/(1-\theta)$ es $\xi(\gamma) = (1+\gamma)^{-2}$.

1.2 Haciendo uso del hecho de que,

$$\frac{\partial}{\partial y}\int_{a(y)}^{b(y)} g(x,y)dx \quad = \quad \frac{\partial b}{\partial y}g(b,y) - \frac{\partial a}{\partial y}g(a,y) + \int_{a(y)}^{b(y)} \frac{\partial g(x,y)}{\partial y}dx$$

demuéstrese que el estimador $\hat{\theta}$ que minimiza la función de pérdida $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$ es la mediana de la distribución $f_{\theta|X}(\theta|x)$ (supuesta ésta última continua, y por tanto la mediana únicamente definida).

(Garthwaite et al. (1995), pág. 118)

Capítulo 2

Procedimientos admisibles y minimax.

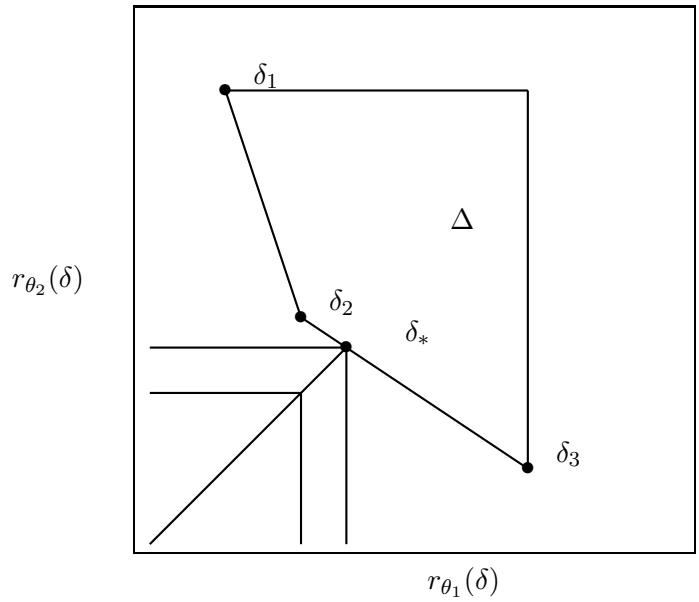
2.1. Minimax y criterios globales.

El criterio de Bayes se justificaba en el Capítulo anterior como un promedio ponderado del riesgo, con ponderación dada por $\xi(\theta)$. Ello presta cierto atractivo a dicho criterio: si un agente se enfrenta al mismo proceso de decisión muchas veces, el minimizar el riesgo medio es una estrategia sensata.

Puede suceder que, o bien desconozcamos $\xi(\theta)$, o bien enfrentemos un proceso de decisión una única vez. En estas circunstancias y algunas otras, puede interesarnos minimizar el mayor de los riesgos que hayamos de afrontar. En otras palabras, podemos diseñar una estrategia consistente en hacer mínimo el riesgo en la situación (es decir, para el θ) más desfavorable. Se trata de una estrategia conservadora, que procura la máxima cobertura frente a la peor catástrofe. La comparación entre procedimientos se hace así sobre la base de *un sólo* valor (el máximo) de las correspondientes funciones de riesgo, en lugar de considerar (promediándolos mediante $\xi(\theta)$) la totalidad de los riesgos.

El empleo de gráficos como los introducidos en la Sección 1.8 es ilustrativo. La Figura 2.1 muestra un procedimiento δ_2 que no es minimax y uno que sí lo es, δ_*. Es muy intuitivo el procedimiento gráfico que debemos seguir para encontrar procedimientos minimax; construiremos cuadrados cuyo vértice inferior izquierdo se apoye sobre el origen, y cuyo vértice superior derecho se apoye sobre la bisectriz del primer cuadrante. En la Figura 2.1 puede verse uno de dichos cuadrados, de lado 2, que no alcanza a intersectar Δ y otro —de lado 2.6— que sí lo hace. El punto de contacto, $(2{,}6, 2{,}6)$, corresponde al procedimiento δ_* minimax. No hay

Figura 2.1: δ_* es minimax. δ_2 no lo es; su riesgo cuando $\theta = \theta_2$ es mayor que el de δ_*.



ningún procedimiento factible con riesgos menores tanto para θ_1 como para θ_2 (un tal procedimiento estaría en *el interior* del cuadrado de lados $(2{,}6, 2{,}6)$ dibujado). Más precisamente, tenemos la siguiente

Definición 2.1 *Se dice que δ_* es un* procedimiento minimax *en una cierta clase de procedimientos Δ si $\forall \delta \in \Delta$:*

$$\sup_\theta r_\theta(\delta_*) \leq \sup_\theta r_\theta(\delta) \tag{2.1}$$

2.2. Caracterización de procedimientos minimax.

Los procedimientos minimax no tienen porqué ser únicos. Tampoco tienen necesariamente que ser admisibles (como la Figura 2.2 pone de manifiesto). El siguiente teorema proporciona una caracterización útil de procedimientos minimax y una condición suficiente para que sean admisibles.

Teorema 2.1 *Si δ_ξ es un procedimiento de Bayes respecto a $\xi(\theta)$, distribución tal que:*

$$\sum_\theta r_\theta(\delta_\xi)\xi(\theta) = \sup_\theta r_\theta(\delta_\xi) \tag{2.2}$$

entonces: (i) δ_ξ es minimax. (ii) Si δ_ξ es la única solución de Bayes con respecto a $\xi(\theta)$, es el único procedimiento minimax.

DEMOSTRACION:

Tomemos cualquier otro procedimiento δ. Entonces,

$$\sup_{\theta\in\Theta} r_\theta(\delta) \geq \sum_{\theta\in\Theta} r_\theta(\delta)\xi(\theta) \geq \sum_{\theta\in\Theta} r_\theta(\delta_\xi)\xi(\theta) = \sup_{\theta\in\Theta} r_\theta(\delta_\xi) \tag{2.3}$$

El apartado (ii) se deduce inmediatamente, si tenemos en cuenta que la unicidad de δ_ξ implica que la segunda desigualdad en (2.3) es estricta.

La distribución definida por $\xi(\theta)$ se denomina *distribución* a priori *más desfavorable*. Da lugar al máximo riesgo de Bayes. En efecto, supongamos cualquier otra distribución *a priori* $\tau(\theta)$, y un procedimiento δ_τ que sea de Bayes respecto a la misma. Entonces:

$$R_\tau(\delta_\tau) = \sum_{\theta\in\Theta} r_\theta(\delta_\tau)\tau(\theta) \leq \sum_{\theta\in\Theta} r_\theta(\delta_\xi)\tau(\theta) \leq \sup_{\theta\in\Theta} r_\theta(\delta_\xi) = R_\xi(\delta_\xi) \tag{2.4}$$

■

Dos consecuencias son inmediatas:

Corolario 2.1 *Un procedimiento de Bayes de riesgo constante es minimax.*

En efecto, basta comprobar que en este caso (2.2) se verifica.

Corolario 2.2 *Sea* $\Theta_\xi = \{\theta' \colon r_{\theta'}(\delta_\xi) = \sup_\theta r_\theta(\delta_\xi)\}$*, es decir, el conjunto de estados para los que el riesgo de* δ_ξ *toma su valor máximo. Entonces,* δ_ξ *es minimax si* Θ_ξ *tiene, de acuerdo con la distribución definida por* $\xi(\theta)$*, probabilidad uno.*

Esto se deduce, como el corolario anterior, de (2.2). Si, excepto para sumandos con probabilidad cero, $r_\theta(\delta_\xi) = \sup_\theta r_\theta(\delta_\xi)$, necesariamente (2.2) se cumple.

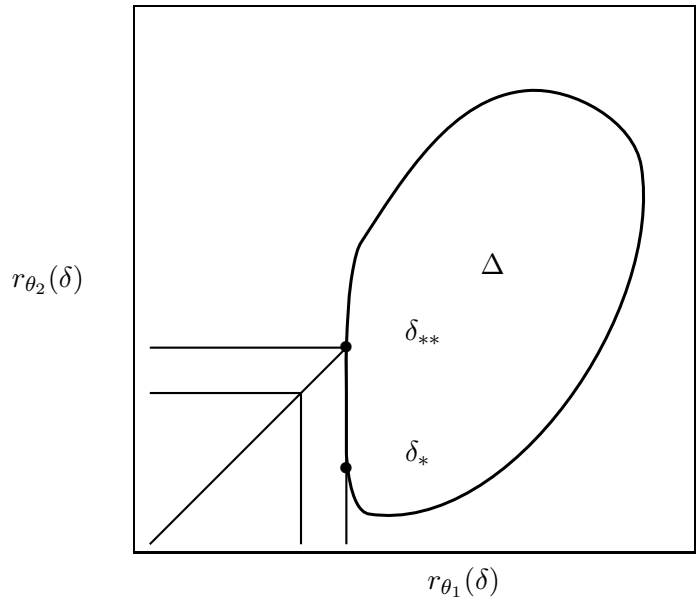
El teorema anterior y ambos corolarios proporcionan medios para caracterizar procedimientos como minimax, caracterización que en general no es fácil.

2.3. Caracterización de procedimientos admisibles.

La noción de admisibilidad se introdujo en la Definición 1.3, (pág. 4). Al igual que la condición de minimax, no es fácil en general demostrar que un procedimiento es admisible. En algunos casos particulares, sin embargo, es sencillo. El siguiente teorema es un instrumento útil para probar admisibilidad.

Teorema 2.2 *Un procedimiento de Bayes relativo a una cierta distribución* a priori *, si es único, es admisible.*

Figura 2.2: δ_{**} es minimax, pero no admisible. Es mejorado por δ_*



En efecto, supongamos un procedimiento de Bayes δ_ξ inadmisible. Existiría otro, δ_0, tal que $r_\theta(\delta_0) \leq r_\theta(\delta_\xi)$. Pero entonces:

$$R_\xi(\delta_0) = \sum_{\theta\in\Theta} r_\theta(\delta_0)\xi(\theta) \leq \sum_{\theta\in\Theta} r_\theta(\delta_\xi)\xi(\theta) = R_\xi(\delta_\xi)$$

contra la hipótesis de que δ_ξ es único de Bayes.

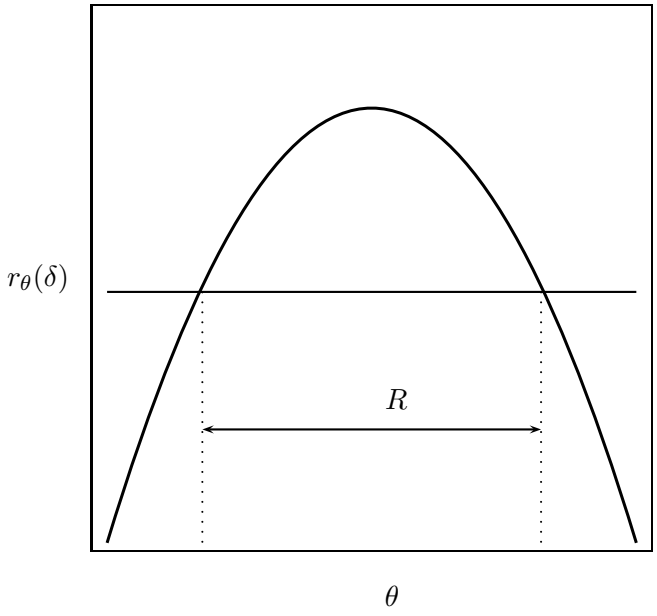
Por tanto, ¿es admisible todo procedimiento Bayes? Si es único, es claro que sí: acabamos de ver que no puede estar dominado por ningún otro. Pero puede ocurrir que para una cierta distribución *a priori* haya más de un procedimiento de Bayes, y sólo uno de ellos sea admisible. El ejemplo que sigue lo aclara.

> **Ejemplo 2.1** Consideremos el caso ilustrado en la Figura 2.2 Ambos procedimientos δ_{**} y δ_* son Bayes respecto a una distribución *a priori* que diera probabilidad uno a θ_1 (las líneas de igual riesgo de Bayes sería entonces verticales. Sólo la abscisa de un punto importaría: el riesgo bajo θ_2 sería irrelevante, porque θ_2 se presenta con probabilidad cero). Sin embargo, δ_* domina a δ_{**} —aunque en términos de riesgo ambos sean equivalentes—.

Situaciones como la que ilustra el ejemplo anterior pueden excluírse imponiendo alguna condición adicional, como sucede en el siguiente teorema.

Teorema 2.3 *Supongamos que $R_\xi(\delta) < \infty$, $\forall\delta$. Si: (i) Θ es discreto y $\xi(\theta) > 0$ para cada $\theta \in \Theta$, o bien: (ii) Θ es un intervalo con $\xi(\theta) > 0$ para todo θ en Θ, y, para cada posible δ, $r_\theta(\delta)$ es una función continua en θ, entonces cada procedimiento de Bayes relativo a $\xi(\theta)$ es admisible.*

Figura 2.3: Comparación de las funciones de riesgo de $\delta_*(\boldsymbol{X})$ y $\overline{Y}$, en el caso en que $n = 10$. R es la región en que el estimador minimax δ_* es mejor que $\overline{Y}$.



La demostración es inmediata. Ambas condiciones alternativas eliminan la posibilidad de múltiples procedimientos de Bayes que difieren sólo con probabilidad cero.

2.4. Búsqueda de procedimientos admisibles y minimax.

Las Secciones anteriores proporcionan algunos instrumentos, pero como se ha indicado la obtención de procedimientos tanto admisibles como minimax es una labor relativamente *ad-hoc*. Las siguientes consideraciones pueden ayudar.

Para probar que un procedimiento es admisible, basta probar que es Bayes y único para alguna distribución *a priori* (Teorema 2.2). Pero puede no ser fácil encontrar una tal distribución.

Una condición suficiente para ser minimax es ser Bayes respecto a la distribución *a priori* más desfavorable (Teorema 2.1), si tal distribución existe[1]. De nuevo puede no ser obvio cuál es esta distribución más desfavorable; pero una ayuda intuitiva es considerar aquellas distribuciones que más incertidumbre crean acerca del estado de la naturaleza prevalente (o que más "esparcen" el parámetro θ, si estamos ante un problema de estimación). Los siguientes dos ejemplos (que pueden encontrarse más desarrollados en Lehmann (1983)) ilustran las dificultades que se encuentran de ordinario.

[1]Nótese que tal existencia es *un supuesto* del Teorema 2.1.

Ejemplo 2.2 *(un procedimiento de Bayes con riesgo constante, y por tanto minimax)* Consideremos el caso en que tenemos una moneda no regular, cuya probabilidad θ de proporcionar "cara" (ó $Y = 1$) queremos estimar. Contamos con una muestra formada por n observaciones independientes, $Y_1, \ldots, Y_n$, y nos preguntamos si el estimador $\delta(\boldsymbol{Y}) = \overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i$ es minimax. Nuestra función de pérdida es cuadrática: $L(\theta, d) = (d - \theta)^2$.

Dado que $E[\overline{Y}] = \theta$, el riesgo (para un θ fijo) es:

$$r_\theta(\delta) = \frac{\theta(1-\theta)}{n}$$

cuyo máximo es $\frac{1}{4n}$, dado que $0 \leq \theta \leq 1$. Si $r_\theta(\delta)$ fuera $\frac{1}{4n}$ para cualquier θ, estaríamos ante un estimador minimax, pero éste no es el caso.

La siguiente cosa que se nos ocurriría es buscar una distribución *a priori* que hiciera el riesgo de Bayes igual a su valor máximo, $\frac{1}{4n}$. Es claro que tal distribución habría de ser la que diera al valor $\theta = \frac{1}{2}$ probabilidad igual a 1, ¡pero con tal distribución *a priori* el estimador de Bayes ya no sería $\overline{Y}$, sino $\frac{1}{2}$!

Ante el fracaso de estos dos intentos, podríamos ir a la búsqueda de una familia de distribuciones *a priori* y encontrar la familia de estimadores de Bayes asociados. Si tuviéramos la suerte de que alguno de ellos fuera único y de riesgo constante, entonces sería minimax (Teorema 2.1). Si tomamos una distribución *a priori* $\beta(r, s)$, el correspondiente procedimiento de Bayes es el que se obtuvo en el Ejemplo 1.9 (la función de riesgo se computó en el Ejemplo 1.10). ¿Hay alguna distribución $\beta(r, s)$ tal que el riesgo asociado al procedimiento de Bayes correspondiente sea constante? Tratemos de encontrar r y s verificando para una constante cualquiera y todo θ que:

$$\left(\frac{n}{n+r+s}\right)^2 \frac{\theta(1-\theta)}{n} + \left(\frac{r+n\theta}{n+r+s} - \theta\right)^2 = k$$

lo que implica, tras reducir a denominador común, que el numerador del lado izquierdo ha de ser constante:

$$n\theta - n\theta^2 + \left[r^2 + (r+s)^2\theta^2 - 2r(r+s)\theta\right] = c$$

Para ello es preciso que los coeficientes de θ y θ^2 sean cero:

$$\begin{aligned} n - 2r(r+s) &= 0 \\ (r+s)^2 - n &= 0 \end{aligned}$$

de donde:

$$r = s = \frac{1}{2}\sqrt{n}$$

Llevando estos dos valores a la fórmula (1.25) obtenemos el procedimiento minimax que buscamos:

$$\delta(\boldsymbol{Y}) = \left(\frac{n}{n+\sqrt{n}}\right)\frac{\sum Y_i}{n} + \frac{\frac{1}{2}\sqrt{n}}{n+\sqrt{n}} \tag{2.5}$$

$$= \frac{\sqrt{n}}{1+\sqrt{n}} \cdot \frac{\sum Y_i}{n} + \frac{1}{2}\frac{1}{1+\sqrt{n}} \tag{2.6}$$

Su riesgo (constante) es:

$$r_\theta(\delta) = r^2 \frac{1}{(n+r+s)^2} = \frac{1}{4(1+\sqrt{n})^2} \tag{2.7}$$

Es interesante comparar este riesgo con el del estimador insesgado habitual, $\overline{X} = n^{-1}\sum_i X_i$, que es $\theta(1-\theta)/n$. En el caso más desfavorable para este último (cuando $\theta = \frac{1}{2}$ y $r_\theta(\delta) = \frac{1}{4n}$, el estimador minimax es mejor. Sin embargo, esta reducción de riesgo en la situación más desfavorable tiene un precio; para otros valores de θ, el estimador minimax puede ser considerablemente peor que el estimador insesgado habitual. La Figura 2.3 (pág. 25) muestra la función de riesgo del estimador minimax (horizontal al nivel 0.01443) y la del estimador $\overline{X}$, ambas correspondientes a un tamaño muestral $n = 10$. Puede verse que para $0{,}18 \leq \theta \leq 0{,}82$ el estimador minimax es de menor riesgo, mientras lo contrario ocurre fuera de dicho intervalo. Es fácil comprobar también que a medida que $n \to \infty$ el intervalo en que el estimador minimax mejora a $\overline{X}$ se va estrechando en torno a $\theta = \frac{1}{2}$.

Ejemplo 2.3 Supongamos que hemos de estimar la media θ desconocida de una distribución normal $N(\theta, \sigma^2)$, cuya varianza supondremos por simplicidad conocida. Supondremos también que la distribución *a priori* de θ es $N(\mu, b^2)$, y la función de pérdida $L(\theta, d) = (d-\theta)^2$. Contamos con una m.a.s. $\boldsymbol{X} = (X_1, \ldots, X_n)$. ¿Cuál es el estimador minimax de θ?

Comencemos por encontrar el estimador de Bayes, y, si fuera de riesgo constante, podríamos entonces afirmar que es minimax.

Según comprobamos en el Ejemplo 1.11, la distribución *a posteriori* de θ es:

$$\theta|\boldsymbol{X} \sim N\left(\frac{\mu\sigma^2 + nb^2\overline{X}}{\sigma^2 + nb^2},\ \frac{b^2\sigma^2}{\sigma^2 + nb^2}\right)$$

De acuerdo con el Teorema 1.1, el procedimiento de Bayes será entonces:

$$\delta(\boldsymbol{X}) = \frac{\mu\sigma^2 + nb^2\overline{X}}{\sigma^2 + nb^2}$$

y su riesgo:

$$r_\theta(\delta) = E_\theta\left[\delta(\boldsymbol{X}) - \theta\right]^2 = \frac{nb^4\sigma^2}{(\sigma^2 + nb^2)^2} + \left(\frac{\mu\sigma^2 + nb^2\theta}{\sigma^2 + nb^2} - \theta\right)^2$$

De esta última expresión deducimos que el riesgo no es constante y por tanto $\delta(\boldsymbol{X})$ no es minimax. Observemos, sin embargo, que $\overline{X}$, límite de procedimientos de Bayes cuando $n \to \infty$, si tiene riesgo constante (=σ^2/n), y por tanto es minimax. La distribución más desfavorable es la distribución *a priori* difusa.

Ejemplo 2.4 *(un procedimiento de Bayes en que los estados más desfavorables totalizan probabilidad 1; y, por tanto, un procedimiento minimax en virtud del Corolario 2.2)* Consideremos el espacio paramétrico $\Theta = \left\{\theta : \frac{1}{3} \leq \theta \leq \frac{2}{3}\right\}$, la función de pérdida

$$L(\hat{\theta}, \theta) \quad = \quad (\hat{\theta} - \theta)^2.$$

Podemos observar una variable aleatoria binaria tal que $P(X = 1) = 1 - P(X = 0) = \theta$. Consideramos el procedimiento estadístico

$$\hat{\theta} = \delta(X) = \begin{cases} a & \text{si } X = 0, \\ b & \text{si } X = 1. \end{cases} \tag{2.8}$$

El riesgo de dicho procedimiento es

$$r_\theta(\hat{\theta}) = (1-\theta)(a-\theta)^2 + \theta(b-\theta)^2. \tag{2.9}$$

Parece que una distribución máximamente desfavorable podría ser

$$\xi(\theta) = \begin{cases} \frac{1}{2} & \text{si } \theta = \frac{1}{3}, \\ \frac{1}{2} & \text{si } \theta = \frac{2}{3}. \end{cases} \tag{2.10}$$

El riesgo de Bayes entonces sería

$$R_\xi(\hat{\theta}) \quad = \quad \frac{5 - 8a + 9a^2 - 10b + 9b^2}{18}$$

Maximizando la expresión anterior respecto a a y b obtenemos $a = \frac{4}{9}$ y $b = \frac{5}{9}$. Sustituyendo estos valores en (2.9) obtenemos

$$r_\theta(\hat{\theta}) \quad = \quad \frac{1}{18} + \frac{7}{9}\left(\theta - \frac{1}{2}\right)^2,$$

que toma idéntico valor en $\theta = \frac{1}{3}$ y en $\theta = \frac{2}{3}$. Por tanto, estamos ante un procedimiento con valor constante para un conjunto de estados cuya probabilidad conjunta es 1. En virtud del Corolario 2.2, dicho procedimiento es minimax.

Capítulo 3

La familia exponencial. Suficiencia

3.1. Familia exponencial.

Definición 3.1 *Sea $F_X(x;\theta)$ una función de distribución dependiendo de un único parámetro. Se dice que pertenece a la familia exponencial si su función de densidad (o cuantía, en su caso) puede expresarse así:*

$$f_X(x;\theta) = \exp\{a(\theta)b(x) + c(\theta) + d(x)\} \tag{3.1}$$

Esto debe ocurrir sobre el soporte de X, y tal soporte no depender de θ.

Puede encontrarse una definición más precisa en Lehmann (1983), p. 26. Un ejemplo de distribución en la que el soporte depende del parámetro es la uniforme $U(0,\theta)$.

En el caso de distribuciones dependiendo de k parámetros, $\boldsymbol{\theta}$, la definición anterior se generaliza de la manera obvia, requiriendo que:

$$f_X(x;\boldsymbol{\theta}) = \exp\left\{\sum_{i=1}^{k} a_i(\boldsymbol{\theta})b_i(x) + c(\boldsymbol{\theta}) + d(x)\right\} \tag{3.2}$$

Ejemplo 3.1 Si $X \sim N(\mu, \sigma^2)$, su función de densidad puede escribirse en la forma:

$$\begin{aligned} f_X(x;\boldsymbol{\theta}) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \\ &= \exp\left\{-\frac{1}{2}\frac{x^2}{\sigma^2} - \frac{1}{2}\frac{\mu^2}{\sigma^2} + \frac{x\mu}{\sigma^2} + \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)\right\} \\ &= \exp\left\{\sum_{i=1}^{2} a_i(\boldsymbol{\theta}) b_i(x) + c(\boldsymbol{\theta}) + d(x)\right\} \end{aligned}$$

con:

$$\begin{aligned} \boldsymbol{\theta} &= (\mu, \sigma^2)' \\ a_1(\boldsymbol{\theta}) &= -\frac{1}{2\sigma^2} \\ a_2(\boldsymbol{\theta}) &= \frac{\mu}{\sigma^2} \\ b_1(x) &= x^2 \\ b_2(x) &= x \\ c(\boldsymbol{\theta}) &= -\frac{1}{2}\frac{\mu^2}{\sigma^2} + \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) \\ d(x) &= 0 \end{aligned}$$

Ejemplo 3.2 Si $X \sim b(p, n)$ tenemos que para $x \in \{0, 1, \ldots, n\}$ y $p \in (0, 1)$:

$$P_X(x;p) = \binom{n}{x} p^x (1-p)^{n-x} = \exp\left\{\log\binom{n}{x} + x\log(p) + (n-x)\log(1-p)\right\} \quad (3.3)$$

que responde a la forma general en (3.1) con:

$$\begin{aligned} \theta &= p \\ a(\theta) &= \log(p) - \log(1-p) = \log\left(\frac{p}{1-p}\right) \\ b(x) &= x \\ c(\theta) &= n\log(1-p) \\ d(x) &= \log\binom{n}{x} \end{aligned}$$

Ejemplo 3.3 La distribución de Weibull tiene por función de densidad,

$$f_X(x;\alpha,\beta) = \frac{\beta}{\alpha^\beta} x^{\beta-1} \exp\left\{-\left(\frac{x}{\alpha}\right)^\beta\right\} \quad (3.4)$$

para $x > 0$, $\alpha > 0$ y $\beta > 0$. Es fácil ver que no puede expresarse en la forma (3.1), y por tanto no pertenece a la familia exponencial.

Se llama *parámetro natural* de la distribución (3.5) a $\eta = a(\theta)$. En términos del parámetro natural, si $a(.)$ es una función 1-1, la expresión (3.1) queda en forma canónica o simplificada:

$$f_X(x, \boldsymbol{\eta}) = \exp\{\eta b(x) + A(\eta) + d(x)\}. \tag{3.5}$$

En el caso de distribuciones k-paramétricas, (3.5) se generaliza a

$$f_X(x; \boldsymbol{\eta}) = \exp\left\{\sum_{i=1}^{k} \eta_i b_i(x) + A(\boldsymbol{\eta}) + d(x)\right\}. \tag{3.6}$$

En una distribución binomial, el parámetro natural es el logaritmo de la razón de probabilidades *(log odds)* (Ejemplo 3.2, más arriba). Véase también el ejemplo que sigue.

Ejemplo 3.4 En una distribución de Poisson, cuya función de probabilidad es

$$f_X(x; \theta) = \frac{e^{-\theta}\theta^x}{x!}$$

con $x = 1, 2, 3, \ldots$ y $\theta > 0$, el parámetro natural es $\log_e \theta$, como se comprueba sin más que reescribir la función de probabilidad en forma canónica:

$$f_X(x; \theta) = \exp\{-\theta + x \log_e \theta - \log_e x!\}.$$

De (3.5), dado que

$$\int f_X(x, \boldsymbol{\eta}) = \int \exp\{\eta b(x) + A(\eta) + d(x)\} = 1,$$

se deduce:

$$e^{A(\eta)} \int \exp\{\eta b(x) + d(x)\} = 1$$

y por tanto

$$A(\eta) = -\log \int \exp\{\eta b(x) + d(x)\}.$$

El conjunto de valores para los cuales la integral anterior es finita se denomina *espacio del parámetro natural;* es el conjunto de valores de η que hacen que (3.5) defina una distribución. Se llama a $b(\boldsymbol{x})$ *estadístico canónico* de la distribución. En el Ejemplo 3.4 el parámetro natural es $\log\theta$ y el espacio del parámetro natural es $(-\infty, +\infty)$.

3.2. Suficiencia.

Definición 3.2 *Sea $\boldsymbol{X} = (X_1, \ldots, X_n)'$ una muestra generada por una distribución $F_X(x;\theta)$. Se dice que $S = S(\boldsymbol{X})$ es un estadístico suficiente respecto de θ (o "suficiente para θ") en la familia $\{F_X(x,\theta),\ \theta \in \Theta\}$ si:*

$$f_{\boldsymbol{X}|S}(\boldsymbol{x}|s) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{f_S(s;\theta)} \tag{3.7}$$

no depende de θ.

La denominación de suficiente para el estadístico S se justifica porque, en cierto sentido, el conocimiento de S proporciona cuanta información existe en la muestra acerca de θ. Podemos imaginar el espacio muestral de $\boldsymbol{X}$ dividido en regiones, cada una de ellas proporcionando el mismo valor de S. Una vez que sabemos el valor de S, la distribución de $\boldsymbol{X}$ condicionada por $S = s$ es independiente de θ, *y por tanto el conocer qué muestra concreta $\boldsymbol{x}$ ha dado lugar a $S = s$ es no informativo acerca de θ*. El siguiente ejemplo aporta plausibilidad intuitiva a la afirmación anterior.

> **Ejemplo 3.5** Supongamos dos urnas, con los siguientes contenidos. La urna A contiene 50 bolas blancas, 20 negras, y 30 azules. La urna B contiene 50 bolas blancas, 40 negras y 10 azules. Si nos presentan una de ambas urnas, sin indicarnos cuál, y al extraer una bola resulta ser blanca, este hecho es no informativo acerca de la identidad de la urna. Ambas pueden generar bola blanca en una extracción al azar con la misma probabilidad. El observar algo que dos o más estados de la naturaleza pueden generar con la misma probabilidad es no informativo acerca de cuál sea el estado de la naturaleza prevalente.

Un segundo ejemplo que exhibe suficiencia en un caso extremadamente simple es el siguiente.

> **Ejemplo 3.6** Sea una población binaria de parámetro θ de la que nos es posible obtener dos observaciones, X_1 y X_2. A efectos de inferencia sobre el parámetro θ (probabilidad de obtener $X_i = 1$) parece que sólo el número total de "unos" obtenidos en las dos observaciones importa, y que es irrelevante, en el caso de obtener un único valor "uno", saber si se ha producido en la primera observación o en la segunda. Ello sugeriría que $S(\boldsymbol{X}) = X_1 + X_2$ es suficiente para θ en la familia de distribuciones binarias. Veamos que éste es efectivamente el caso, comprobando que al condicionar sobre $S(\boldsymbol{X})$ la

distribución resultante no depende de θ:

$$\begin{aligned}
\text{Prob}\{\boldsymbol{X} = (0,0)|X_1 + X_2 = 0\} &= 1 \\
\text{Prob}\{\boldsymbol{X} = (0,0)|X_1 + X_2 \neq 0\} &= 0 \\
\text{Prob}\{\boldsymbol{X} = (0,1)|X_1 + X_2 = 1\} &= \frac{1}{2} \\
\text{Prob}\{\boldsymbol{X} = (1,0)|X_1 + X_2 = 1\} &= \frac{1}{2} \\
\text{Prob}\{\boldsymbol{X} = (1,1)|X_1 + X_2 = 2\} &= 1
\end{aligned}$$

probabilidades que, en todos los casos, son independientes de θ. Las probabilidades no recogidas en la relación anterior son todas cero, de manera también independiente de θ.

El siguiente teorema, de inmediata demostración, muestra que la noción realmente relevante es la de *partición suficiente*, y que un estadístico suficiente no hace sino "etiquetar" las clases de una tal partición.

Teorema 3.1 *Todo estadístico $T = \gamma(S)$ función 1-1 de un estadístico suficiente S es suficiente.*

DEMOSTRACION:

En efecto,

$$\begin{aligned}
\text{Prob}\{\boldsymbol{X} = \boldsymbol{x}|\gamma(S(\boldsymbol{X})) = b;\theta\} &= \text{Prob}\left\{\boldsymbol{X} = \boldsymbol{x}|S(\boldsymbol{X}) = \gamma^{-1}(b);\theta\right\} \\
&= \text{Prob}\left\{\boldsymbol{X} = \boldsymbol{x}|S(\boldsymbol{X}) = \gamma^{-1}(b)\right\}
\end{aligned}$$

en que la omisión en el último término de la igualdad de θ como argumento se justifica por la suficiencia de $S(\boldsymbol{X})$.

Si definimos $A_S = \{a_s\}$, partición asociada al estadístico suficiente S, como el conjunto de clases de equivalencia formadas por puntos $\boldsymbol{x}$ con igual valor de $S(\boldsymbol{x})$, vemos que lo que realmente interesa saber a efectos de inferencia sobre el parámetro θ no es cuál es el valor tomado por S, un determinado estadístico suficiente, sino la clase de equivalencia en la que está $\boldsymbol{x}$.

Es también claro que cualquier partición "mas fina" que A_S (es decir, cualquier partición formada por clases de equivalencia $b_{s'}$ con la propiedad de que para cualquier $b_{s'}$ hay un $a_s \text{ tal que } b_{s'} \subseteq a_s$) es también suficiente. Intuitivamente, si el saber en que clase a_s esta $\boldsymbol{x}$ es cuanto necesitamos a efectos de hacer inferencia sobre θ, el saber que $\boldsymbol{x} \in b_{s'} \subseteq a_s$ es *a fortiori* suficiente. Un argumento formal sería el proporcionado por el teorema a continuación.

Teorema 3.2 *Si A_S es una partición suficiente y $B_{s'}$ es una partición más fina, entonces $B_{s'}$ es también una partición suficiente.*

DEMOSTRACION:

Existe a_s verificando $b_{s'} \subseteq a_s$. Se tiene entonces que:

$$\begin{aligned}
\text{Prob}\{\boldsymbol{X} = \boldsymbol{x}|b_{s'}\} &= \frac{\text{Prob}\{(\boldsymbol{X} = \boldsymbol{x}) \cap (\boldsymbol{X} \in b_{s'})\}}{\text{Prob}\{\boldsymbol{X} \in b_{s'}\}} \\
&= \frac{\text{Prob}\{(\boldsymbol{X} = \boldsymbol{x}) \cap (\boldsymbol{X} \in (b_{s'} \cap a_s))\} / \text{Prob}\{a_s\}}{\text{Prob}\{\boldsymbol{X} \in (b_{s'} \cap a_s)\} / \text{Prob}\{a_s\}} \\
&= \frac{\text{Prob}\{(\boldsymbol{X} = \boldsymbol{x}) \cap (\boldsymbol{X} \in b_{s'})|\boldsymbol{X} \in a_s\}}{\text{Prob}\{\boldsymbol{X} \in b_{s'}|\boldsymbol{X} \in a_s\}}
\end{aligned}$$

y esta última expresión es independiente de θ por suficiencia de A_S, lo que implica que Prob $\{\boldsymbol{X} = \boldsymbol{x}|b_{s'}\}$ también lo es.

El teorema anterior tiene una consecuencia inmediata: si un estadístico S suficiente puede expresarse como función de otro estadístico T, entonces T es también suficiente. En efecto, si $T(\boldsymbol{x}) = T(\boldsymbol{y})$, entonces $S(\boldsymbol{x}) = S(\boldsymbol{y})$; dos muestras que den lugar al mismo valor de T dan lugar al mismo valor de S, y, en consecuencia, es indiferente obtener una u otra a efectos de inferencia sobre θ.

Un estadístico suficiente que puede obtenerse como función de cualquier otro estadístico suficiente, se dice que es *mínimo suficiente*. La partición del espacio muestral en clases cada una de las cuales da lugar al mismo valor de un estadístico mínimo suficiente, es la partición menos fina que conserva la suficiencia.

Los siguientes ejemplos de estadísticos y particiones suficientes ilustran los conceptos anteriores.

Ejemplo 3.7 Consideremos la estimación del parámetro media en una distribución uniforme $U(0, 2\theta)$ (cuya media, por tanto, es θ). Podemos tomar una muestra $\boldsymbol{X} = (X_1, \ldots, X_n)'$, cuyos valores ordenados denominaremos por $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. Es fácil ver que un estadístico suficiente para 2θ lo es también para θ, y viceversa. Es también muy intuitivo que la media aritmética —estimador habitual de la media poblacional— no es suficiente en nuestro caso. Supongamos que $n = 3$ y que los tres valores muestrales en una experimentación concreta son: 1.2, 1.1, y 6.7. La media aritmética sería $(1{,}2 + 1{,}1 + 6{,}7)/3 = 3{,}0$. Sin embargo, es claro que hay información en la muestra que permite mejorar nuestra estimación de θ sobre la proporcionada por la media aritmética. El saber que una observación es 6.7 nos muestra que $2\theta \geq 6{,}7$, y por tanto $\theta \geq 3{,}35$.

El argumento anterior sugiere que $X_{(n)}$ —el mayor de los valores muestrales, o *n-ésimo estadístico de orden*— es particularmente informativo acerca de θ en la clase de distribuciones uniformes $U(0, 2\theta)$. Haciendo uso de la Definición 3.2 vamos a demostrar que tal estadístico es suficiente.

Sea $S = X_{(n)}$. Entonces,

$$\begin{aligned}
F_S(s;\theta) &= \text{Prob}\left\{X_{(n)} \leq s\right\} \\
&= \text{Prob}\{\cap_{i=1}^{n}(X_i \leq s)\} \\
&= \prod_{i=1}^{n} \text{Prob}\{X_i \leq s\} \\
&= \left(\frac{s}{2\theta}\right)^n
\end{aligned}$$

Derivando esta última expresión tenemos:

$$f_S(s;\theta) = \frac{ns^{n-1}}{(2\theta)^n}, \qquad (0 < s < 2\theta)$$

Por otra parte:

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^{n} f_X(x;\theta) = \frac{1}{(2\theta)^n}$$

Por consiguiente:

$$f_{\boldsymbol{X}|S}(\boldsymbol{x}|s) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{f_S(s;\theta)} = \frac{1}{ns^{n-1}}$$

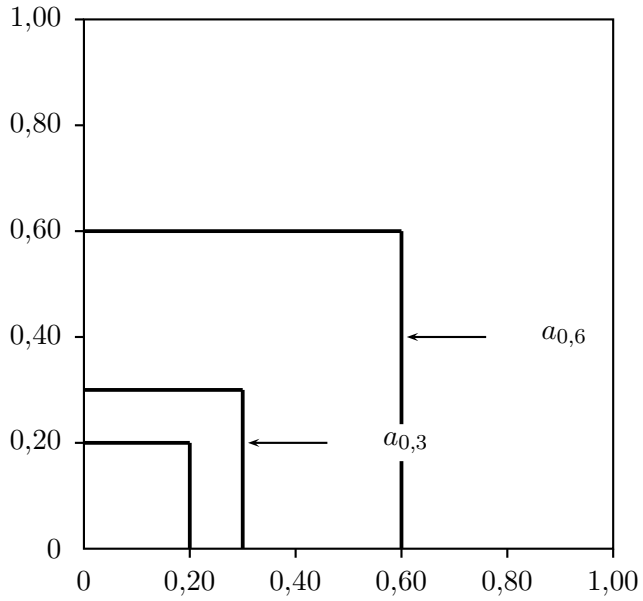
expresión independiente de θ lo que, de acuerdo con con la Definición 3.2, establece la suficiencia de $S = X_{(n)}$.

En este caso, las clases de equivalencia en que queda dividido el espacio muestral son las de expresión genérica siguiente:

$$a_s = \left\{ \boldsymbol{x} : \max_i x_i = s \right\}$$

Cuando $n = 2$ dichas clases serían las que ilustra la Figura 3.1; bordes superior y derecho de cuadrados de lado s apoyados sobre los ejes de coordenadas.

Figura 3.1: Clases de equivalencia en la partición mínima suficiente. Distribución $U(0, 2\theta)$ con $n = 2$. $a_{0,3}$ y $a_{0,6}$ denotan las clases correspondientes a $s = 0{,}3$ y $s = 0{,}6$ del estadístico suficiente $S = \max\{X_1, X_2\}$



Ejemplo 3.8 Consideremos ahora el caso de una muestra aleatoria simple $\boldsymbol{X} = (X_1, \ldots, X_n)'$ procedente de una distribución de Poisson, $P(\lambda)$. Comprobemos que $\overline{X}$ o, alternativamente, $\sum_{i=1}^n X_i$ es un estadístico suficiente para la media, λ. Como la suma de n v.a. independientes con distribución $P(\lambda)$ se distribuye como $P(n\lambda)$, tenemos que si $S = \sum_{i=1}^n X_i$:

$$P_S(s;\lambda) = \frac{e^{-n\lambda}(n\lambda)^s}{s!}$$

Por otra parte:

$$P_{\boldsymbol{X}}(\boldsymbol{x};\lambda) = \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^s}{\prod_{i=1}^n x_i!}$$

En consecuencia:

$$f_{\boldsymbol{X}|S}(\boldsymbol{x}|s) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x};\lambda)}{f_S(s;\lambda)} = \frac{s!}{n^s \prod_{i=1}^n x_i!}$$

que es independiente del parámetro λ. Se trata de una distribución multinomial de parámetros $\frac{1}{n}, \ldots, \frac{1}{n}, s$.

La comparación de este ejemplo con el anterior muestra que lo que en una familia de distribuciones es un estadístico suficiente para la media, puede no serlo en otra.

Observación 3.1 Esto obliga a ser cauto en el trabajo estadístico aplicado, y a no apelar alegremente a la noción de suficiencia para prescindir de información. Un estadístico suficiente contiene cuanta información puede la muestra aportar sobre un parámetro *si nuestros supuestos sobre la familia de distribuciones generadora de la muestra son correctos*. No en otro caso. Y, en la práctica, esta certeza acerca del modelo teórico adecuado rara vez se tiene. Por el contrario, es frecuente el caso de distribuciones difícilmente distinguibles cuando sólo se cuenta con muestras pequeñas o moderadas, que tienen muy diferentes estadísticos suficientes. Un caso claro lo ofrecerían las distribuciones $N(\theta, \sigma^2)$ y de Cauchy con parámetro de localización θ, $\mathcal{C}(\theta)$.

Ejemplo 3.9 Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple y denotemos sus correspondientes valores ordenados por $(X_{(1)}, \ldots, X_{(n)})$. Conocidos $(X_{(1)}, \ldots, X_{(n)})$, cualquiera de las permutaciones dando lugar a tales valores ordenados puede haberse presentado con la misma probabilidad. Por consiguiente:

$$\text{Prob}\left\{(X_1, \ldots, X_n)|(X_{(1)}, \ldots, X_{(n)})\right\} = \frac{1}{n!}$$

sea cual fuere la distribución generadora $F_X(x;\theta)$. Por lo tanto, $(X_{(1)}, \ldots, X_{(n)})$ es un estadístico suficiente.

Ejemplo 3.10 Consideremos el caso en que $\Theta = \{\theta_0, \theta_1\}$ y las dos posibles distribuciones $F_X(x;\theta)$ tienen soporte común. Entonces, la razón de verosimilitudes:

$$R(\boldsymbol{x}) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)}{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_1)}$$

es un estadístico mínimo suficiente. En efecto,

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}|R(\boldsymbol{x})=r;\theta_0) &= \frac{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)}{\displaystyle\int_{R(\boldsymbol{X})=r} f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)d\boldsymbol{x}} \\ &= \frac{r f_{\boldsymbol{X}}(\boldsymbol{x};\theta_1)}{\displaystyle\int_{R(\boldsymbol{X})=r} r f_{\boldsymbol{X}}(\boldsymbol{x};\theta_1)d\boldsymbol{x}} \\ &= \frac{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_1)}{\displaystyle\int_{R(\boldsymbol{X})=r} f_{\boldsymbol{X}}(\boldsymbol{x};\theta_1)d\boldsymbol{x}} \\ &= f_{\boldsymbol{X}}(\boldsymbol{x}|R(\boldsymbol{x})=r;\theta_1) \end{aligned}$$

lo que muestra que la densidad condicionada no depende del valor de θ.

3.3. Caracterización de estadísticos suficientes.

La aplicación directa de la Definición 3.2 es con frecuencia tediosa, y por otra parte requiere una conjetura previa acerca de qué estadístico S puede ser suficiente. El siguiente teorema es de aplicación frecuentemente mucho más rápida y directa.

Teorema 3.3 (Teorema de factorización) *Una condición necesaria y suficiente para que $S = S(\boldsymbol{X})$ sea suficiente para θ en la familia de distribuciones $\{F_X(x;\theta), \theta \in \Theta\}$ es que la verosimilitud de la muestra pueda factorizarse así:*

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = g_S(s;\theta)h(\boldsymbol{x}) \tag{3.8}$$

siendo $g_S(s;\theta)$ la función de densidad de S y $h(\boldsymbol{x})$ una función dependiente sólo de $\boldsymbol{x}$, pero no de θ.

DEMOSTRACION:

i) (Necesidad). Supongamos que S es suficiente. Ello quiere decir, de acuerdo con la Definición 3.2, que:

$$f_{\boldsymbol{X}|S}(\boldsymbol{x}|s) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{f_S(s;\theta)} \tag{3.9}$$

y por tanto:

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \underbrace{f_{\boldsymbol{X}|S}(\boldsymbol{x}|s)}_{h(\boldsymbol{x})}\underbrace{f_S(s;\theta)}_{g_S(s;\theta)} \tag{3.10}$$

ii) (Suficiencia). Denominemos $\Delta(s)$ el conjunto formado por todos los posibles valores muestrales $\boldsymbol{x}$ dando lugar al valor $S = s$, y supongamos que (3.8) se

verifica. Entonces:

$$\begin{aligned} f_{\boldsymbol{X}|S}(\boldsymbol{x}|s) &= \frac{f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{f_S(s;\theta)} = \frac{f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{\sum_{\boldsymbol{x}\in\Delta(s)} f_{\boldsymbol{X}}(\boldsymbol{x};\theta)} \\ &= \frac{g_S(s;\theta)h(\boldsymbol{x})}{g_S(s;\theta)\sum_{\boldsymbol{x}\in\Delta(s)} h(\boldsymbol{x})} = \frac{h(\boldsymbol{x})}{\sum_{\boldsymbol{x}\in\Delta(s)} h(\boldsymbol{x})} \end{aligned}$$

y el último término de la derecha es independiente de θ, lo que establece la suficiencia de S en virtud de la Definición 3.2. El anterior argumento supone que X es una variable discreta y $\Delta(s)$ un conjunto de probabilidad no nula; en el caso de una distribución continua, los sumatorios en la expresión anterior deben reemplazarse por integrales.

Ejemplo 3.11 Sea una distribución $N(\theta, 1)$, y una muestra formada por n observaciones de la misma, $X_1, \ldots, X_n$. La verosimilitud puede escribirse así:

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x};\theta) &= \exp\left\{-\frac{1}{2}\sum_i (x_i-\theta)^2 + n\log\frac{1}{\sqrt{2\pi}}\right\} \\ &= \exp\left\{-\frac{1}{2}\sum_i \left(x_i^2 - 2x_i\theta + \theta^2\right) + n\log\frac{1}{\sqrt{2\pi}}\right\} \\ &= \exp\left\{\sum_i x_i\theta - \frac{1}{2}n\theta^2\right\}\exp\left\{-\frac{1}{2}\sum_i x_i^2 + n\log\frac{1}{\sqrt{2\pi}}\right\} \end{aligned}$$

Podemos en la anterior expresión identificar sin dificultad $\sum_i x_i$ como estadístico suficiente para θ, de acuerdo con el teorema de factorización.

Ejemplo 3.12 En el Ejemplo 3.9, pág. 36, se comprobó que la $(X_{(1)}, \ldots, X_{(n)})$, la muestra ordenada, era suficiente. Ciertamente, es un estadístico suficiente bastante trivial, que no efectúa una gran reducción de la muestra. En ocasiones, sin embargo, es todo lo lejos que se puede ir.

La distribución de Cauchy con parámetro de localización θ, $\mathcal{C}(\theta)$, proporciona una ilustración simple de ello. La densidad de una muestra $(x_1, \ldots, x_n)$ es de la forma

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^{n}\left[\frac{1}{\pi}\frac{1}{1+(x_i-\theta)^2}\right],$$

para $-\infty < x_i < \infty$, e $i = 1, \ldots, n$. Puede verse fácilmente que cualquier intento de factorizar la expresión anterior obliga a englobar en $g_S(s;\theta)$ una función s de la muestra que depende de todos los valores muestrales. No es posible ninguna reducción: $S = (X_{(1)}, \ldots, X_{(n)})$ es mínimo suficiente.

Ejemplo 3.13 En el Ejemplo 3.7, pág. 34, se comprobó que en el caso de una distribución uniforme $U(0, 2\theta)$ el mayor estadístico de orden $X_{(n)}$ es suficiente para θ. Podemos llegar al mismo resultado haciendo uso del teorema de factorización. En efecto,

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = (2\theta)^{-n}H(2\theta - x_{(n)})$$

con $H(z) = 1$ cuando $z > 0$ y $H(z) = 0$ en caso contrario. Por tanto, $-2\theta^n H(2\theta - x_{(n)})$ juega el papel de $g_S(s;\theta)$ en (3.8), y $x_{(n)}$ es suficiente.

Ejemplo 3.14 La minimalidad en el Ejemplo 3.10 también es simple de establecer haciendo uso del teorema de factorización. Bastará para ello comprobar que, sea cual fuere el estadístico suficiente U que consideremos, $R(\boldsymbol{X}) = H(U)$ para alguna función $H()$. Esto sucede:

$$R(\boldsymbol{X}) = \frac{f_{\boldsymbol{X}}(\boldsymbol{X};\theta_0)}{f_{\boldsymbol{X}}(\boldsymbol{X};\theta_1)} = \frac{g_U(U;\theta_0)h(\boldsymbol{X})}{g_U(U;\theta_1)h(\boldsymbol{X})} = H(U)$$

3.4. Completitud, ancilaridad, y suficiencia.

Asociadas a la noción de suficiencia están las de ancilaridad y completitud.

Definición 3.3 *Dada una familia de distribuciones $\{F_X(x;\theta), \theta \in \Theta\}$ se dice que $V(\boldsymbol{X})$ es un estadístico* ancilar *si su distribución es independiente de θ. Es* ancilar de primer orden *si su valor medio no depende de θ.*

De acuerdo con el argumento esbozado inmediatamente después de la Definición 3.2, podemos considerar que un estadístico ancilar carece, por si mismo, de contenido informativo acerca de θ. Obsérvese, sin embargo, que un estadístico ancilar puede, en compañía de otro, ser muy informativo —quizá incluso suficiente— .

Ejemplo 3.15 Sea $X_{(1)}, \ldots, X_{(n)}$ una muestra aleatoria simple procedente de una población $U(0,\theta)$. Entonces, de modo enteramente análogo a como sucede en el Ejemplo 3.7 (pág. 34), $X_{(n)}$ es suficiente para θ, y es claro además que $X_{(1)}$ no es suficiente. Se puede demostrar, sin embargo, que $X_{(n)}/X_{(1)}$ sigue una distribución que para nada depende de θ, y es por tanto ancilar. ¡Y sin embargo, $X_{(1)}, X_{(n)}/X_{(1)}$ es suficiente! Vemos aquí como un estadístico ancilar, en compañía de otro que por sí sólo es bastante poco informativo acerca de θ, proporciona un estadístico suficiente. El ejemplo 8.11 en Garín y Tusell (1991) muestra con más detalle un caso similar.

Definición 3.4 *Un estadístico T es* completo *en la familia $\{F_X(x;\theta), \theta \in \Theta\}$ si no existe ninguna función de él (salvo la función constante, $\ell(T) = c$) que sea ancilar de primer orden. Es decir, si de $E_\theta[\ell(T)] = c$, $\forall\theta \in \Theta$, se deduce necesariamente que $\ell(T) = c$. Un estadístico es* acotado completo *si lo anterior se verifica para cualquier función $\ell()$ acotada.*

De nuevo la definición anterior tiene un contenido intuitivo notable. Un estadístico es completo si ninguna función de él —salvo la función constante— está desprovista de información acerca de θ. El significado de esto es más claro si consideramos un estadístico que *no* sea completo.

Ejemplo 3.16 Sea una distribución $N(\theta, 1)$, y una muestra formada por dos observaciones de la misma, (X_1, X_2). Claramente, $(X_2 - X_1)$ sigue una distribución que no depende de θ: $N(0, \sigma^2 = 2)$. Por tanto, $T = (X_1, X_2)$ no será un estadístico completo, y $\ell(T) = X_2 - X_1$ es ancilar de primer orden.

Ejemplo 3.17 El estadístico $X_{(1)}, X_{(n)}/X_{(1)}$ en el Ejemplo 3.15 no es completo; una parte de él, $X_{(n)}/X_{(1)}$ es ancilar.

3.5. Suficiencia y familia exponencial.

La inspección de la forma general de la densidad (o cuantía) de una distribución en la familia exponencial,

$$f_X(x;\theta) = \exp\{a(\theta)b(x) + c(\theta) + d(x)\}$$

muestra que, si se cumplen las condiciones que permiten aplicar el teorema de factorización (Teorema 3.3), se tendrá:

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x};\theta) &= \exp\left\{a(\theta)\sum_{i=1}^{n} b(x_i) + nc(\theta) + \sum_{i=1}^{n} d(x_i)\right\} \\ &= \exp\left\{a(\theta)\sum_{i=1}^{n} b(x_i) + nc(\theta)\right\}\exp\left\{\sum_{i=1}^{n} d(x_i)\right\} \\ &= g_S(s;\theta)h(\boldsymbol{x}) \end{aligned}$$

con:

$$s = \sum_{i=1}^{n} b(x_i)$$

La generalización al caso multiparamétrico es obvia, teniéndose entonces que:

$$\left(\sum_{i=1}^{n} b_1(x_i), \ldots, \sum_{i=1}^{n} b_k(x_i)\right)$$

son estadísticos conjuntamente suficientes para $(a_1(\theta), \ldots, a_k(\theta))$.

En general, pues, salvo en casos patológicos en que está vedado el empleo del Teorema 3.3, las distribuciones en la familia exponencial poseen estadísticos suficientes. La relación entre la pertenencia a dicha familia y la existencia de estadísticos suficientes va más allá sin embargo, como se desprende del siguiente,

Teorema 3.4 *(Teorema de Darmois) Sea X una variable aleatoria con densidad $f_X(x;\theta)$, $\theta \in \Theta$. Supongamos que el dominio de variación de X es independiente de θ, y que $(X_1, \ldots, X_n)$ es una m.a.s. de tamaño n de dicha variable. Entonces:*

i) Si existe $n > 1$ tal que $(X_1, \ldots, X_n)$ admite un estadístico suficiente,

$$f_X(x;\theta) = \exp\{a(\theta)b(x) + c(\theta) + d(x)\}.$$

ii) Si $f_X(x;\theta) = \exp\{a(\theta)b(x) + c(\theta) + d(x)\}$ y la aplicación $x_1 \to \sum_{i=1}^{n} b(x_i)$ es biunívoca para todo $x_1, \ldots, x_n$, entonces para $n \geq 1$ admite un estadístico suficiente. En particular, $r = \sum_{i=1}^{n} b(x_i)$ es uno.

La demostración puede hallarse en Fourgeaud y Fuchs (1967), p. 192.

Observación 3.2 El enunciado del teorema anterior puede sugerir que, en la familia exponencial, cuando hay un único parámetro, hay un estadístico suficiente escalar; o, más generalmente, que la dimensión del vector de parámetros y del estadístico suficiente son iguales. Ello es frecuentemente el caso, pero no siempre. Por ejemplo, consideremos el caso en que la probabilidad de que un sujeto sobreviva más de t unidades de tiempo es:

$$\text{Prob}\{T > t\} = e^{-\beta t}$$

y por tanto, la función de distribución de T, "tiempo de vida", es:

$$F_T(t) = 1 - e^{-\beta t}$$

Si en una muestra de N sujetos se producen d muertes en los momentos t_i, $(i = 1, \ldots, d)$, y los restantes $s = N - d$ sujetos permanecen todavía vivos en los momentos u_j, $(j = d+1, \ldots, N)$, la densidad conjunta puede escribirse así:

$$f_{\boldsymbol{T},\boldsymbol{U}}(\boldsymbol{t},\boldsymbol{u}) = \beta^d \exp\left\{-\beta\left(\sum_{i=1}^{d} t_i + \sum_{j=d+1}^{N} u_j\right)\right\} \tag{3.11}$$

$$= \exp\left\{-\beta\left(\sum_{i=1}^{d} t_i + \sum_{j=d+1}^{N} u_j\right) + d\log\beta\right\} \tag{3.12}$$

Hay un sólo parámetro, β. Sin embargo, como estadístico suficiente necesitamos tanto d como $\left(\sum_{i=1}^{d} t_i + \sum_{j=d+1}^{N} u_j\right)$; ambos conjuntamente son un estadístico suficiente. Se dice que estamos ante una *distribución curvada*; hay un sólo parámetro, pero es como si existieran dos (β y $\log\beta$). Este ejemplo concreto procede de Berkson (1980). Otro ejemplo puede verse en Lehmann (1983), pág. 45. En Cox y Hinkley (1974) pág. 28 y ss. se ofrecen ejemplos adicionales que muestran que el número de parámetros (q) y el de estadísticos suficientes (m) no tienen necesariamente que coincidir: tanto $m > q$ como $q > m$ son situaciones posibles.

3.6. Estadísticos suficientes y soluciones de Bayes.

Hemos justificado en la Sección 3.2 el interés de emplear estadísticos suficientes apelando a la intuición. Pueden ahora darse argumentos adicionales.

Recordemos (Sección 1.10) que estamos interesados en la clase de procedimientos de Bayes y sus límites, como punto de partida para localizar procedimientos admisibles. Pues bien: de acuerdo con (1.18), especificada una función de pérdida, el procedimiento de Bayes depende de $\boldsymbol{X}$ sólo a través de $f_{\theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$, que a

su vez depende de $\boldsymbol{X}$ sólo a través del estadístico suficiente $S(\boldsymbol{X})$. En efecto:

$$\begin{aligned} f_{\theta|\boldsymbol{X}}(\theta|\boldsymbol{x}) &= \frac{f_{\boldsymbol{X}|\theta}(\boldsymbol{x}\,|\theta)\xi(\theta)}{f_{\boldsymbol{X}}(\boldsymbol{x})} \\ &= \frac{g_S(s;\theta)h(\boldsymbol{x})\xi(\theta)}{\int g_S(s;\theta')h(\boldsymbol{x})\xi(\theta')d\theta'} \\ &= \frac{g_S(s;\theta)\xi(\theta)}{\int g_S(s;\theta')\xi(\theta')d\theta'} \\ &= G(s;\theta) \end{aligned}$$

Una vez constatado que el limitar nuestra atención a procedimientos que son función de estadísticos suficientes nos da acceso a todos los procedimientos de Bayes, es claro que desearemos la máxima simplificación, limitándonos a considerar estadísticos no sólo suficientes sino mínimos suficientes.

3.7. Caracterización de la suficiencia minimal.

Hemos visto (comentario tras el Teorema 3.1, pág. 33) que la noción realmente importante es la de partición suficiente. La *partición mínima suficiente* será la partición suficiente menos fina posible. Tenemos entonces el siguiente resultado.

Teorema 3.5 *Sea $X_1, \ldots, X_n$ una muestra generada por una distribución en la familia $\{F_X(x;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$. Sea $\mathcal{S}$ la partición del espacio muestral que se obtiene al agrupar en clases de equivalencia los puntos cuya razón de verosimilitudes no depende de Θ; es decir, denotando por $\sim$ la pertenencia a la misma clase de equivalencia, aquella partición tal que*

$$\boldsymbol{x} \sim \boldsymbol{y} \iff \frac{f_{\boldsymbol{X}}(\boldsymbol{y};\boldsymbol{\theta})}{f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})} = m(\boldsymbol{x},\boldsymbol{y}). \tag{3.13}$$

Entonces, $\mathcal{S}$ es mínima suficiente, y cualquier estadístico T tomando valores diferentes en cada clase $\mathcal{S}_t \in \mathcal{S}$ es mínimo suficiente.

DEMOSTRACION:

En lo que sigue, se hace la demostración para el caso de una distribución discreta; el caso continuo es sustancialmente idéntico en esencia, pero formalmente mas difícil de tratar. Comprobemos en primer lugar que la partición es suficiente. Sea,

$$g(t,\boldsymbol{\theta}) = \sum_{\boldsymbol{y}\in\mathcal{S}_t} f_{\boldsymbol{X}}(\boldsymbol{y};\boldsymbol{\theta}) \tag{3.14}$$

y definamos

$$h(\boldsymbol{x}|t) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})}{\sum_{\boldsymbol{y}\in\mathcal{S}_t} f_{\boldsymbol{X}}(\boldsymbol{y};\boldsymbol{\theta})} = \left[\sum_{\boldsymbol{y}\in\mathcal{S}_t} m(\boldsymbol{x},\boldsymbol{y})\right]^{-1}. \tag{3.15}$$

Es claro entonces que,

$$f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta}) \;=\; g(t,\boldsymbol{\theta})h(\boldsymbol{x}|t) \tag{3.16}$$

Como $g(t,\boldsymbol{\theta})$ depende de la muestra sólo a través de t y $h(\boldsymbol{x}|t)$ no depende de $\boldsymbol{\theta}$, el Teorema 3.3 garantiza la suficiencia de T.

Tenemos ahora que ver que T es mínimo suficiente. Bastaría para ello probar que, para cualquier otro estadístico suficiente U, $U(\boldsymbol{x}) = U(\boldsymbol{y}) \Longrightarrow T(\boldsymbol{x}) = T(\boldsymbol{y})$. Pero esto se deduce sin dificultad: como U es suficiente,

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta}) &= g_1(\boldsymbol{u}(\boldsymbol{x}),\theta)g_2(\boldsymbol{x}) \\ f_{\boldsymbol{X}}(\boldsymbol{y};\boldsymbol{\theta}) &= g_1(\boldsymbol{u}(\boldsymbol{y}),\theta)g_2(\boldsymbol{y}), \end{aligned}$$

y

$$\frac{f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})}{f_{\boldsymbol{X}}(\boldsymbol{y};\boldsymbol{\theta})} \;=\; \frac{g_1(\boldsymbol{u}(\boldsymbol{x}),\boldsymbol{\theta})g_2(\boldsymbol{x})}{g_1(\boldsymbol{u}(\boldsymbol{y}),\boldsymbol{\theta})g_2(\boldsymbol{y})} = \frac{g_2(\boldsymbol{x})}{g_2(\boldsymbol{y})}.$$

Como este último término es función exclusivamente de $\boldsymbol{x}$ y de $\boldsymbol{y}$, es claro que $\boldsymbol{x} \sim \boldsymbol{y}$ y en consecuencia $T(\boldsymbol{x}) = T(\boldsymbol{y})$.

■

Ejemplo 3.18 Consideremos una distribución binaria de la que se obtiene una muestra de tamaño n. Estarán en la misma clase de la partición mínima suficiente aquellos puntos verificando

$$\frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i}} = m(\boldsymbol{x},\boldsymbol{y});$$

ello requiere $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$.

Hay algunos otros resultados que permiten en ocasiones caracterizar la suficiencia minimal. Los enunciamos a continuación.

Teorema 3.6 *Si un estadístico es suficiente y acotado completo, es minimal suficiente.*

Una demostración puede encontrarse en Fourgeaud y Fuchs (1967).

Ejemplo 3.19 Comprobemos que $S = X_{(n)}$ es minimal suficiente en una distribución $U(0,2\theta)$. En el Ejemplo 3.7 vimos que S es suficiente para θ en dicha distribución, y que su función de densidad es

$$f_S(s;\theta) = \frac{ns^{n-1}}{(2\theta)^n};$$

podemos pues limitarnos ahora a comprobar que es acotado completo. De acuerdo con la Definición 3.4, pág. 39, basta que comprobemos que de $E[\ell(S)] = 0$ para todo θ se deduce necesariamente $\ell(S) = 0$. Y así es, pues derivando la igualdad

$$E[\ell(S)] = \int_0^{2\theta} \ell(s) \frac{n s^{n-1}}{(2\theta)^n} ds = 0 \tag{3.17}$$

respecto de su límite superior, obtenemos

$$\ell(2\theta) \frac{n(2\theta)^{n-1}}{(2\theta)^n} = 0$$

de donde se sigue que $\ell(2\theta) = 0$.

En la familia exponencial, es simple establecer suficiencia minimal. Es evidente en virtud del teorema de factorización y de la expresión (3.1) (ó (3.2), si estamos ante una familia multiparamétrica) que $\sum_j b(X_j)$ (o, en el caso multiparamétrico, $\sum_j b_1(X_j), \ldots, \sum_j b_k(X_j)$) son estadísticos suficientes. El siguiente teorema permite establecer suficiencia minimal.

Teorema 3.7 *Si X sigue una distribución en la familia exponencial y de rango completo*[1]*, entonces*

$$\left(\sum_j b_1(X_j), \ldots, \sum_j b_k(X_j) \right) \tag{3.18}$$

es mínimo suficiente.

DEMOSTRACION:Puede demostrarse como corolario del Teorema 3.5. En efecto, la condición de suficiencia mínima (3.13) requiere en el caso de distribuciones en la familia exponencial

$$\begin{aligned} \frac{f_{\boldsymbol{X}}(\boldsymbol{y};\boldsymbol{\theta})}{f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})} &= \frac{\exp\left\{\sum_{j=1}^k \left(a_j(\boldsymbol{\theta}) \sum_{i=1}^n b_j(y_i)\right) + nc(\boldsymbol{\theta}) + \sum_{i=1}^n d(y_i)\right\}}{\exp\left\{\sum_{j=1}^k \left(a_j(\boldsymbol{\theta}) \sum_{i=1}^n b_j(x_i)\right) + nc(\boldsymbol{\theta}) + \sum_{i=1}^n d(x_i)\right\}} \\ &= \exp\left\{ \sum_{j=1}^k a_j(\boldsymbol{\theta}) \left[\sum_{i=1}^n b_j(x_i) - \sum_{i=1}^n b_j(y_i) \right] + \sum_{i=1}^n d(x_i) - \sum_{i=1}^n d(y_i) \right\}. \end{aligned}$$

En el caso de rango completo, para que la expresión anterior no dependa de $\boldsymbol{\theta}$ sera preciso que

$$\sum_{i=1}^n b_j(x_i) = \sum_{i=1}^n b_j(y_i) \qquad (i = 1, 2, \ldots, k.)$$

[1]Se dice que la familia es de rango completo si $(a_1(\boldsymbol{\theta}), \ldots, a_k(\boldsymbol{\theta}))$ genera un conjunto conteniendo un rectángulo de dimensión k cuando $\boldsymbol{\theta}$ toma valores en Θ.

Por tanto, cada vector k-dimensional

$$\left(\sum_{i=1}^{n} b_1(x_i), \sum_{i=1}^{n} b_2(x_i), \ldots, \sum_{i=1}^{n} b_k(x_i)\right)$$

determina una clase de la partición mínima suficiente.

■

Ejemplo 3.20 Sea $X_1, \ldots, X_n$ una m.a.s. generada por una distribución $N(\mu, \sigma^2)$. Entonces, $(\overline{X}, S^2)$ es un estadístico mínimo suficiente para (μ, σ^2). En efecto,

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}, \mu, \sigma^2) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\ &= \exp\left\{-\frac{1}{2}\frac{\sum_{i=1}^n x_i^2}{\sigma^2} - \frac{n\mu^2}{2\sigma^2} + \frac{\mu\sum_{i=1}^n x_i}{\sigma^2} + n\log_e\left(\frac{1}{\sigma\sqrt{2\pi}}\right)\right\} \end{aligned}$$

La expresión anterior puede escribirse en la forma canónica de las densidades de la familia exponencial (véase (3.2) y Ejemplo 3.1),

$$f_X(x;\boldsymbol{\theta}) = \exp\left\{\sum_{i=1}^{k} a_i(\boldsymbol{\theta}) b_i(x) + nc(\boldsymbol{\theta}) + d(x)\right\}, \tag{3.19}$$

con

$$\begin{aligned} \boldsymbol{\theta} &= (\mu, \sigma^2) \\ a_1(\boldsymbol{\theta}) &= -\frac{1}{2\sigma^2} \\ a_2(\boldsymbol{\theta}) &= \frac{\mu}{\sigma^2} \\ \sum_{i=1}^{n} b_1(x_i) &= \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} b_2(x_i) &= \sum_{i=1}^{n} x_i \\ c(\boldsymbol{\theta}) &= -\frac{n\mu^2}{2\sigma^2} + n\log_e\left(\frac{1}{\sigma\sqrt{2\pi}}\right). \end{aligned}$$

Por consiguiente, en aplicación del Teorema 3.7, $(\sum x_i, \sum x_i^2)$ —o cualquier función biunívoca de él— es un estadístico suficiente para (μ, σ^2).

Ejemplo 3.21 Podríamos también llegar al mismo resultado del ejemplo anterior mediante aplicación del Teorema 3.5. La partición mínima suficiente sería aquélla que pusiera en la misma clase de equivalencia puntos $\boldsymbol{x}$, $\boldsymbol{y}$ verificando

$$\frac{f_{\boldsymbol{X}}(\boldsymbol{y};\boldsymbol{\theta})}{f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})} = m(\boldsymbol{x}, \boldsymbol{y}).$$

En nuestro caso,

$$\begin{aligned}\frac{f_{\boldsymbol{X}}(\boldsymbol{y};\boldsymbol{\theta})}{f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})} &= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(x_i-\mu)^2-(y_i-\mu)^2\right]\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}y_i^2-2\mu\left(\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}y_i\right)\right]\right\}.\end{aligned}$$

Para que esta función no dependa de μ ni de σ^2 todo lo que se requiere es que

$$\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}y_i^2 \tag{3.20}$$

$$\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}y_i \tag{3.21}$$

Por consiguiente $(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$, o cualquier función biunívoca de dicho estadístico, como por ejemplo $(\overline{x}, \sum_{i=1}^n (x_i - \overline{x})^2)$, es un estadístico mínimo suficiente.

CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

3.1 Utilícese el procedimiento en el Ejemplo 3.21 para mostrar que al estimar el modelo lineal ordinario $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ con las condiciones habituales más la de normalidad, $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{Y}$ y $SSE = (\boldsymbol{Y} - X\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - X\hat{\boldsymbol{\beta}})$ son conjuntamente suficientes para los parámetros $(\boldsymbol{\beta}, \sigma^2)$

3.2 En la familia de distribuciones uniformes, $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. encuéntrese un estadístico suficiente para θ. ¿Es completo?

3.3 Sea $X_1, \ldots, X_n$ una m.a.s. procedente de una distribución con densidad

$$f_X(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{si } x > \theta, \\ 0 & \text{n otro caso.} \end{cases}$$

Muéstrese que $X_{(1)}$ es suficiente para θ.

3.4 Sea $X_1, \ldots, X_n$ una m.a.s. procedente de una distribución beta con densidad

$$f_X(x;r,s) = \frac{1}{\beta(r,s)}x^{r-1}(1-x)^{s-1}$$

en que $0 < x < 1$, $r > 0$, $s > 0$ y $\beta(r,s)$ es la constante de normalización. Compruébese que $(\sum_i \log(X_i), \sum_i \log(1 - X_i))$ es suficiente para r y s.

3.5 Sean $Y_1, \ldots, Y_n$ variables aleatorias independientes con densidades respectivas $\lambda_j e^{-\lambda_j y_j}$, $\lambda_j > 0, j = 1, \ldots, n$. Supongamos que $\log(\lambda_j) = \theta x_j$, $j = 1, \ldots, n$, y que $x_1, \ldots, x_n$ son constantes fijas y positivas. Muéstrese que no es de rango completo.

Capítulo 4

Procedimientos insesgados.

4.1. La condición de insesgadez.

Vimos (Ejemplo 1.6, pág. 5) que la búsqueda de un procedimiento mejor que cualquier otro estaba condenada al fracaso. Pero se apuntó allí que quizá si nos restringimos a una clase de procedimientos "razonable", que excluya comportamientos excelentes en casos aislados y muy malos en todos los demás estados de la naturaleza, sí podríamos encontrar un procedimiento óptimo.

La restricción de insesgadez es una forma de imponer tal comportamiento "razonable" a los procedimientos que estamos dispuestos a considerar[1].

En un problema de decisión, se dice que el procedimiento $\delta(\boldsymbol{X})$ es *insesgado* si:

$$E_\theta L(\theta', \delta(\boldsymbol{X})) \geq E_\theta L(\theta, \delta(\boldsymbol{X})) \qquad \forall \theta, \theta' \in \Theta \tag{4.1}$$

Restringir nuestra atención a procedimientos que verifican (4.1) elimina de nuestra consideración procedimientos como $\delta_2(\boldsymbol{X})$ en el referido Ejemplo 1.6.

En problemas de estimación puntual de una función $\gamma(\theta)$ se dice que $\delta(\boldsymbol{X})$ es un procedimiento insesgado si:

$$E_\theta(\delta(\boldsymbol{X})) = \gamma(\theta) \qquad \forall \theta \in \Theta \tag{4.2}$$

Ambas condiciones de insesgadez (la dada por (4.1) y la dada por (4.2)) pueden reconciliarse fácilmente, dado que, salvo en condiciones bastante anómalas, se implican mutuamente. El siguiente ejemplo lo ilustra.

[1] En palabras de Lehmann (ver Lehmann (1983)) es una condición de "imparcialidad".

Ejemplo 4.1 Supongamos un problema de estimación puntual con función de pérdida cuadrática. La condición de insesgadez (4.1) requiere:

$$E_\theta(\theta' - \delta(\boldsymbol{X}))^2 \geq E_\theta(\theta - \delta(\boldsymbol{X}))^2 \qquad \forall \theta, \theta' \in \Theta \tag{4.3}$$

Sumando y restando $E_\theta(\delta(\boldsymbol{X}))$ en el interior de cada paréntesis y tomando valor medio, tras simplificar tenemos:

$$E_\theta\left[\theta' - E_\theta\delta(\boldsymbol{X})\right]^2 \geq E_\theta\left[\theta - E_\theta\delta(\boldsymbol{X})\right]^2 \qquad \forall \theta, \theta' \in \Theta \tag{4.4}$$

que se verifica sólo si $E_\theta\delta(\boldsymbol{X}) = \theta$. La equivalencia entre (4.1) y (4.2) va más lejos de lo que el argumento anterior deja entrever: (4.1) y (4.2) son equivalentes en condiciones bastante generales (ver Lehmann (1959), p. 22). En lo sucesivo, cuando hablemos de insesgadez en un contexto de estimación de parámetros, nos estaremos refiriendo a estimadores verificando (4.2).

En problemas de contraste de hipótesis, al igual que sucede en problemas de estimación, se define insesgadez mediante una condición estrechamente relacionada con (4.1), cuya discusión abordaremos en el Capítulo 8.

Es importante darse cuenta de que la insesgadez, siendo como es una propiedad intuitivamente atrayente, no es un requerimiento imprescindible, ni necesariamente deseable. En ocasiones, ni siquiera existen procedimientos insesgados. Los siguientes ejemplos ilustran estas ideas.

Ejemplo 4.2 *(un estimador insesgado claramente indeseable)* Sea $\delta = \delta(\boldsymbol{X}))$ un estimador de θ con pérdida cuadrática, $L(\theta, t) = (t - \theta)^2$. Supongamos que la distribución de δ es tal que:

$$\text{Prob}\{\delta = \theta + 100\} = \text{Prob}\{\delta = \theta - 100\} = \frac{1}{2}$$

Tal estimador es insesgado. Sin embargo, siempre tendrá mayor pérdida que otro, δ', acaso sesgado pero verificando:

$$\text{Prob}\{|\delta' - \theta| \leq 5\} = 1$$

En consecuencia, δ es inadmisible.

Existen otros muchos ejemplos de estimadores de Bayes que son sesgados, menos artificialmente simples que el presente. La teoría de Modelos Lineales muestra que, si la pérdida es cuadrática, un estimador sesgado (el estimador *ridge*) puede ser preferible al (insesgado óptimo) proporcionado por mínimos cuadrados ordinarios, y que la mejora derivada de tolerar algún sesgo puede ser notable (en los casos de acusada multicolinealidad). Garthwaite et al. (1995), pág. 35, proporciona un ejemplo alternativo a éste.

Ejemplo 4.3 *(un estimador insesgado puede ocasionalmente dar resultados absurdos)* La insesgadez, cuando el estimando está constreñido a estar en un cierto rango, da lugar a situaciones anómalas. Supongamos que se desea estimar ρ^2 o coeficiente de correlación al cuadrado entre dos variables. Por definición, $0 \leq \rho^2 \leq 1$. Si obligamos a un estimador $\hat{\rho}^2$ a ser insesgado, nos encontraremos con que podemos obtener $\hat{\rho}^2 < 0$ ó sobre otras $\hat{\rho}^2 > 1$. En efecto, el ser insesgado cuando $\rho^2 = 0$ obliga a que eventualmente $\hat{\rho}^2 < 0$ (si siempre fuera $\hat{\rho}^2 \geq 0$, $E[\hat{\rho}^2] > 0$ contra el supuesto de insesgadez). Lo mismo ocurre cuando $\rho^2 = 1$.

Ejemplo 4.4 *(no existencia de estimadores insesgados de una cierta función)* Consideremos una moneda cuya probabilidad de dar cara al ser arrojada es θ. Estamos interesados en estimar no θ, sino la razón de probabilidades cara/cruz, es decir, $\gamma(\theta) = \theta/(1-\theta)$, y contamos con una muestra formada por n observaciones independientes $X_1, \ldots, X_n$. Sea $S(\boldsymbol{X}) = X_1 + \ldots X_n$.

No existe un estimador insesgado. Si lo hubiera, debería verificar:

$$E_\theta \delta(\boldsymbol{X}) = \sum_{\boldsymbol{x} \in \mathcal{X}} \delta(\boldsymbol{x}) \theta^{s(\boldsymbol{x})} (1-\theta)^{n-s(\boldsymbol{x})} = \frac{\theta}{1-\theta} \tag{4.5}$$

en que $s(\boldsymbol{x}) = \sum x_i$ y $\mathcal{X}$ es el conjunto formado por todas las posibles n-tuplas de ceros y unos. Sin embargo, el lado izquierdo de la igualdad anterior es un polinomio de grado finito en θ, en tanto que el lado derecho puede escribirse como $\theta(1 + \theta + \theta^2 + \ldots)$; ningún polinomio puede igualar a la serie de potencias en el lado derecho para cualquier valor de θ.

4.2. Funciones convexas.

Una función $\phi(x)$ real-valorada en el intervalo (a, b) $(-\infty \leq a < b \leq \infty)$ es *convexa* si para cualesquiera x, y, con $a < x < y < b$ y para cualquier $0 < \gamma < 1$ se verifica:

$$\phi(\gamma x + (1-\gamma)y) \leq \gamma\phi(x) + (1-\gamma)\phi(y) \tag{4.6}$$

Decimos que es una función *estrictamente convexa* si la desigualdad en la expresión anterior es estricta. Una función $\phi(x)$ es cóncava en $[a, b]$ si $-\phi(x)$ es convexa en el mismo intervalo. Es inmediato ver que, en el caso de funciones derivables, $\phi'(x)$ monótona no decreciente es condición necesaria y suficiente de convexidad; $\phi''(x) \geq 0$ es condición suficiente pero no necesaria para la convexidad de $\phi(x)$.

Las siguientes propiedades de las funciones convexas, enunciadas como teoremas, serán de utilidad.

Teorema 4.1 *Si $\phi(x)$ es convexa en (a, b) y $t \in (a, b)$, siempre existe una recta de ecuación $y = L(x) = c(x - t) + \phi(t)$ a través de $(t, \phi(t))$ tal que: $L(x) \leq \phi(x)$, $\forall x \in (a, b)$.*

La prueba es sencilla, y resulta innecesaria a la vista de un gráfico. Todo lo que el teorema establece es que para cualquier t en el intervalo de convexidad podemos trazar una tangente[2] a una función convexa que queda siempre por debajo.

Teorema 4.2 *(Desigualdad de Jensen) Si $\phi(x)$ es una función convexa en el intervalo soporte de la v.a. X, y X tiene momento de primer orden finito, se tiene que:*

$$\phi\left(E(X)\right) \leq E\left[\phi(X)\right] \tag{4.7}$$

[2]Estrictamente, podría no ser una tangente en el sentido habitual, y limitarse a tocar a la función convexa en un punto donde ésta es angulosa.

DEMOSTRACION:

Sea $y = L(x)$ la recta aludida en el teorema anterior, con $t = E(X)$. Entonces:

$$\begin{aligned} E\left[\phi(X)\right] &\geq E\left[L(X)\right] \\ &= E\left[c(X-t)\right] + \phi(t) \\ &= \phi(E[X]) \end{aligned}$$

■

Ejemplo 4.5 Una situación en que la desigualdad de Jensen es de aplicación inmediata es aquélla en que el regresando en un modelo lineal es una función cóncava (o convexa) de la variable que resulta de interés predecir. Por ejemplo, podemos tener:

$$Y_i = \log Z_i = \boldsymbol{x}_i{}'\beta + \epsilon_i$$

De acuerdo con el teorema de Gauss-Markov, sabemos entonces que una predicción insesgada y de varianza mínima del valor y_* del regresando es $\boldsymbol{x}'_*\hat{\beta} + \epsilon$. Es decir:

$$E[\boldsymbol{x}_*{}'\hat{\beta}] = y_* \tag{4.8}$$

Sin embargo, la variable que deseamos predecir es $z_* = e^{y_*}$. Como la función exponencial es convexa, de acuerdo con la desigualdad de Jensen se tiene:

$$E[Z_*] \geq e^{E[Y_*]}$$

Si $\boldsymbol{x}'_*\hat{\beta}$ estima insesgadamente el exponente del lado derecho en la expresión anterior, $e^{\boldsymbol{x}'_*\hat{\beta}}$ será un estimador sesgado por defecto de $E[Z_*]$.

Si quisiéramos corregir este sesgo, podríamos quizá linealizar la función logaritmo. En la práctica, el sesgo suele ser de entidad lo suficientemente reducida en comparación con la varianza de la predicción como para no ser considerado.

4.3. Estimación insesgada puntual.

Demostraremos en lo que sigue algunos resultados de gran alcance, que muestran la forma de obtener estimadores insesgados óptimos con funciones de pérdida bastante generales (convexas[3], lo que en particular incluye la estimación mínimo-cuadrática).

[3]La convexidad es una propiedad intuitivamente plausible en una función de pérdida. En esencia supone, en un problema de estimación paramétrica, que la pérdida en que se incurre al estimar un parámetro crece más que proporcionalmente al error cometido en la estimación.

Teorema 4.3 (Rao - Blackwell) *Sea X una v.a. con distribución $\{F_X(x,\theta), \theta \in \Theta\}$, y $S = S(\boldsymbol{X})$ un estadístico suficiente para θ. Sea $\hat{\theta}(\boldsymbol{X})$ un estimador de θ, y $L(\hat{\theta},\theta)$ la función de pérdida, convexa en $\hat{\theta}$. Si $\hat{\theta}(\boldsymbol{X})$ tiene media finita y riesgo:*

$$r_\theta(\hat{\theta}) = E_\theta\left[L(\hat{\theta},\theta)\right] < \infty$$

y definimos:

$$\hat{\eta}(s) = E\left[\hat{\theta}(\boldsymbol{X})|S = s\right]$$

entonces:

$$r_\theta(\hat{\eta}(s)) < r_\theta(\hat{\theta})$$

DEMOSTRACION:

Es una aplicación de la desigualdad de Jensen:

$$\begin{aligned} L(\hat{\eta},\theta) &= L(E_{\boldsymbol{X}|S}\left[\hat{\theta}(\boldsymbol{X})\right],\theta) \\ &= \phi(E_{\boldsymbol{X}|S}\left[\hat{\theta}(\boldsymbol{X})\right]) \\ &\leq E_{\boldsymbol{X}|S}\left[\phi(\hat{\theta}(\boldsymbol{X}))\right] \\ &= E_{\boldsymbol{X}|S}\left[L(\hat{\theta}(\boldsymbol{X}),\theta)\right] \end{aligned}$$

Tomando ahora valor medio respecto de la distribución de S tenemos:

$$E_S\left[L(\hat{\eta},\theta)\right] \leq E_S\left[E_{\boldsymbol{X}|S}\left[L(\hat{\theta}(\boldsymbol{X}),\theta)\right]\right]$$

y como $E_S\left[E_{\boldsymbol{X}|S}\left[\cdot\right]\right] = E_{\boldsymbol{X}}\left[\cdot\right]$ obtenemos en definitiva:

$$\begin{aligned} E_S\left[L(\hat{\eta},\theta)\right] &\leq E\left[L(\hat{\theta}(\boldsymbol{X}),\theta)\right] \\ r_\theta(\hat{\eta}) &\leq r_\theta(\hat{\theta}) \end{aligned}$$

La desigualdad es estricta si la función de pérdida es estrictamente convexa.

■

Observemos, de paso, que, si $\hat{\theta}(\boldsymbol{X})$ es insesgado, la aplicación del teorema de Rao-Blackwell proporciona un $\hat{\eta}(S)$ también insesgado. En efecto:

$$\theta = E_\theta\left[\hat{\theta}(\boldsymbol{X})\right] = E_S\left[E_{\boldsymbol{X}|S}\left[\hat{\theta}(\boldsymbol{X})|S\right]\right] = E_S\left[\hat{\eta}(S)\right]$$

Observación 4.1 ¿Dónde se ha hecho uso de la suficiencia de S? Parece a primera vista que en ninguna parte, y que bastaría condicionar sobre cualquier cosa para que el teorema de Rao-Blackwell surtiera efecto.

Observemos que ello no es así. Si queremos que $\hat{\eta}(S)$ sea un estimador, **no** debe depender del parámetro θ. Si S es suficiente,

$$\hat{\eta}(S) = E_{\boldsymbol{X}|S}\left[\hat{\theta}(\boldsymbol{X})|S\right] = \int \hat{\theta}(\boldsymbol{X}) f_{\boldsymbol{X}|S}(\boldsymbol{x}|s) d\boldsymbol{x}$$

y se verifica esta condición de no dependencia de θ (pues, por definición de suficiencia, $f_{\boldsymbol{X}|S}(\boldsymbol{x}|s)$ no depende de dicho parámetro). No podría afirmarse lo mismo si S no fuera suficiente.

Cuando en un problema de estimación puntual con pérdida convexa se dispone de un estadístico que no sólo es suficiente sino también completo, puede afirmarse la existencia de un estimador único y de riesgo mínimo para cualquier función estimable de θ (es decir, para cualquier $\gamma(\theta)$ para la que exista *alguna* función de la muestra verificando $E_\theta\left[\delta(\boldsymbol{X}))\right] = \gamma(\theta), \forall\theta \in \Theta$). El siguiente teorema proporciona los detalles.

Teorema 4.4 *Sea X una variable aleatoria con distribución $F_X(x;\theta)$, y S un estadístico suficiente para θ en la familia $\{F_X(x;\theta),\ \theta \in \Theta\}$. Entonces, cualquier función estimable $\gamma(\theta)$ posee un estimador insesgado que depende sólo de S. Si S es completo además de suficiente, este estimador es único.*

DEMOSTRACION:

Por hipótesis existe $\delta(\boldsymbol{X})$ tal que $E_\theta\left[\delta(\boldsymbol{X})\right] = \gamma(\theta)$. Condicionando sobre S obtenemos $\hat{\eta}(S)$ que conserva la insesgadez. ¿Podría existir otro estimador insesgado, $\hat{\alpha}(S)$? No. Si lo hubiera, tendríamos (por insesgadez de ambos) que:

$$E_\theta\left[\hat{\eta}(S)\right] = E_\theta\left[\hat{\alpha}(S)\right] \Longrightarrow E_\theta\underbrace{[\hat{\eta}(S) - \hat{\alpha}(S)]}_{g(S)} = 0$$

Pero la condición de completo de S permite entonces concluír que $E_\theta\left[g(S)\right] = 0 \Rightarrow g(S) = 0$ con probabilidad 1, y por tanto $\hat{\eta}(S) = \hat{\alpha}(S)$ (con probabilidad 1). ■

Si a las condiciones del teorema anterior unimos convexidad de la función de pérdida, tenemos el siguiente interesante resultado.

Teorema 4.5 *En las condiciones del Teorema 4.4, si $L(\hat{\theta}(\boldsymbol{X}),\theta)$ es estrictamente convexa y $r_\theta(\hat{\theta})$ es finito, el único estimador insesgado obtenido es uniformemente de mínimo riesgo insesgado. En particular, se trata del estimador insesgado de mínima varianza uniforme*[4].

[4] En ocasiones llamado UMVU (UMVU = Uniformly Minimum Variance Unbiased).

DEMOSTRACION:

En efecto: consideremos $\hat{\eta}(S)$ y cualquier otro posible estimador insesgado $\hat{\theta}(\boldsymbol{X})$. Una aplicación del teorema de Rao-Blackwell a $\hat{\theta}(\boldsymbol{X})$ producirá un $\hat{\alpha}(S)$ mejor que $\hat{\theta}(\boldsymbol{X})$ y que necesariamente coincide con $\hat{\eta}(S)$. Por tanto, éste último es mejor que $\hat{\theta}(\boldsymbol{X})$.

■

Los Teoremas 4.3 y 4.4 muestran dos vías para obtener estimadores insesgados de riesgo mínimo. La primera consistiría en buscar un estadístico suficiente completo S y, a continuación, una función de él que fuera insesgada. El Teorema 4.4 garantiza que este modo de operar conduce al (esencialmente único) estimador insesgado de riesgo mínimo.

El inconveniente de este método es que a veces puede no ser fácil de llevar a cabo la corrección de sesgo aludida, dependiendo del estadístico suficiente que tomemos como punto de partida.

Hay una segunda vía que a menudo permite llegar al mismo resultado de modo más simple. Una vez que hemos encontrado un estadístico suficiente completo S, podemos tomar *cualquier* estimador insesgado $\hat{\theta}$ del parámetro de interés y calcular $E[\hat{\theta}|S]$. El Teorema 4.3 garantiza que el resultado es el estimador insesgado de riesgo mínimo, *sin importar cuál haya sido el estimador insesgado de partida.*

Ejemplo 4.6 Volvamos sobre el Ejemplo 3.8, pág. 36. Vimos allí que $S = \sum_{i=1}^{n} X_i$ (y, equivalentemente, $\overline{X}$) es un estadístico suficiente para λ en la clase de distribuciones de Poisson, $P(\lambda)$. Además, $\overline{X}$ es un estadístico completo.

El Teorema 4.4 (pág. 52) muestra entonces que $\overline{X}$ es el único estimador insesgado de mínima varianza de λ (más generalmente, de mínimo riesgo para cualquier función de pérdida convexa).

Ejemplo 4.7 Consideremos de nuevo el caso de una distribución $U(0, 2\theta)$ y una m.a.s. $X_1, \ldots, X_n$ procedente de ella. Vimos (Ejemplo 3.7, pág. 34) que $X_{(n)}$ es suficiente para θ y además completo (Ejemplo 3.19, pág. 43). Sea $S = X_{(n)}$. Entonces,

$$E_\theta[S] \;=\; \int_0^{2\theta} \frac{n s^{n-1}}{(2\theta)^n} s\,ds = \frac{n}{(2\theta)^n}\left[\frac{s^{n+1}}{n+1}\right]_0^{2\theta} = \frac{2n}{n+1}\theta.$$

Por tanto, $(2n)^{-1}(n+1)X_{(n)}$ es un estimador insesgado de θ que depende sólo del estadístico suficiente $X_{(n)}$. Es insesgado de mínima varianza.

En este caso, ha sido fácil aplicar la primera vía aludida en el texto: buscar una función del estadístico suficiente, calcular su sesgo y corregirlo.

El ejemplo siguiente hace también uso de la primera vía: imponer la insesgadez a una función de un estadístico completo suficiente.

Ejemplo 4.8 *(estimador insesgado de mínima varianza de la varianza de una distribución binaria)* Consideremos una distribución binaria de parámetro p; su varianza es $pq = p(1-p)$. Sea $\hat{p}$ el estimador habitual de p,

$$\hat{p} = n^{-1} \sum_{i=1}^{n} X_i. \tag{4.9}$$

Es fácil ver que $\hat{p}$ es insesgado para p y también suficiente y completo. Sin embargo, el estimador de la varianza $\hat{p}(1-\hat{p})$ no es insesgado. En efecto, en virtud de la desigualdad de Jensen (Sección 4.2, pág. 49),

$$E\left[\hat{p}(1-\hat{p}\right] = E\left[\Phi(\hat{p})\right] \leq \Phi(E(\hat{p})) = p(1-p),$$

dado que $\Phi(.)$ es una función cóncava.

Podemos sin embargo acometer en este caso la corrección directa del sesgo. Sea $T = \sum_{i=1}^{n} X_i$ (completo suficiente) y $\delta(T)$ una función arbitraria de dicho estadístico. Dado que T sigue una distribución binomial, el valor medio de $\delta(T)$ es:

$$E\left[\delta(T)\right] = \sum_{t=0}^{n} \delta(t) \binom{n}{t} p^t (1-p)^{n-t}.$$

Definiendo $\rho = p(1-p)^{-1}$ (por tanto $p = \rho(1+\rho)^{-1}$ y $(1-p) = (1+\rho)^{-1}$),

$$\begin{aligned} E\left[\delta(T)\right] &= \sum_{t=0}^{n} \delta(t) \binom{n}{t} p^t (1-p)^{n-t} \\ &= \sum_{t=0}^{n} \delta(t) \binom{n}{t} \frac{\rho^t}{(1+\rho)^t} \frac{1}{(1+\rho)^{n-t}}. \end{aligned} \tag{4.10}$$

Igualando (4.10) a $p(1-p)$ y simplificando tenemos:

$$\begin{aligned} \sum_{t=0}^{n} \delta(t) \binom{n}{t} \frac{\rho^t}{(1+\rho)^t} \frac{1}{(1+\rho)^{n-t}} &= \frac{\rho}{(1+\rho)^2} \\ \sum_{t=0}^{n} \delta(t) \binom{n}{t} \rho^t &= \rho(1+\rho)^{n-2} \\ \sum_{t=0}^{n} \delta(t) \binom{n}{t} \rho^t &= \rho \left[\binom{n-2}{0} + \binom{n-1}{1} \rho + \ldots + \binom{n-2}{n-2} \rho^{n-2} \right] \\ \sum_{t=0}^{n} \delta(t) \binom{n}{t} \rho^t &= \sum_{t=1}^{n-1} \binom{n-2}{t-1} \rho^t. \end{aligned} \tag{4.11}$$

Igualando términos de igual orden a ambos lados de (4.11) vemos que debe verificarse:

$$\delta(t) \binom{n}{t} = \binom{n-2}{t-1} \Longrightarrow \delta(t) = \frac{t(n-t)}{n(n-1)} \tag{4.12}$$

para $t = 1, \ldots, n-1$ (y $\delta(0) = \delta(n) = 0$, que ya quedan recogidos en la expresión general).

Ejemplo 4.9 Supongamos que la v.a. X sigue una distribución de Poisson y que el parámetro que tenemos interés en estimar es $\theta = e^{-\lambda} = \text{Prob}\{X = 0\}$. Definamos $\hat{\theta}(X)$ así: $\hat{\theta}(X) = 1$ si $X = 0$ y $\hat{\theta}(X) = 0$ en otro caso. Entonces, $\hat{\theta}(X)$ es un estimador insesgado de θ, función de un estadístico completo suficiente, y por tanto uniformemente de mínima varianza, de acuerdo con el Teorema 4.4. Veámoslo.

1. El estadístico X es suficiente; claro, puesto que la totalidad de la muestra es siempre suficiente.
2. El estadístico X es completo en la familia de distribuciones de Poisson $\mathcal{P}(\lambda)$. Comprobémoslo. Sea una función $g(x)$ tal que $E[g(X)] = c$. Ello significaría que:

$$\sum_{j=0}^{\infty} g(j)\frac{e^{-\lambda}\lambda^j}{j!} = c \quad \Longrightarrow \quad \sum_{j=0}^{\infty} [g(j) - c]\,\frac{e^{-\lambda}\lambda^j}{j!} = 0$$

   y por tanto:

$$[g(j) - c] = 0 \quad \forall j \text{ entero} \quad \Longrightarrow \quad g(j) = c \quad \forall j \text{ entero}$$

   En consecuencia, la única función $g(x)$ verificando $E[g(X)] = c$ es la función constante.
3. Finalmente, observemos que:

$$E[\hat{\theta}(X)] = 1 \cdot \frac{e^{-\lambda}\lambda^0}{0!} + 0 \cdot \text{Prob}\{X > 0\} = e^{-\lambda}$$

   luego $\hat{\theta}(X)$ es insesgado.

Este ejemplo o similares han sido objeto de debate en la literatura. El estimador sólo puede proporcionar dos estimaciones: 0 ó 1. Ello es particularmente molesto cuando $\theta = e^{-\lambda}$ no puede alcanzar ninguno de ambos extremos: $0 < \theta < 1$ si $0 < \lambda < \infty$. Junto con los ejemplos 4.2 y 4.4, éste muestra que en algunos casos (en general, bastante anómalos) la elección de un estimador insesgado, incluso de mínima varianza, puede no ser una buena idea.

El siguiente ejemplo, reproducido de Cox y Hinkley (1974), pág. 259, amplía el precedente considerando n observaciones. Ilustra la segunda vía referida más arriba para obtener estimadores insesgados de riesgo mínimo: condicionar *cualquier* estimador insesgado sobre el valor que toma un estadístico completo suficiente.

Ejemplo 4.10 Consideremos la misma situación examinada en el Ejemplo 4.9, pero suponiento ahora que disponemos de una muestra formada por n observaciones independientes, $X_1, \ldots, X_n$. Si deseáramos estimar λ, $\overline{X}$ sería un estimador insesgado. Pero, para estimar $\theta = e^{-\lambda}$, el estimador obvio $e^{-\overline{X}}$ es sesgado (desigualdad de Jensen); y no es inmediato el valor de su sesgo ni la forma de eliminarlo.

Sin embargo, lo cierto es que $\overline{X}$ (o, equivalentemente, $S = X_1 + \ldots + X_n$) es un estadístico completo suficiente (lo que se puede demostrar de modo exactamente análogo al empleado en el Ejemplo 4.9).

Busquemos un estimador insesgado *cualquiera* de $\theta = e^{-\lambda}$; recordando que $\theta = \text{Prob}\{X = 0\}$ vemos que:

$$\hat{\theta}(\boldsymbol{X}) = \begin{cases} 1 & \text{si } X_1 = 0 \\ 0 & \text{en otro caso.} \end{cases}$$

es efectivamente insesgado. Entonces, de acuerdo con el Teorema 4.3 tenemos[5] que:

$$\hat{\theta}^*(S) = E[\hat{\theta}(\boldsymbol{X})|S] = \left(1 - \frac{1}{n}\right)^S \tag{4.13}$$

es el estimador insesgado (esencialmente único) de mínima varianza. ¡A la vista de (4.13) es claro que el indagar directamente qué función de S (o de $\overline{X}$) es insesgada no hubiera tenido grandes posibilidades de éxito!

4.4. El jackknife

En ocasiones puede ser difícil encontrar un estimador insesgado de partida y aplicar el procedimiento de Rao-Blackwell para obtener el estimador insesgado de varianza mínima. Quenouille (1956) propuso un procedimiento para, partiendo de un estimador sesgado, obtener otro insesgado o con sesgo muy reducido respecto al estimador inicial. Es la técnica conocida como *jackknifing*.

Supongamos que el estimador $\hat{\theta}_n$, basado en una muestra de tamaño n, tiene un sesgo de orden $O(n^{-1})$ —como es lo habitual—. Supongamos que

$$E[\hat{\theta}_n] \quad = \quad \theta + \sum_{i=1}^{\infty} \frac{a_i}{n^i}$$

en que los coeficientes a_i pueden depender de θ (pero no de n) y al menos el primero es distinto de cero (de forma que el orden del sesgo es el estipulado). El procedimiento de jackknifing consiste en lo siguiente:

1. Recalcular el estimador n veces, dejando cada vez fuera una observación. Esto proporcionará n versiones del estimador que denotaremos por $\hat{\theta}_{n-1,i}$, $i = 1, \ldots, n$, en que el primer subíndice alude al tamaño de muestra empleado y el segundo a la observación omitida.

2. Computar la media aritmética $\overline{\theta}_{n-1}$ de las n versiones del estimador calculadas en el apartado anterior.

3. Definir el estimador jackknife así:

$$\hat{\theta}_n^J \quad = \quad \hat{\theta}_n + (n-1)(\hat{\theta}_n - \overline{\theta}_{n-1}) \tag{4.14}$$

$$\quad = \quad n\hat{\theta}_n - (n-1)\overline{\theta}_{n-1} \tag{4.15}$$

[5] Condicionalmente sobre S, la distribución de $\boldsymbol{X}$ es multinomial (véase Ejemplo 3.8, pág. 36), y por tanto la distribución de X_1 condicionado por S es binomial de parámetros $\frac{1}{n}, s$.

Es fácil comprobar que el sesgo de $\hat{\theta}_n^J$ es de menor orden que el de $\hat{\theta}_n$. En efecto,

$$E[\hat{\theta}_n^J] = n(\theta + \sum_{i=1}^{\infty} \frac{a_i}{n^i}) - (n-1)\left(\theta + \sum_{i=1}^{\infty} \frac{a_i}{(n-1)^i}\right) \quad (4.16)$$

$$= \frac{-a_2}{n(n-1)} + O(n^{-3}). \quad (4.17)$$

Por consiguiente, el sesgo original que era $O(n^{-1})$ ha quedado reducido a $O(n^{-2})$.

Ejemplo 4.11 *(estimación de θ^2 en una distribución binaria $b(\theta)$)* Si disponemos de una muestra de n observaciones, sabemos que $X = X_1 + \ldots + X_n$ (o, alternativamente, $\hat{\theta}_n = \overline{X} = X/n$) son estadísticos suficientes para θ. Es claro no obstante que, si bien $\hat{\theta}_n$ es insesgado para θ, $\hat{\eta} = \hat{\theta}_n^2 = \overline{X}^2$ es sesgado para $\eta = \theta^2$ (consecuencia inmediata de la desigualdad de Jensen). Veamos cuál es este sesgo y cómo eliminarlo o reducirlo haciendo uso del *jackknife*. Dado que

$$E[\overline{X}^2] = \text{Var}(\overline{X}) + \left[E(\overline{X})\right]^2 = \frac{\theta(1-\theta)}{n} + \theta^2 \quad (4.18)$$

vemos que $\text{Sesgo}(\hat{\eta}) = E[\overline{X}^2] - \theta^2 = n^{-1}\theta(1-\theta)$.

Dejando de lado la observación i-ésima sólo se pueden obtener dos valores para $\hat{\eta}_{n-1,i}$:

$$\hat{\eta}_{n-1,i} = \begin{cases} \left(\frac{x-1}{n-1}\right)^2 & \text{con probabilidad } x/n \\ \left(\frac{x}{n-1}\right)^2 & \text{con probabilidad } \frac{n-x}{n}; \end{cases}$$

por consiguiente, el cálculo del $\overline{\eta}_{n-1}$ puede hacerse directamente sin necesidad de recomputar n veces el estimador y promediar los resultados:

$$\overline{\eta}_{n-1} = \frac{x}{n}\left(\frac{x-1}{n-1}\right)^2 + \frac{n-x}{n}\left(\frac{x}{n-1}\right)^2$$

$$= \frac{(n-2)x^2 + x}{n(n-1)^2}$$

El estimador *jackknife* es por tanto:

$$\hat{\eta}_n^J = n\hat{\eta}_n - (n-1)\overline{\eta}_{n-1} \quad (4.19)$$

$$= n\left(\frac{x}{n}\right)^2 - (n-1)\frac{(n-2)x^2 + x}{n(n-1)^2} \quad (4.20)$$

$$= \frac{x(x-1)}{n(n-1)} \quad (4.21)$$

Puede verificarse con facilidad que, en este caso particular, el jackknife no sólo ha reducido el orden del sesgo, sino que lo ha cancelado en su totalidad. Recordemos que, de acuerdo con (4.18), el sesgo de $\overline{X}^2$ es $n^{-1}\theta(1-\theta)$; por tanto, la remoción del sesgo de orden $O(n)$ supone la remoción de *todo* el sesgo.

CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

4.1 En la situación descrita en el Ejemplo 4.9, obténgase un estimador insesgado de mínima varianza para $\theta = \text{Prob}\{\sum_{i=1}^{n} X_i \leq 1\}$.

4.2 Se cuenta con dos observaciones independientes X_1 y X_2 procedentes de una distribución con densidad

$$f_X(x;\lambda) \quad = \quad \lambda e^{-\lambda x}.$$

Hállese el estimador de mínima varianza insesgado de $\theta = \text{Prob}\{X \geq 1\}$.

4.3 Sea una m.a.s. $X_1, \ldots, X_n$ procedente de una distribución cuya densidad es,

$$\begin{cases} f_X(x,\theta) & \text{para } x \in [a, b(\theta)], \\ 0 & \text{en otro caso.} \end{cases}$$

El parámetro a estimar es θ; a es una constante y $b(\theta)$ una función fija de θ. Compruébese que, si existe un estadístico suficiente, debe ser $X_{(n)}$, y que una condición suficiente para ello es que $f_{\boldsymbol{X}}(\boldsymbol{x},\theta) = g(\boldsymbol{x})h(\theta)$.

(Garthwaite et al. (1995), pág. 37)

4.4 Sea una m.a.s. $X_1, \ldots, X_n$ procedente de una distribución cuya densidad es,

$$f_X(x,\theta) = \begin{cases} \theta^{-1} e^{x/\theta} & \text{si } x \geq 0, \\ 0 & \text{en otro caso.} \end{cases}$$

Indíquese cuáles de los siguientes estimadores de θ: i) $\hat{\theta} = X_1$; ii) $\hat{\theta} = n^{-1}\sum_{i=1}^{n} X_i$; iii) $\hat{\theta} = (n+1)^{-1}\sum_{i=1}^{n} X_i$; iv) $\hat{\theta} = nX_{(1)}$; v) $\hat{\theta} = X_1$; vi) $\hat{\theta} = X_{(n)}$, son: a) Insesgados, b) Función de estadísticos suficientes y c) De mínima varianza insesgados.

4.5 Sean $X_1, \ldots, X_n$, variables aleatorias con densidad común $f_{X|\theta}(x\,|\theta) = \theta x^{\theta-1}$, en que $0 \leq x \leq 1$ y $\theta > 0$.

i) Encuéntrese un estadístico suficiente para θ.

ii) Compruébese que $-\log X_1$ es un estimador insesgado de θ^{-1}.

iii) Haciendo uso del hecho establecido en el apartado anterior, utilícese el teorema de Rao-Blackwell para encontrar el estimador insesgado de mínima varianza de θ^{-1}.

4.6 Sean $X_1, \ldots, X_n$ variables independientes con densidad común $f_X(x|\theta_1,\theta_2)$. Supongamos dos estadísticos T_1 y T_2 tales que T_1 es suficiente para θ_1 cuando θ_2 está dado, y T_2 es suficiente para θ_2 cuando θ_1 está dado. Compruébese que $T = (T_1, T_2)$ es conjuntamente suficiente para (θ_1, θ_2).

Capítulo 5

Eficiencia. La cota de Cramér-Rao.

5.1. Introducción

La teoría que precede, y en particular el Teorema 4.5, muestran el modo de establecer optimalidad de un estimador insesgado.

En lo que sigue, probaremos un resultado de menor alcance: bajo ciertas condiciones de regularidad, si $\hat{\theta}$ es un estimador de θ se verifica

$$\text{Var}_\theta(\hat{\theta}) \geq H(\theta), \tag{5.1}$$

en que $H(\theta)$ es una función que podemos obtener fácilmente. Entonces, si para un estimador $\hat{\theta}$ insesgado de θ tuviéramos

$$\text{Var}_\theta(\hat{\theta}) = H(\theta), \tag{5.2}$$

no existiría ningún otro de varianza menor, y podríamos declarar $\hat{\theta}$ óptimo (en términos de varianza y en la clase de los insesgados, no se olvide).

Este procedimiento es inferior al proporcionado por el Teorema 4.5 por varias razones. En primer lugar, son precisas condiciones de regularidad —básicamente, la función de verosimilitud debe ser lo suficientemente "suave", en un sentido que quedará claro más abajo—. En segundo lugar, (5.1) se refiere sólo a pérdidas cuadráticas. Finalmente, (5.1) no es una desigualdad "ajustada", en el sentido de que puede suceder que, para todo $\hat{\theta}$ insesgado,

$$\text{Var}_\theta(\hat{\theta}) > H(\theta). \tag{5.3}$$

Es decir, el lado derecho es una cota inferior, no necesariamente alcanzable, de la varianza en la estimación insesgada de θ por $\hat{\theta}$.

Sin embargo, la utilización de (5.1) es cómoda en muchas ocasiones, y para su obtención haremos uso de algunos resultados de interés en sí mismos. Son los que se demuestran a continuación.

5.2. Algunos resultados instrumentales

Lema 5.1 *Consideremos la función de verosimilitud, es decir, $f_{\boldsymbol{X}}(\boldsymbol{x};\theta)$ como función de θ, y supongamos que se verifica*

$$\frac{\partial}{\partial\theta}\int f_{\boldsymbol{X}}(\boldsymbol{x};\theta)d\boldsymbol{x} \;=\; \int \frac{\partial}{\partial\theta} f_{\boldsymbol{X}}(\boldsymbol{x};\theta)d\boldsymbol{x}. \tag{5.4}$$

Entonces,

$$E_{\theta_0}\left[\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta}\right]_{\theta=\theta_0} \;=\; 0. \tag{5.5}$$

DEMOSTRACION:

En efecto, observemos que

$$\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{\partial\theta} \;=\; \frac{\left(\frac{\partial}{\partial\theta} f_{\boldsymbol{X}}(\boldsymbol{x};\theta)\right)}{f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}.$$

Por consiguiente,

$$\begin{aligned}
E_{\theta_0}\left[\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta}\right]_{\theta=\theta_0} &= \int f_{\boldsymbol{X}}(\boldsymbol{x};\theta)\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{\partial\theta}d\boldsymbol{x} \\
&= \int f_{\boldsymbol{X}}(\boldsymbol{x};\theta)\frac{\frac{\partial}{\partial\theta} f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}d\boldsymbol{x} \\
&= \int \frac{\partial}{\partial\theta} f_{\boldsymbol{X}}(\boldsymbol{x};\theta)d\boldsymbol{x} \\
&= \frac{\partial}{\partial\theta}\int f_{\boldsymbol{X}}(\boldsymbol{x};\theta)d\boldsymbol{x} \\
&= 0.
\end{aligned}$$

■

Ejemplo 5.1 Ilustramos (5.5) en el caso simple en que $X \sim N(\theta, \sigma^2 = 1)$ y $X_1, \ldots, X_n$ es una muestra aleatoria simple. Entonces,

$$f_{\boldsymbol{X}}(\boldsymbol{X};\theta) \;=\; \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}}\exp\left\{-(X_i-\theta)^2/2\right\}\right),$$

y

$$\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial \theta} = -\sum_{i=1}^{n}(X_i-\theta).$$

Tomando valor medio de esta última expresión comprobamos que se anula:

$$E_\theta\left[\sum_{i=1}^{n}(X_i-\theta)\right] = n\theta - n\theta = 0.$$

Obsérvese que ello es cierto sólo si coinciden los valores del parámetro que se sustrae de cada X_i y el valor del parámetro para el cuál se toma el valor medio.

Observación 5.1 En el Lema 5.1 se ha empleado la notación

$$E_{\theta_0}\left[\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial \theta}\right]_{\theta=\theta_0}$$

para enfatizar el hecho de que se toma el valor medio de la derivada del logaritmo de la verosimilitud *evaluada para el valor θ_0 del parámetro θ*, y que este valor medio lo es con respecto a la densidad $f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)$. Nótese que esto es crítico para que el Lema 5.1 sea válido.

En lo que sigue, para aligerar la notación, θ denota a un tiempo el valor del parámetro y la variable respecto de la que se deriva, sin que esta notación deba inducir a error. Además, salvo expresa mención en contrario, las derivadas respecto a θ se suponen también evaluadas en el valor del parámetro.

Lema 5.2 *Bajo condiciones de regularidad*[1] *se tiene:*

$$\mathrm{Var}_\theta\left(\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial \theta}\right) = E_\theta\left[\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial \theta}\right]^2 \quad (5.6)$$

$$= -E_\theta\left[\frac{\partial^2 \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial \theta^2}\right]. \quad (5.7)$$

DEMOSTRACION:

[1] Que permitan intercambiar el orden de las operaciones de derivación e integración en los casos en que esto se hace en la demostración. Las condiciones de regularidad también incluyen que el recorrido de la distribución no dependa del parámetro θ (como sucedería, por ejemplo, en una $U(0,\theta)$).

Se tiene que:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial\theta}(0) && (5.8)\\
&= \frac{\partial}{\partial\theta} E_\theta\left[\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta}\right] && (5.9)\\
&= \int \frac{\partial}{\partial\theta}\left(\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{\partial\theta} f_{\boldsymbol{X}}(\boldsymbol{x};\theta)\right) d\boldsymbol{x} && (5.10)\\
&= \int \left(f_{\boldsymbol{X}}(\boldsymbol{x};\theta)\frac{\partial^2 \log f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{\partial\theta^2} + \frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{\partial\theta}\frac{\partial f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{\partial\theta}\right) d\boldsymbol{x} && (5.11)\\
&= E_\theta\left[\frac{\partial^2 \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta^2}\right] + \int \left(\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{\partial\theta}\right)^2 f_{\boldsymbol{X}}(\boldsymbol{x};\theta) d\boldsymbol{x} && (5.12)\\
&= E_\theta\left[\frac{\partial^2 \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta^2}\right] + E_\theta\left[\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta}\right]^2. && (5.13)
\end{aligned}
$$

Se ha hecho uso de

$$\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{\partial\theta} = \frac{1}{f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}\frac{\partial f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{\partial\theta}$$

para pasar de (5.10) a (5.11). Del hecho de ser (5.13) igual a cero, se deduce

$$E_\theta\left[\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta}\right]^2 = -E_\theta\left[\frac{\partial^2 \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta^2}\right].$$

■

5.3. Información de Fisher. Cota de Cramér-Rao

Definición 5.1 *Consideremos la variable aleatoria*

$$\frac{\partial \log f_X(X,\theta)}{\partial\theta}.$$

Su varianza se denota por $I_X(\theta)$ y se denomina información de Fisher *asociada a una observación. De acuerdo con el lema anterior:*

$$I_X(\theta) = E_\theta\left[\frac{\partial \log f_X(X,\theta)}{\partial\theta}\right]^2 = -E_\theta\left[\frac{\partial^2 \log f_X(X,\theta)}{\partial\theta^2}\right]$$

Observación 5.2 El nombre de *información* dado a $I_X(\theta)$ encuentra en parte su justificación en el papel que $I_X(\theta)$ juega en la desigualdad de Cramér-Frechet-Rao (Teorema 5.1, pág. 64). Una justificación alternativa, que puede tener cierto atractivo intuitivo, sería la siguiente.

Consideremos una familia de distribuciones, $\{f_X(x,\theta), \theta \in \Theta\}$, y dos miembros de la misma correspondientes a sendos valores del parámetro, θ_0 (el "correcto") y $\theta' = \theta_0 + d\theta$. Pueden proponerse diversas formas de medir la distancia o discrepancia entre $f_X(x,\theta_0)$ y $f_X(x,\theta')$. Una de ellas sería:

$$\ell(\theta_0,\theta') = E_{\theta_0}\left[\log f_X(X,\theta_0) - \log f_X(X,\theta')\right] \tag{5.14}$$

Si suponemos $f_X(x,\theta')$ suficientemente derivable respecto a θ y la sustituimos por su desarrollo en serie de Taylor hasta términos de segundo orden, (5.14) se convierte en:

$$\begin{aligned}
\ell(\theta_0,\theta') &\simeq E_{\theta_0}\left[\log f_X(X,\theta_0) - \log f_X(x,\theta_0) - \left(\frac{\partial \log f_X(X,\theta)}{\partial\theta}\right)_{\theta=\theta_0} d\theta \right.\\
&\qquad \left. -\frac{1}{2}\left(\frac{\partial^2 \log f_X(X,\theta)}{\partial\theta^2}\right)_{\theta=\theta_0} (d\theta)^2\right]\\
&= E_{\theta_0}\left[\frac{1}{2}\frac{\partial^2 \log f_X(X,\theta)}{\partial\theta^2}\right](d\theta)^2\\
&= \frac{1}{2} I_X(\theta_0)(d\theta)^2
\end{aligned}$$

Ello muestra $I_X(\theta)$ como el coeficiente de $(d\theta)^2$ en la medida aproximada de la distancia entre las dos distribuciones. Cuando $I_X(\theta)$ es grande, una alteración de $d\theta$ en el valor del parámetro da lugar a dos distribuciones muy separadas, y cada observación es muy informativa. El caso extremo contrario se presentaría cuando $I_X(\theta)$ fuera cero. Entonces, ambas distribuciones serían (hasta términos de segundo orden) iguales, y las observaciones de X sería nulamente informativas (si los dos valores del parámetro, θ y θ', dan lugar a distribuciones idénticas, el observar los valores que toma X no permite discriminar entre una y otra).

El argumento esbozado no depende de manera crítica de la medida de discrepancia $\ell(\theta,\theta')$ escogida; se llegaría al mismo resultado con otras muchas. Véase al respecto Rao (1965), pág. 271.

Observación 5.3 Vimos en el Ejemplo 1.7, pág. 6, que no era obvio el modo en que debe escogerse una distribución *a priori* no informativa. Una opción muy empleada consiste en emplear la distribución *a priori* no informativa de Jeffreys: véase Jeffreys (1961). Consiste en tomar para una función $\phi = \phi(\theta)$ tal que $I_X(\phi)$ sea constante una distribución *a priori* $\xi(\phi) \propto k$ (quizá impropia, por consiguiente). Ello equivale a tomar sobre el parámetro de interés θ una distribución *a priori* $\xi(\theta) \propto I_X(\theta)^{\frac{1}{2}}$.

Lema 5.3 *La información de Fisher $I_{\boldsymbol{X}}(\theta)$ asociada a una muestra aleatoria simple $\boldsymbol{X}$ formada por n observaciones, es $nI_X(\theta)$.*

DEMOSTRACION:

Si la muestra es aleatoria simple,

$$f_{\boldsymbol{X}}(\boldsymbol{X};\theta) = f_X(X_1,\theta)\cdot\ldots\cdot f_X(X_n,\theta) \tag{5.15}$$

y por consiguiente:

$$\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial \log f_X(X_i,\theta)}{\partial\theta} \tag{5.16}$$

Tomando el valor medio del cuadrado de la expresión anterior, tenemos en el lado izquierdo la información de Fisher correspondiente a la muestra $\boldsymbol{X}$:

$$\begin{aligned} E_\theta\left[\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta}\right]^2 &= \sum_{i=1}^{n} E_\theta\left[\frac{\partial \log f_X(X_i,\theta)}{\partial\theta}\right]^2 \\ &\quad +2\sum_{i=1}^{n}\sum_{j=i+1}^{n} E_\theta\left[\frac{\partial \log f_X(X_i,\theta)}{\partial\theta}\right] E_\theta\left[\frac{\partial \log f_X(X_j,\theta)}{\partial\theta}\right] \\ &= nI_X(\theta) \end{aligned}$$

habida cuenta de que $E_\theta\left[\frac{\partial \log f_X(X_j,\theta)}{\partial\theta}\right] = 0$ (Lema 5.1, pág. 60).

■

Con ayuda de los lemas anteriores podemos ahora fácilmente probar el siguiente teorema.

Teorema 5.1 *Sea $\hat\theta = \hat\theta(\boldsymbol{X})$ un estimador del parámetro θ y $\psi(\theta)$ su valor medio, $\psi(\theta) = E_\theta\left[\hat\theta\right]$. Entonces, bajo condiciones de regularidad,*

$$Var_\theta(\hat\theta) \geq \frac{[\psi'(\theta)]^2}{E_\theta\left[\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{\partial\theta}\right]^2} \tag{5.17}$$

DEMOSTRACION:

$$\begin{aligned} \psi'(\theta) &= \frac{\partial}{\partial\theta} E_\theta\left[\hat\theta(\boldsymbol{X})\right] \\ &= \frac{\partial}{\partial\theta}\int \hat\theta(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x};\theta)d\boldsymbol{x} \\ &= \int \hat\theta(\boldsymbol{x})\frac{\partial}{\partial\theta} f_{\boldsymbol{X}}(\boldsymbol{x};\theta)d\boldsymbol{x} \\ &= \int \hat\theta(\boldsymbol{x})\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{\partial\theta} f_{\boldsymbol{X}}(\boldsymbol{x};\theta)d\boldsymbol{x} \\ &= E_\theta\left[\hat\theta(\boldsymbol{X})\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta}\right] \\ &= E_\theta\left[(\hat\theta(\boldsymbol{X}) - \psi(\theta))\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta}\right] \end{aligned}$$

En el último paso se ha tenido en cuenta (Lema 5.1, pág. 60) que

$$E_\theta\left[\frac{\partial \log f_X(X,\theta)}{\partial\theta}\right] = 0.$$

Elevando al cuadrado la igualdad anterior tenemos:

$$\begin{aligned}
\left[\psi'(\theta)\right]^2 &= \left(E_\theta\left[(\hat\theta(\boldsymbol{X})-\psi(\theta))\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta}\right]\right)^2 && (5.18)\\
&\leq E_\theta\left[(\hat\theta(\boldsymbol{X})-\psi(\theta))^2\right] E_\theta\left[\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta}\right]^2 && (5.19)\\
&= \mathrm{Var}_\theta(\hat\theta)\cdot E_\theta\left[\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta}\right]^2 && (5.20)
\end{aligned}$$

en que el $\leq$ resulta de aplicar la desigualdad de Schwarz a la expresión precedente. Despejando $\mathrm{Var}_\theta(\hat\theta)$ se llega a la tesis del teorema.

■

Observación 5.4 En el caso particular de que $\hat\theta(\boldsymbol{X})$ sea insesgado para cualquier valor de θ, $\psi(\theta)=\theta$, y el numerador de (5.17) es la unidad. Si $\boldsymbol{X}$ es una muestra formada por observaciones independientes, el denominador de (5.17) es, de acuerdo con el Lema 5.3, $nI_X(\theta)$. En el caso de que ambas cosas se verifiquen —estimador $\hat\theta(\boldsymbol{X})$ insesgado y muestra formada por observaciones independientes—, la desigualdad (5.17) adopta por consiguiente la forma:

$$\mathrm{Var}_\theta(\hat\theta) \geq \frac{1}{nI_X(\theta)} \qquad (5.21)$$

Observación 5.5 Por analogía con la definición de información de Fisher sobre θ contenida en $\boldsymbol{X}$, podemos definir *información de Fisher sobre θ contenida en $\hat\theta$* así:

$$I_{\hat\theta}(\theta) = E_\theta\left[\frac{\partial \log f_{\hat\theta}(\hat\theta;\theta)}{\partial\theta}\right]^2$$

Hagamos el cambio de variables $\boldsymbol{X}\to(\boldsymbol{\xi},\hat\theta)$ (siendo $\boldsymbol{\xi}$ variables cualesquiera, que, junto con $\hat\theta$, permiten recuperar $\boldsymbol{X}$; véase Cramér (1960), pág. 548 y siguientes). Entonces:

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = f_{\boldsymbol{\xi}|\hat\theta}(\boldsymbol{\xi}|\hat\theta;\theta) f_{\hat\theta}(\hat\theta;\theta)\left|\frac{\partial(\boldsymbol{\xi},\hat\theta)}{\partial\boldsymbol{x}}\right|$$

y se tiene que:

$$\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta} = \frac{\partial \log f_{\boldsymbol{\xi}|\hat\theta}(\boldsymbol{\xi}|\hat\theta;\theta)}{\partial\theta} + \frac{\partial \log f_{\hat\theta}(\hat\theta;\theta)}{\partial\theta}$$

ya que el jacobiano de la transformación no depende de θ. Elevando al cuadrado y tomando valor medio:

$$\begin{aligned} I_{\boldsymbol{X}}(\theta) &= E_\theta\left[\frac{\partial \log f_{\boldsymbol{\xi}|\hat{\theta}}(\boldsymbol{\xi}|\hat{\theta};\theta)}{\partial\theta}\right]^2 + E_\theta\left[\frac{\partial \log f_{\hat{\theta}}(\hat{\theta};\theta)}{\partial\theta}\right]^2 \\ &\quad +2E_\theta\left[\frac{\partial \log f_{\boldsymbol{\xi}|\hat{\theta}}(\boldsymbol{\xi}|\hat{\theta};\theta)}{\partial\theta}\frac{\partial \log f_{\hat{\theta}}(\hat{\theta};\theta)}{\partial\theta}\right] \\ &= E_\theta\left[\frac{\partial \log f_{\boldsymbol{\xi}|\hat{\theta}}(\boldsymbol{\xi}|\hat{\theta};\theta)}{\partial\theta}\right]^2 + I_{\hat{\theta}}(\theta) \end{aligned} \tag{5.22}$$

ya que:

$$\begin{aligned} &E_\theta\left[\frac{\partial \log f_{\boldsymbol{\xi}|\hat{\theta}}(\boldsymbol{\xi}|\hat{\theta};\theta)}{\partial\theta}\frac{\partial \log f_{\hat{\theta}}(\hat{\theta};\theta)}{\partial\theta}\right] \\ &\quad = E_{\hat{\theta}}\left[E_{\boldsymbol{\xi}|\hat{\theta}}\left(\frac{\partial \log f_{\boldsymbol{\xi}|\hat{\theta}}(\boldsymbol{\xi}|\hat{\theta};\theta)}{\partial\theta}\frac{\partial \log f_{\hat{\theta}}(\hat{\theta};\theta)}{\partial\theta}\right)\right] \\ &\quad = E_{\hat{\theta}}\left(\frac{\partial \log f_{\hat{\theta}}(\hat{\theta};\theta)}{\partial\theta}\left[E_{\boldsymbol{\xi}|\hat{\theta}}\left(\frac{\partial \log f_{\boldsymbol{\xi}|\hat{\theta}}(\boldsymbol{\xi}|\hat{\theta};\theta)}{\partial\theta}\right)\right]\right) \end{aligned}$$

y el término en el corchete es cero (Lema 5.1, pág. 60). De (5.22) se desprende que $I_{\hat{\theta}}(\theta) \leq I_{\boldsymbol{X}}(\theta)$, y que para que se verifique la igualdad es necesario que:

$$E_\theta\left(\frac{\partial \log f_{\boldsymbol{\xi}|\hat{\theta}}(\boldsymbol{\xi}|\hat{\theta};\theta)}{\partial\theta}\right)^2 = 0 \tag{5.23}$$

Ahora bien, (5.23) se verifica siempre que $\hat{\theta}$ es un estadístico suficiente (pues entonces, condicionalmente en $\hat{\theta}$, el "resto" de la muestra $\boldsymbol{\xi}$ tiene distribución independiente de θ).

Observación 5.6 Relacionada con la observación anterior, tenemos la siguiente: si $I_{\hat{\theta}}(\theta) = I_{\boldsymbol{X}}(\theta)$, es decir, si $\hat{\theta}$ es suficiente, la aplicación del Teorema 5.1 a la variable aleatoria $\hat{\theta}$, supuesta insesgada, proporciona:

$$E_\theta(\hat{\theta}-\theta)^2 \geq \frac{1}{I_{\hat{\theta}}(\theta)} = \frac{1}{I_{\boldsymbol{X}}(\theta)} \tag{5.24}$$

La última igualdad está garantizada por la suficiencia, pero ello todavía no implica que el primer término y el último sean iguales. La suficiencia no garantiza que un estimador alcance la cota de Cramér-Rao. Para que ello ocurra es preciso, *además*, que

$$E_\theta(\hat{\theta}-\theta)^2 = \frac{1}{I_{\hat{\theta}}(\theta)}. \tag{5.25}$$

El Problema 5.2 proporciona una condición necesaria y suficiente (bajo condiciones de regularidad) para que ello ocurra.

Examinemos a continuación casos simples en que la cota de Cramér-Rao permite concluir que estamos ante estimadores insesgados de mínima varianza entre los que verifican condiciones de regularidad.

Ejemplo 5.2 Consideremos $X \sim N(\theta, \sigma^2 = 1)$. Vimos en el Ejemplo 5.1, pág. 60, que

$$\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta} = \sum_{i=1}^{n}(X_i - \theta).$$

Tomando valor medio en dicha expresión,

$$I_{\boldsymbol{X}}(\theta) = E_\theta\left[\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial\theta}\right]^2 = E_\theta\left[\sum_{i=1}^{n}(X_i-\theta)\right]^2 = n\sigma^2 = n.$$

Por consiguiente, la varianza de cualquier estimador insesgado regular está acotada inferiormente por $I_{\boldsymbol{X}}(\theta)^{-1} = n^{-1}$. Como quiera que $\text{Var}(\overline{X}) = n^{-1}$ e insesgado, tenemos que $\overline{X}$ es insesgado de mínima varianza regular.

Nótese que al mismo resultado se puede llegar a partir del teorema de Rao-Blackwell sin requerir condiciones de regularidad: basta notar el carácter de insesgado de $\overline{X}$ y que es función de un estadístico completo suficiente.

Ejemplo 5.3 *(cota de Cramér-Rao para el parámetro de una Poisson)* Sea $X \sim P_X(x;\lambda) = e^{-\lambda}\lambda^x(x!)^{-1}$. Entonces,

$$\begin{aligned}
\frac{\partial \log P_X(X;\lambda)}{\partial\lambda} &= -1 + \lambda^{-1}X \\
E_\lambda\left(\frac{\partial \log P_X(X;\lambda)}{\partial\lambda}\right)^2 &= E_\lambda\left(X\lambda^{-1}-1\right)^2 \\
&= E_\lambda\left(\frac{X-\lambda}{\lambda}\right)^2 \\
&= \lambda^{-1}.
\end{aligned}$$

Por consiguiente, $I_X(\lambda) = \lambda^{-1}$ y la cota de Cramér-Rao para cualquier estimador $\hat{\lambda}$ basado en n observaciones independientes es

$$\text{Var}(\hat{\lambda}) \geq \frac{1}{n\lambda^{-1}} = \frac{\lambda}{n}.$$

Como quiera que $\overline{X}$ tiene varianza precisamente λ/n, concluimos que es estimador insesgado de mínima varianza.

5.4. Eficiencia

En relación con la Observación 5.4, tenemos la siguiente definición.

Definición 5.2 *Se llama* eficiencia *(o, a veces,* eficiencia de Bahadur*) de un estimador insesgado al cociente*

$$\frac{1/I_{\boldsymbol{X}}(\theta)}{\mathit{Var}(\hat{\theta})}$$

Un estimador que alcance la cota de Cramér-Rao tiene pues eficiencia 1; se dice que es eficiente.

Es preciso notar que la eficiencia así definida no implica optimalidad en un sentido demasiado amplio, y, de hecho, es quizá un nombre no muy afortunado. En efecto, un estimador eficiente es mejor sólo:

- En la clase de estimadores regulares insesgados.
- Si adoptamos como función de pérdida una mínimo cuadrática (recuérdese que, en cambio, el Teorema 4.3 proporcionaba estimadores insesgados óptimos para cualquier función de pérdida convexa y sin supuestos de regularidad).

Es también interesante señalar que la noción de eficiencia surge de la comparación de la varianza de un estimador insesgado con un *óptimo optimorum* (en la clase de los estimadores regulares insesgados) que no tiene porqué ser alcanzable. Puede así darse el caso de que un estimador sea ineficiente de acuerdo con la definición anterior, y sin embargo no exista ninguno mejor en la clase de los insesgados. El siguiente ejemplo lo pone de manifiesto.

Ejemplo 5.4 *(un estimador insesgado de varianza mínima que, sin embargo, no alcanza la cota de Cramér-Rao para estimadores insesgados)* Como ejemplo de situación descrita en la observación anterior, puede tomarse el siguiente (ver Romano y Siegel (1986), ejemplo 9.4). Consideremos de nuevo el Ejemplo 4.9 (pág. 55), que a su vez hacía referencia al Ejemplo 3.8 (pág. 36). Nos planteábamos allí el problema de estimar insesgadamente el $\theta = e^{-\lambda} = \text{Prob}\{X = 0\}$ en una distribución de Poisson $\mathcal{P}(\lambda)$. Si sólo se dispone de una observación, el estimador:

$$\hat{\theta} = \begin{cases} 1 & \text{si } X = 0 \\ 0 & \text{en otro caso} \end{cases}$$

vimos que era insesgado y de varianza mínima. Esta varianza es la de una binaria de parámetro $\theta = e^{-\lambda}$, es decir, $e^{-\lambda}(1 - e^{-\lambda})$. En términos de θ, la función de cuantía de X es:

$$P_X(x;\theta) = \frac{\theta(-\log\theta)^x}{x!}$$

y el cálculo de la cota de Cramér-Rao es simple:

$$\begin{aligned} \frac{\partial \log P_X(X;\theta)}{\partial\theta} &= \frac{1}{\theta} + X\frac{(-\log\theta)'}{(-\log\theta)} \\ &= \frac{1}{\theta} + X\frac{(-1/\theta)}{-\log\theta} \\ &= \frac{1}{\theta}\left(\frac{\log\theta + X}{\log\theta}\right) \end{aligned}$$

Por tanto:

$$\begin{aligned} I_X(\theta) &= E\left(\frac{\partial \log P_X(X;\theta)}{\partial\theta}\right)^2 \\ &= \frac{1}{\theta^2}E\left(\frac{X-\lambda}{-\lambda}\right)^2 \\ &= \frac{1}{\theta^2\lambda^2}E(X-\lambda)^2 \\ &= \frac{1}{\theta^2\lambda} \end{aligned}$$

y en consecuencia, la varianza de un estimador insesgado $\hat{\theta}$ haciendo uso de una única observación es:

$$E(\hat{\theta}-\theta)^2 \geq \frac{\lambda\theta^2}{1} = \frac{\lambda e^{-2\lambda}}{1}$$

Fácilmente se comprueba que $e^{-\lambda}(1-e^{-\lambda}) > \lambda e^{-2\lambda}$ (viendo que las funciones a ambos lados de la desigualdad toman el valor 0 cuando $\lambda = 0$ y que la derivada del lado izquierdo es mayor que la del lado derecho). La cota de Cramér-Rao no es por tanto alcanzable en este caso por ningún estimador insesgado.

En el mismo espíritu que la Definición 5.2 tenemos la siguiente.

Definición 5.3 *Se llama* eficiencia relativa *de un estimador $\hat{\theta}_1$ respecto a otro $\hat{\theta}_2$ al cociente*

$$\frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}.$$

Las eficiencias, relativas o no, pueden variar con el tamaño muestral, por lo que en ocasiones se recurre a especificarlas para muestras "muy grandes". Ello da lugar a las nociones de *eficiencia asintótica* y *eficiencia asintótica relativa*, que encontraremos en la Sección 6.5.

Ejemplo 5.5 *(eficiencia relativa de varios estimadores de θ en una distribución $U(0,2\theta)$)* Consideremos de nuevo el caso de una distribución uniforme $U(0,2\theta)$. Dada una m.a.s. $X_1,\ldots,X_n$ procedente de esta distribución hemos visto que $X_{(n)}$ es suficiente (Ejemplo 3.7, pág. 34), completo (Ejemplo 3.19, pág. 43) y puede dar lugar, mediante la oportuna corrección de su sesgo, a un estimador insesgado de mínima varianza de θ,

$$\hat{\theta} = \frac{n+1}{2n}X_{(n)},$$

(Ejemplo 4.7, pág. 53). Examinemos ahora la eficiencia relativa de $\hat{\theta}_1 = \overline{X}$. Ambos estimadores $\hat{\theta}$ y $\hat{\theta}_1$ son insesgados. La varianza del segundo es

$$\text{Var}(\hat{\theta}_1) = n^{-2}\sum_{i=1}^{n}\text{Var}(X_i) = n^{-2}\sum_{i=1}^{n}\frac{(2\theta-0)^2}{12} = \frac{\theta^2}{3n}.$$

La varianza de $\hat{\theta}$ se calcula también con facilidad. Tenemos

$$E\left[\hat{\theta}^2\right] = \left(\frac{n+1}{2n}\right)^2 \int_0^{2\theta} \frac{n\hat{\theta}^{n+1}}{(2\theta)^n} d\hat{\theta} = (n+1)^2(n+2)^{-1}\theta^2;$$

la varianza de $\hat{\theta}$ es por tanto

$$\text{Var}(\hat{\theta}) = (n+1)^2(n+2)^{-1}\theta^2 - \theta^2 = \frac{\theta^2}{n(n+2)}.$$

Comparando, vemos que el estimador $\hat{\theta}$ tiene varianza igual (cuando $n = 1$) ó menor, y tanto menor cuanto mayor es n. De hecho, la varianza de $\hat{\theta}$ tiende a cero con orden $O(n^{-2})$, mientras que la de $\hat{\theta}_1$ tiende a cero linealmente.

La eficiencia relativa de $\hat{\theta}_1$ respecto de $\hat{\theta}$ es

$$\text{Ef.rel.}(\hat{\theta}_1; \hat{\theta}) = \frac{n^{-1}(n+2)^{-1}\theta^2}{(3n)^{-1}\theta^2} = \frac{3}{n+2}.$$

Ejemplo 5.6 *(cuando fallan las condiciones de regularidad, la varianza de un estimador puede descender por debajo de la cota de Cramér-Rao)* En el Ejercicio 5.5 se ha calculado la varianza del estimador insesgado de mínima varianza. Podemos ahora comprobar que dicha varianza es inferior a la cota que resultaría de una aplicación mecánica (e incorrecta) de la cota de Cramér-Rao.

En efecto:

$$f_{\boldsymbol{X}}(\boldsymbol{X}; \theta) = \begin{cases} (2\theta)^{-1} & \text{si } 2\theta > X_{(n)}, \\ 0 & \text{en otro caso,} \end{cases}$$

$$\frac{\partial \log f_X(X, \theta)}{\partial \theta} = \begin{cases} -1/\theta & \text{si } 2\theta > X_{(n)}, \\ 0 & \text{en otro caso.} \end{cases}$$

Hay que señalar que la derivada no existe en el punto anguloso $\theta = X_{(n)}$. Si ahora calculamos la "información de Fisher", obtenemos:

$$I_X(\theta) = \int_0^{2\theta} \left(-\frac{1}{\theta}\right)^2 \frac{1}{2\theta} dx = \frac{1}{\theta^2}.$$

Por consiguiente, la "cota de Cramér-Rao" daría

$$\text{Var}(\hat{\theta}) \geq \frac{1}{n\theta^{-2}} = \frac{\theta^2}{n},$$

mientras que en el Ejemplo 5.5 hemos comprobado que el estimador insesgado óptimo tiene varianza $\theta^2 n^{-1}(n+2)^{-1}$.

La razón por la que la desigualdad de Cramér-Rao no es de aplicación aquí, es que fallan las condiciones de regularidad. En efecto,

$$\frac{\partial}{\partial \theta} \int f_X(x, \theta) dx = 0,$$

mientras que

$$\int \frac{\partial}{\partial \theta} f_X(x, \theta) dx = \int \frac{\partial}{\partial \theta} \frac{1}{\theta} dx = \int -\frac{1}{\theta^2} dx \neq 0.$$

CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

5.1 Demuéstrese que la expresión (5.14), pág. 63, propuesta como distancia entre $f_X(x, \theta_0)$ y $f_X(x, \theta')$ toma valores no negativos, y es cero si y sólo si $f_X(x, \theta_0)$ y $f_X(x, \theta')$ son iguales, salvo acaso en un conjunto de puntos con probabilidad cero.

5.2 Para que la desigualdad de Schwarz

$$[E(XY)]^2 \quad \leq \quad E[X^2]E[Y^2]$$

se verifique, es condición suficiente que $X \propto Y$, salvo en un conjunto de puntos con probabilidad cero. Haciendo uso de este hecho y observando el uso que de la desigualdad de Schwarz se ha hecho en la ecuación (5.19), demuéstrese que para que un estimador insesgado regular $\hat{\theta}$ alcance la cota de Cramer-Rao es precisa, además de la suficiencia, que

$$(\hat{\theta} - \theta) \propto \frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{x}; \theta)}{\partial \theta}.$$

(Garthwaite et al. (1995), pág. 14)

5.3 Completando el problema anterior, verifíquese que bajo condiciones de regularidad, un estimador insesgado alcanza la cota de Cramér-Rao si, y sólo si,

$$(\hat{\theta} - \theta) = I_{\boldsymbol{X}}(\theta)^{-1} \frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{x}; \theta)}{\partial \theta}.$$

5.4 Sea una distribución de Poisson con función de cuantía $P_X(x; \lambda)$. Haciendo uso de que $E[X(X-1)] = \lambda^2$, obténgase:

1. El mejor estimador insesgado de λ^2 basado en una única observación X.
2. El mejor estimador insesgado de λ^2 basado en n observaciones.

5.5 Sea $X_1, \ldots, X_n$ una m.a.s. procedente de una distribución $N(\mu, \Sigma^2)$. Compruébese que $S^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \overline{N})^2$ no alcanza la cota de Cramér-Rao, pero la diferencia entre su varianza y dicha cota tiende hacia cero cuando $n \to \infty$.

Capítulo 6

Máxima verosimilitud

6.1. La lógica máximo verosímil

En (Trocóniz, 1987, pág. 214) se propone el siguiente ejemplo:

> "Supongamos que se dispone de tres urnas simbolizadas por
>
> $$\begin{aligned} U_4 &= \begin{cases} 4 & \text{bolas blancas} \\ 96 & \text{bolas negras} \end{cases} \\ U_{50} &= \begin{cases} 50 & \text{bolas blancas} \\ 50 & \text{bolas negras} \end{cases} \\ U_{99} &= \begin{cases} 99 & \text{bolas blancas} \\ 1 & \text{bolas negras} \end{cases} \end{aligned}$$
>
> y que nos presentan una muestra de cuatro bolas tomadas de una de las urnas U_4, U_{50} ó U_{99}; las cuatro bolas resultaron ser blancas.
>
> Con cierta lógica, si debiéramos emitir un juicio sobre la urna de procedencia nos inclinaríamos por U_{99}, pues es grande la probabilidad de que esta urna proporcione una muestra de cuatro bolas blancas, y pequeña la probabilidad en las urnas U_4 y U_{50}. [...] La lógica que contiene esta forma de decidir es la lógica de la máxima verosimilitud."

Es lo cierto que difícilmente alguien podría, confrontado con el mismo problema, resolver de diferente modo. Ello dice mucho de la fuerte base intuitiva que subyace a la lógica de la máxima verosimilitud.

Examinemos algunas cuestiones de interés, y tratemos de racionalizar el comportamiento que parece tan intuitivamente correcto. En primer lugar, podemos pensar en las urnas como "estados de la Naturaleza" que generan observables. Ello nos devuelve al marco de la teoría esbozada en capítulos anteriores.

Si las bolas sacadas hubieran sido cinco, y las cinco blancas, ello haría de inmediato descartable la urna U_4. No podemos considerar un estado de la Naturaleza como plausible si es incapaz de generar la evidencia que hemos observado. Observemos que la lógica máximo verosímil va un paso más allá, y permite manejar casos en que la conclusión no puede alcanzarse con absoluta certeza. No es *imposible* que la urna U_4 genere cuatro bolas blancas en un muestreo, pero si *muy raro*; y por lo tanto adoptamos como estado de la Naturaleza otro (en el ejemplo propuesto, U_{99}) que genera la evidencia observada con mayor facilidad. Podemos pues ver la lógica máximo verosímil como una extensión de la lógica ordinaria que nos obliga a excluir hipótesis o explicaciones que no dan cuenta de lo observado.

Observemos también que, en un sentido vago e impreciso, que será perfilado en el Capítulo 9, la lógica máximo verosímil conduce a escoger el estado de la Naturaleza o hipótesis explicativa menos "compleja." El razonamiento subyacente al enfrentarnos al ejemplo de las tres urnas es: "¿Por qué habríamos de aceptar que la urna generadora de las cuatro bolas blancas es U_4 —que sólo rarísimamente genera cuatro bolas blancas— cuando la urna U_{99} genera el mismo observable con gran frecuencia? ¿Por qué admitir que ha ocurrido algo muy raro cuando hay una explicación alternativa que lo hace frecuente?"

En otras palabras, lo que hacemos es escalafonar los posibles estados de la Naturaleza, considerando más "complejos" (y por ello menos deseables) a aquéllos que más raramente generan evidencia como la observada. Veremos (en el Capítulo 9) que esta intuición se puede precisar considerablemente en una noción de complejidad.

En parte por su atrayente contenido intuitivo y en parte por las buenas propiedades asintóticas de que disfruta, el método de estimación máximo verosímil alcanzó enseguida una enorme popularidad. En lo que sigue se examinan las propiedades asintóticas del estimador, destacando que las mismas no siempre se trasladan a pequeñas muestras, donde el estimador MV puede ser marcadamente ineficiente.

6.2. Verosimilitud y estimación máximo verosímil.

Sea $f_{\boldsymbol{X}}(\boldsymbol{X};\theta)$ la función de densidad conjunta de una muestra $\boldsymbol{X} = X_1, \ldots, X_n$. Si consideramos fija la muestra en los valores observados, tenemos una función $f_{\boldsymbol{X}}(\boldsymbol{x};\theta)$ de θ llamada función de verosimilitud. Proporciona la densidad (o cuantía en el caso de variables aleatorias discretas) que correspondería a la muestra fija considerada bajo cada posible valor de θ.

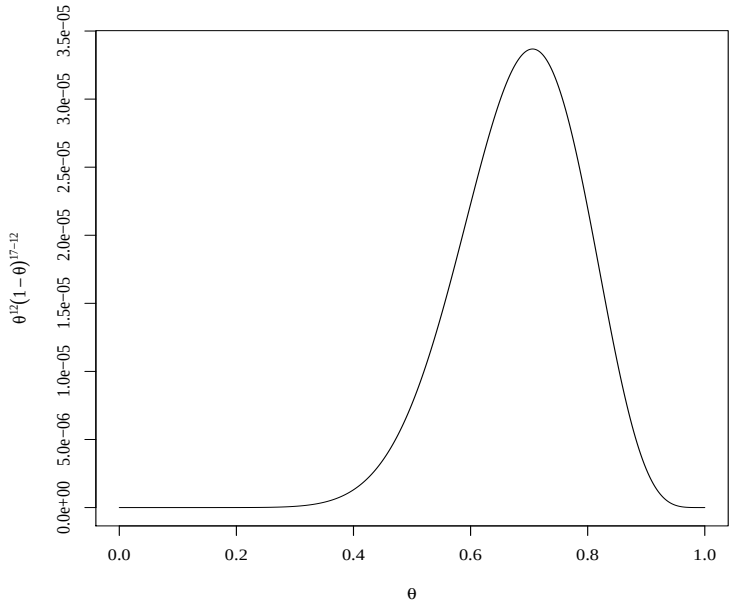
Ejemplo 6.1 Sea una muestra aleatoria simple $(X_1, \ldots, X_n)$ procedente de una distribución $N(\theta, \sigma_0^2)$, de la que se conoce la varianza σ_0^2. Fija-

dos en el muestreo los n valores $(x_1, \ldots, x_n)$, la verosimilitud es:

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \left(\frac{1}{\sigma_0\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\theta)^2\right\} \tag{6.1}$$

Como función de θ, es una distribución normal con varianza σ_0^2 centrada sobre $\overline{x}$.

Figura 6.1: Verosimilitud asociada a una muestra $(x_1, \ldots, x_{17})$, cuando X es binaria de parámetro θ y $\sum_{i=1}^{17} x_i = 12$.



Ejemplo 6.2 Sea una muestra aleatoria simple $(X_1, \ldots, X_n)$ procedente de una distribución binaria de parámetro θ. Sea $s = x_1 + \ldots + x_n$. La función de cuantía conjunta es:

$$P_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \binom{n}{s}\theta^s(1-\theta)^{n-s} \tag{6.2}$$

Como función de θ, su forma es la que muestra la Figura 6.1. El máximo se alcanza sobre s/n (que en el caso representado en la Figura 6.1 es $12/17$).

Definición 6.1 *Llamamos estimador máximo verosímil $\hat{\theta}_{\text{MV}}$ del parámetro θ en la familia de distribuciones $\{f_{X|\theta}(x\,|\theta), \theta \in \Theta\}$ a*

$$\hat{\theta}_{\text{MV}} \overset{def}{=} \arg\max_{\theta} f_{\boldsymbol{X}\,|\theta}(\boldsymbol{x}\,|\theta).$$

Puede ocurrir que $\hat{\theta}_{\text{MV}}$ no esté unívocamente definido. Cuando necesitemos enfatizar la dependencia de $\hat{\theta}_{\text{MV}}$ del tamaño muestral escribiremos $\hat{\theta}_{\text{MV},n}$.

Se sigue inmediatamente de la Definición 6.1 que si $\hat{\theta}_{\text{MV}}$ es el estimador máximo verosímil de θ y $g(.)$ es cualquier función 1-1 de θ, entonces $g(\hat{\theta}_{\text{MV}})$ es el estimador máximo verosímil de $g(\theta)$ (Ejercicio 6.3).

Observación 6.1 Es de interés comprobar que, como cabe esperar de cualquier estimador "sensato", si hay un estadístico suficiente $S = S(\boldsymbol{X})$ para θ y $\hat{\theta}_{\text{MV}}$ es único, entonces $\hat{\theta}_{\text{MV}} = \ell(S)$. En efecto, como consecuencia del teorema de factorización (Teorema 3.3, pág. 37),

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = g_S(s,\theta)h(\boldsymbol{x})$$

Como función de θ, dada $\boldsymbol{x}$, $f_{\boldsymbol{X}}(\boldsymbol{x};\theta)$ tiene un perfil idéntico al de $g_S(s,\theta)$; $h(\boldsymbol{x})$ es un mero factor de escala. Por tanto, $f_{\boldsymbol{X}}(\boldsymbol{x};\theta)$ alcanza su máximo dondequiera que $g_S(s,\theta)$ alcance el suyo. Este último depende de $\boldsymbol{x}$ sólo a través de s, y por tanto, $\hat{\theta}_{\text{MV}}$ ha de ser función de s solamente. Si $\hat{\theta}_{\text{MV}}$ no es único, cabría imaginar un estimador máximo verosímil que no dependería de la muestra sólo a través de s: véase Romano y Siegel (1986), Ejemplo 8.13, o Levy (1985).

Observación 6.2 Relacionada con la anterior observación está la siguiente: si hay un estadístico suficiente y el estimador máximo verosímil es único, entonces éste no puede ser mejorado con ayuda del método de Rao-Blackwell. En efecto: de acuerdo con la observación precedente, el estimador MV en este caso sería una función unívoca del estadístico suficiente, y el condicionar sobre el valor que toma éste nos daría de nuevo el estimador MV.

Observación 6.3 En ocasiones se dice que "el estimador máximo verosímil extrae cuanta información hay en la muestra", lo que sugiere una especie de suficiencia automática. Esto es frecuentemente, pero no necesariamente, cierto. Es cierto cuando el estimador MV es función $1-1$ de un estadístico suficiente (en cuyo caso es suficiente; véase Sección 3.2). Pero éste no tiene porqué ser necesariamente el caso. Consideremos de nuevo el ejemplo propuesto en la Observación 3.2, pág. 41. La verosimilitud era

$$f_{\boldsymbol{T},\boldsymbol{U}}(\boldsymbol{t},\boldsymbol{u}) = \beta^d \exp\left\{-\beta\left(\sum_{i=1}^{d} t_i + \sum_{j=d+1}^{N} u_j\right)\right\}.$$

Se puede comprobar que el estimador máximo verosímil es

$$\hat{\beta}_{\text{MV}} = \frac{d}{\sum_{i=1}^{d} t_i + \sum_{j=d+1}^{N} u_j}.$$

Fácilmente se ve que $\hat{\beta}_{\text{MV}}$ no es suficiente; un mismo valor de $\hat{\beta}_{\text{MV}}$ es compatible con multitud de valores del estadístico (2-dimensional) suficiente $(d, (\sum t_i + \sum u_j))$.

6.3. Consistencia fuerte del estimador máximo verosímil.

Decimos que un estimador $\hat{\theta}_n$ basado en una muestra de tamaño n es consistente para el parámetro θ si: $\hat{\theta}_n \xrightarrow{p} \theta$. Decimos que es *fuertemente consistente* si la convergencia anterior es casi segura: $\hat{\theta}_n \xrightarrow{c.s.} \theta$.

El lema a continuación hace uso de la desigualdad de Jensen para establecer un resultado instrumental.

Lema 6.1 *Supongamos que $f_X(x;\theta_*) = f_X(x;\theta_0)$ (salvo acaso sobre un conjunto de medida nula) sólo cuando $\theta_* = \theta_0$. Sea θ_0 el verdadero valor del parámetro θ. Entonces,*

$$E_{\theta_0}\left[\log \frac{f_X(X;\theta_*)}{f_X(X;\theta_0)}\right] \quad < \quad \log E_{\theta_0}\left[\frac{f_X(X;\theta_*)}{f_X(X;\theta_0)}\right] = 0. \tag{6.3}$$

DEMOSTRACION:

Como $\log()$ es una función estrictamente cóncava, la desigualdad es consecuencia directa de la de Jensen. La nulidad del lado derecho es también fácil de establecer. En efecto,

$$\begin{aligned} \log E_{\theta_0}\left[\frac{f_X(X;\theta_*)}{f_X(X;\theta_0)}\right] &= \log \int f_X(x;\theta_0)\frac{f_X(x;\theta_*)}{f_X(x;\theta_0)}dx \\ &= \log \int f_X(x;\theta_*)dx \\ &= \log(1) = 0; \end{aligned}$$

si la distribución fuera discreta, las integrales se convertirían en sumatorios.

■

Teorema 6.1 *En las condiciones bajo las que se verifica el Lema anterior, $\hat{\theta}_{\text{MV}} \xrightarrow{c.s.} \theta_0$.*

DEMOSTRACION:

Como

$$E_{\theta_0}\left[\log \frac{f_X(X;\theta_*)}{f_X(X;\theta_0)}\right] = c < 0$$

según el Lema anterior, en virtud de la ley fuerte de grandes números (A.3) tenemos que para todo $\theta_* \neq \theta_0$,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\left[\log \frac{f_X(X_i;\theta_*)}{f_X(X_i;\theta_0)}\right] \xrightarrow{c.s.} c \quad &< \quad 0 \\ \text{Prob}\left\{\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\log\left[\frac{f_X(X_i;\theta_*)}{f_X(X_i;\theta_0)}\right] < 0\right\} \quad &= \quad 1 \\ \text{Prob}\left\{\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\log f_X(X_i;\theta_*) < \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\log f_X(X_i;\theta_0)\right\} \quad &= \quad 1 \end{aligned}$$

Sin embargo, de acuerdo con la definición de $\hat{\theta}_{\text{MV}}$, ha de suceder:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log f_X(X_i, \hat{\theta}_{\text{MV},n}) \quad \geq \quad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log f_X(X_i; \theta_0)$$

Las dos desigualdades anteriores sólo pueden reconciliarse si $\hat{\theta}_{\text{MV},n} \xrightarrow{c.s.} \theta_0$, lo que prueba la consistencia fuerte del estimador MV.

■

6.4. Información de Kullback-Leibler y estimación máximo verosímil

Hay una relación interesante entre la estimación máximo verosímil y la información de Kullback-Leibler. La ilustraremos mediante un caso muy simple.

Supongamos que $\Theta = \{\theta_0, \theta_1\}$, y que la variable aleatoria X se distribuye según $F_X(x; \theta_i)$, $i = 0$ ó 1. Llamamos *información en una observación* X para discriminar entre θ_0 y θ_1 a:

$$-\log \left[\frac{f_X(X; \theta_1)}{f_X(X; \theta_0)} \right] \tag{6.4}$$

Observemos que si $X = x$ tuviera exactamente la misma densidad bajo θ_0 que bajo θ_1, la observación en cuestión carecería de información a efectos de discriminar entre ambos estados de la naturaleza, y (6.4) sería cero. El caso opuesto se presenta cuando la densidad bajo un estado y otro es muy diferente: en este caso, la observación podría considerarse como muy informativa acerca del estado de la naturaleza, y (6.4) sería grande en valor absoluto.

Una medida razonable de la "separación" entre $F_X(x; \theta_0)$ y $F_X(x; \theta_1)$ podría ser la información *media* que proporciona una observación:

$$d(\theta_0, \theta_1) \quad = \quad -\int f_X(x; \theta_0) \log \left[\frac{f_X(x; \theta_1)}{f_X(x; \theta_0)} \right] dx \tag{6.5}$$

o, en el caso de variables discretas:

$$d(\theta_0, \theta_1) \quad = \quad -\sum P_X(x, \theta_0) \log \left[\frac{P_X(x, \theta_1)}{P_X(x, \theta_0)} \right] \tag{6.6}$$

Llamamos a (6.5)-(6.6) *información de Kullback-Leibler para la discriminación entre θ_0 y θ_1 contenida en una observación*. De nuevo, obsérvese que se trata de una definición intuitivamente plausible. En particular, si $f_X(x; \theta_0) = f_X(x; \theta_1)$ para todo valor x tendríamos que $d(\theta_0, \theta_1) = 0$, y sería imposible discriminar.

Observación 6.4 La información de Kullback-Leibler esta relacionada con la de Fisher, que puede verse como una aproximación de segundo orden: véase la Observación 5.2, pág. 62.

Observación 6.5 La expresión (6.5) toma valor no negativo (mismo argumento que el empleado en el Lema 6.1) y puede verse por ello como una medida de separación o distancia. No es sin embargo simétrica en sus argumentos, a diferencia de una distancia.

Es interesante ver el problema de estimación máximo verosímil como un problema de selección de una distribución en una familia paramétrica, $\{F_X(x;\theta), \theta \in \Theta\}$. Razonemos sobre el caso en que X es una variable aleatoria discreta.

La muestra $(x_1, \ldots, x_n)$ puede verse como generando una *distribución empírica* $F_X^*(x)$, que atribuye probabilidad $1/n$ a cada uno de los valores muestrales observados (ó k/n a aquéllos que se han repetido k veces). Es decir,

$$F_X^*(x) \quad = \quad \frac{(\text{Total observaciones } \leq x)}{n}.$$

De aquí podemos obtener

$$P_X^*(x) \quad = \quad F_X^*(x) - F_X^*(x^-).$$

Podríamos pensar en estimar θ seleccionando en la clase paramétrica $\{F_X(x;\theta), \theta \in \Theta\}$ aquella distribución que minimiza la distancia de Kullback-Leibler a la distribución empírica observada, es decir, que minimiza:

$$\begin{aligned} -\sum_{i=1}^{n} P_X^*(x_i) \log \frac{P_X(x_i;\theta)}{P_X^*(x_i)} &= \sum_{i=1}^{n} P_X^*(x_i) \log \frac{P_X^*(x_i)}{P_X(x_i;\theta)} \\ &= \sum_{i=1}^{n} \frac{1}{n} \log \frac{1/n}{P_X(x_i;\theta)} \\ &= \frac{1}{n}\sum_{i=1}^{n} \log \frac{1}{n} - \frac{1}{n}\sum_{i=1}^{n} \log P_X(x_i;\theta) \end{aligned}$$

Como el primer sumando del lado derecho es constante, la minimización de la expresión anterior llevaría a hacer máximo $\sum_{i=1}^{n} \log P_X(x_i;\theta)$ lo que da lugar al estimador máximo verosímil de θ.

6.5. Eficiencia y eficiencia asintótica

Vimos (Teorema 5.1, pág. 64) que un estimador $\hat{\theta}_n$ insesgado de θ basado en una muestra aleatoria simple formada por n observaciones tenía su varianza acotada inferiormente:

$$\text{Var}_\theta(\hat{\theta}_n) \geq \frac{1}{nI_X(\theta)} \tag{6.7}$$

y decíamos que $\hat{\theta}_n$ es *eficiente* (Definición 5.2, pág. 68) si la relación anterior se verifica con igualdad. Es claro que un estimador eficiente no puede ser mejorado (en términos de varianza) por ningún otro en la clase de los insesgados regulares, pues el que lo hiciera violaría (6.7).

Consideremos una sucesión estimadora $\{\hat{\theta}_n\}$ cada uno de cuyos términos estima insesgadamente θ, y supongamos que se dan las condiciones de regularidad necesarias. Entonces, (6.7) se verifica para cada $\hat{\theta}_n$, $n = 1, 2, \ldots$, y $\mathrm{Var}_\theta(\sqrt{n}\hat{\theta}_n)$ (ó, equivalentemente, $\mathrm{Var}_\theta(\sqrt{n}[\hat{\theta}_n - \theta]) = n\mathrm{Var}_\theta(\hat{\theta}_n)$) ha de ser mayor o igual que $1/I_{\boldsymbol{X}}(\theta)$. Cabría esperar que si

$$\sqrt{n}[\hat{\theta}_n - \theta] \xrightarrow{\mathcal{L}} N(0, v(\theta)),$$

en que $\xrightarrow{\mathcal{L}}$ designa convergencia en distribución (Definición A.1, p. 144), la varianza de la distribución asintótica verificase:

$$v(\theta) \geq \frac{1}{I_{\boldsymbol{X}}(\theta)} \tag{6.8}$$

Este no es el caso. La aparente paradoja se desvanece cuando observamos que la varianza asintótica (= varianza de la distribución asintótica) no necesariamente tiene mucho que ver con el límite de la sucesión de varianzas. El siguiente ejemplo lo ilustra.

Ejemplo 6.3 Sea $\{Y_n\}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas como $N(0,1)$, y $\{X_n\}$ una sucesión de variables aleatorias definidas así:

$$X_n = \begin{cases} Y_n & \text{con probabilidad } 1 - \frac{1}{n}, \\ n & \text{con probabilidad } \frac{1}{n} \end{cases}$$

Entonces, es evidente que $X_n \xrightarrow{\mathcal{L}} X$, siendo X una variable $N(0,1)$, la media asintótica es 0 y la varianza asintótica 1. Sin embargo:

$$\begin{aligned} E[X_n] &= 0 \cdot \left(1 - \frac{1}{n}\right) + n \cdot \frac{1}{n} = 1 \\ \mathrm{Var}(X_n) &= E[X_n^2] - (E[X_n])^2 = \left(1 - \frac{1}{n}\right) \cdot 1 + n^2 \frac{1}{n} - 1^2 = \left(n - \frac{1}{n}\right) \end{aligned}$$

Mientras que la media y varianza de la distribución asintótica son respectivamente 0 y 1, los límites de la sucesión de medias y varianzas son:

$$\begin{aligned} \lim_{n\to\infty} E[X_n] &= 1 \\ \lim_{n\to\infty} \mathrm{Var}(X_n) &= \infty \end{aligned}$$

En general, se verifica (véase Lehmann (1983), pág. 405) que la varianza asintótica es menor o igual que el límite inferior de la sucesión de varianzas.

El ejemplo anterior muestra que límite de la sucesión de varianzas y varianza asintótica no tienen por qué coincidir. Una sucesión estimadora todos cuyos términos alcanzan la correspondiente cota de Cramér-Rao, podría dar lugar a una varianza asintótica *menor* que la que se deduciría de dicha cota. De nuevo un ejemplo aclara la situación.

Ejemplo 6.4 Sea $X_1, \ldots, X_n$ una muestra formada por observaciones $N(\theta, 1)$, y consideremos el siguiente estimador de θ:

$$\hat{\theta}_n = \begin{cases} \overline{X} & \text{si } |\overline{X}| \geq n^{-1/4}, \\ b\overline{X} & \text{si } |\overline{X}| < n^{-1/4}. \end{cases} \tag{6.9}$$

Entonces encontramos la siguiente situación: $\hat{\theta}_n$ se distribuye asintóticamente como $N(\theta, \sigma^2 = \frac{1}{n})$, *salvo* si $\theta = 0$. En este último caso, la distribución asintótica es $N(0, b^2/n)$, lo que mejora la varianza de $\overline{X}$ si $b^2 < 1$. ¡Tenemos un estimador de θ tan bueno como $\overline{X}$ —que sabemos insesgado de mínima varianza, y alcanzando la cota de Cramér-Rao— pero asintóticamente mejor para algunos valores del parámetro! En este caso, para $\theta = 0$. En efecto: $\sqrt{n}[\hat{\theta}_n - 0]$ converge en distribución a una variable aleatoria Z tal que:

$$\mathrm{Var}(Z) = b^2 < 1 = \frac{1}{I(\theta)}$$

El punto $\theta = 0$ en que el estimador considerado ve su varianza asintótica decrecer por debajo de $1/I(\theta)$ se dice que es de *supereficiencia*. Este ejemplo se debe a J. Hodges (ver Romano y Siegel (1986), pág. 229).

La existencia de puntos de supereficiencia, en que la varianza asintótica de un estimador regular puede descender por debajo de la cota de Cramer-Rao, es un fenómeno sin mayor interés práctico. En realidad, (6.8) *casi* es cierta, en el sentido de que el conjunto de puntos θ para los cuales no se verifica es de medida de Lebesgue cero. Por otra parte, el comportamiento supereficiente para algunos θ va siempre asociado a un comportamiento no eficiente en la vecindad de los mismos (ver Lehmann (1983), p. 408).

6.6. Normalidad y eficiencia asintótica del estimador máximo verosímil.

En condiciones bastante generales, el estimador MV no sólo es fuertemente consistente, sino que su distribución asintótica es normal. El siguiente resultado, cuya demostración meramente bosquejamos, muestra las condiciones necesarias para ello.

Teorema 6.2 *Sean $(X_1, \ldots, X_n)$ independientes e idénticamente distribuidas, con densidad común $f_X(x; \theta)$. Supongamos que se verifican las siguientes condiciones de regularidad:*

1. *El espacio paramétrico Θ es un intervalo abierto —no necesariamente finito—*
.

2. *Las funciones de densidad $f_X(x;\theta)$ tienen soporte común, que no depende de θ.*

3. *Las funciones de densidad $f_X(x;\theta)$ son tres veces diferenciables respecto a θ para cada x, y las derivadas son continuas en Θ.*

4. *La integral $\int f_X(x;\theta)dx$ puede ser diferenciada dos veces bajo el símbolo integral.*

5. *La información de Fisher verifica $0 < I(\theta) < \infty$.*

6. *La tercera derivada de $\log f_X(x;\theta)$ respecto a θ está acotada superiormente por una función $M(x)$ tal que $E_{\theta_0}[M(x)] < \infty$.*

Entonces, cualquier sucesión consistente $\hat{\theta}_n$ de soluciones de la ecuación de verosimilitud (y el estimador máximo verosímil proporciona una) satisface:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, I(\theta_0)^{-1}) \tag{6.10}$$

DEMOSTRACION:

Designemos, para aligerar la notación,

$$U_j(\theta) = \frac{\partial \log f_X(X_j, \theta)}{\partial \theta} \tag{6.11}$$

Desarrollando $\sum_{j=1}^n U_j(\hat{\theta}_{\text{MV},n})$ en torno a θ_0, obtenemos:

$$\begin{aligned} \sum_{j=1}^n U_j(\hat{\theta}_{\text{MV},n}) &= \sum_{j=1}^n U_j(\theta_0) + \sum_{j=1}^n U_j'(\theta_0)(\hat{\theta}_{\text{MV},n} - \theta_0) \\ &\quad + \frac{1}{2}\sum_{j=1}^n U_j''(\tilde{\theta})(\hat{\theta}_{\text{MV},n} - \theta_0)^2 \end{aligned} \tag{6.12}$$

en que $\tilde{\theta}$ es un punto intermedio entre $\hat{\theta}_{\text{MV},n}$ y θ_0, es decir, $|\tilde{\theta} - \theta_0| < |\hat{\theta}_{\text{MV},n} - \theta_0|$.

Pero $\hat{\theta}_{\text{MV},n}$, bajo condiciones de regularidad, anula el lado izquierdo de (6.12). Por tanto, tenemos que:

$$\sum_{j=1}^n U_j(\theta_0) = -\sum_{j=1}^n U_j'(\theta_0)(\hat{\theta}_{\text{MV},n} - \theta_0) - \frac{1}{2}\sum_{j=1}^n U_j''(\tilde{\theta})(\hat{\theta}_{\text{MV},n} - \theta_0)^2 \tag{6.13}$$

Sabemos (Lema 5.1, pág. 60) que $E_{\theta_0}[U_j(\theta_0)] = 0$. Por otra parte,

$$E_{\theta_0}[-U_j'(\theta_0)] = E_{\theta_0}[U_j(\theta_0)]^2 = I(\theta_0)$$

(Lema 5.2, pág. 61 y definición inmediatamente posterior). Dividiendo (6.13) entre $\sqrt{nI(\theta_0)}$ tenemos la igualdad:

$$\begin{aligned}\frac{\sum_{j=1}^n U_j(\theta_0)}{\sqrt{nI(\theta_0)}} &= \sqrt{nI(\theta_0)}(\hat{\theta}_{\text{MV},n}-\theta_0)\left[\frac{-\sum_{j=1}^n U_j'(\theta_0)}{nI(\theta_0)}\right.\\ &\left.-\frac{1}{2}\frac{\sum_{j=1}^n U_j''(\tilde{\theta})}{nI(\theta_0)}(\hat{\theta}_{\text{MV},n}-\theta_0)\right] \qquad (6.14)\end{aligned}$$

Los Lemas invocados y el teorema central del límite muestran que el lado izquierdo de (6.14) converge en distribución a una $N(0,1)$, y el primer término del corchete converge en probabilidad a 1 (ley débil de los grandes números, Teorema A.2). Como $U_j''(\tilde{\theta})$ tiene valor medio finito (condición 6 del enunciado) y $\hat{\theta}_{\text{MV},n}\xrightarrow{p}\theta_0$, el segundo término del corchete converge en probabilidad a cero. En consecuencia, reescribiendo (6.14) así:

$$\begin{aligned}\sqrt{nI(\theta_0)}(\hat{\theta}_{\text{MV},n}-\theta_0) &= \frac{\sum_{j=1}^n U_j(\theta_0)}{\sqrt{nI(\theta_0)}}\left[\frac{-\sum_{j=1}^n U_j'(\theta_0)}{nI(\theta_0)}\right.\\ &\left.-\frac{1}{2}\frac{\sum_{j=1}^n U_j''(\tilde{\theta})}{nI(\theta_0)}(\hat{\theta}_{\text{MV},n}-\theta_0)\right]^{-1}\end{aligned}$$

vemos que $\sqrt{nI(\theta_0)}(\hat{\theta}_{\text{MV},n}-\theta_0)$ es el producto de una sucesión aleatoria que converge en probabilidad a 1 y una sucesión aleatoria que converge en distribución a una $N(0,1)$. El Teorema A.1 permite entonces asegurar

$$\sqrt{nI(\theta_0)}(\hat{\theta}_{\text{MV},n}-\theta_0)\xrightarrow{\mathcal{L}} N(0,1)$$

que equivale a (6.10) en el enunciado del teorema.

■

Observación 6.6 Si $g(.)$ es función 1-1 de θ se ha mencionado ya que el estimador máximo verosímil de $g(\theta)$ es $g(\hat{\theta}_{\text{MV}})$. Supongamos además que para el verdadero valor del parámetro, θ_0, se verifica que $g'(\theta_0)\neq 0$. Entonces el teorema anterior admite la siguiente generalización:

$$\sqrt{n}(g(\hat{\theta}_{\text{MV}}-g(\theta_0))\xrightarrow{\mathcal{L}} N(0,I(\theta_0)^{-1}[g'(\theta_0)]^2).$$

La demostración es muy simple y se bosqueja a continuación. Desarrollando en serie $g(\hat{\theta}_{\text{MV}})$ hasta términos de primer orden,

$$g(\hat{\theta}_{\text{MV}})=g(\theta_0)+(\hat{\theta}_{\text{MV}}-\theta_0)\left[g'(\theta_0)+R_n\right],$$

en que R_n es el término complementario. Pero $R_n\xrightarrow{p}0$ cuando $\hat{\theta}_{\text{MV}}\xrightarrow{p}\theta_0$. Por consiguiente, siempre en uso del Teorema A.1, tenemos:

$$\sqrt{n}(g(\hat{\theta}_{\text{MV}})-g(\theta_0))\xrightarrow{\mathcal{L}} g'(\theta_0)\sqrt{n}(\hat{\theta}_{\text{MV}}-\theta_0)$$

y por tanto

$$\sqrt{n}(g(\hat{\theta}_{\text{MV}}) - g(\theta_0)) \xrightarrow{\mathcal{L}} g'(\theta_0)N(0, I(\theta_0)^{-1})$$

equivalente a la tesis.

6.7. Estimación máximo verosímil: inconvenientes

El desarrollo anterior muestra la estimación máximo verosímil desde una perspectiva muy favorable. No sólo es consistente —cualidad compartida con muchos otros tipos de estimadores, y ciertamente con cualquiera que estemos dispuestos a considerar—, sino también asintóticamente eficiente. Su distribución asintótica es normal sea cual fuere la de la población muestreada. Estas propiedades se verifican de modo bastante general, como los enunciados de los teoremas anteriores dejan traslucir.

Es importante ver, sin embargo, que se trata de propiedades que operan *en grandes muestras*. En pequeñas muestras, el comportamiento del estimador máximo verosímil puede ser bastante pobre. En ocasiones, la obtención del estimador máximo verosímil puede ser computacionalmente infactible. En otras, puede sencillamente no existir un máximo de la función de verosimilitud. Los ejemplos y observaciones que siguen tienen por objeto mostrar tales problemas en algunas situaciones. Ilustran algunos de los inconvenientes con que se puede tropezar al emplear estimadores máximo verosímiles.

Ejemplo 6.5 *(un estimador máximo verosímil de inviable utilización práctica)* Consideremos una variable aleatoria X con distribución de Cauchy y parámetro de localización θ. La verosimilitud asociada a una muestra de tamaño n es:

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^{n} \frac{1}{\pi} \frac{1}{1+(x_i-\theta)^2}$$

Tomando logaritmos, derivando, e igualando la derivada a cero, tenemos:

$$\begin{aligned} \frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{x};\theta)}{\partial \theta} &= -\sum_{i=1}^{n} \frac{2(x_i-\theta)(-1)}{1+(x_i-\theta)^2} && (6.15)\\ &= 2\frac{\sum_{i=1}^{n}(x_i-\theta)\prod_{j\neq i}\left[1+(x_j-\theta)^2\right]}{\prod_{j=1}^{n}\left[1+(x_j-\theta)^2\right]} && (6.16)\\ &= 0 && (6.17) \end{aligned}$$

El estimador máximo verosímil $\hat{\theta}_{\text{MV},n}$ ha de hacer que la igualdad anterior se verifique. Obsérvese que el numerador —que ha de anularse— es un polinomio de grado $2n-1$. La búsqueda de todas sus raíces para seleccionar entre ellas $\hat{\theta}_{\text{MV},n}$ es infactible a poco grande que sea n.

En ocasiones, el estimador máximo verosímil no existe, porque la verosimilitud no está acotada. Un caso trivial sería el de una variable aleatoria $X \sim N(\mu, \sigma^2)$,

de la que tenemos una única observación. Si quisiéramos estimadores máximo verosímiles de μ y σ^2, habríamos de maximizar:

$$\log f_X(x;\mu,\sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

Esta función no está acotada: tiende a ∞ cuando $\sigma^2 \to 0$.

El caso anterior es irrelevante a efectos prácticos, dado que nunca nos propondríamos estimar los dos parámetros de una distribución normal con una sóla observación. Sin embargo, el siguiente ejemplo muestra que situaciones similares son plausibles en la práctica.

Ejemplo 6.6 *(función de verosimilitud no acotada)* Supongamos una situación en que la variable aleatoria aleatoria X sigue habitualmente una distribución $N(\mu,1)$. Sin embargo, con probabilidad p, X puede proceder de una distribución $N(\mu,\sigma^2)$, con varianza desconocida. La descripción anterior podría convenir, por ejemplo, a un fenómeno en que la variable X está sujeta esporádicamente a cambios de régimen, dando lugar a *outliers*, u observaciones anómalas. La función de verosimilitud sería:

$$f_{\boldsymbol{X}}(\boldsymbol{x};\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \left[\frac{p}{\sigma} \exp\left\{ -\frac{(x_i-\mu)^2}{2\sigma^2} \right\} + (1-p)\exp\left\{ -\frac{(x_i-\mu)^2}{2} \right\} \right]$$

Observemos que dicho producto involucra términos que no están acotados. En efecto, consideremos un término tal como

$$\frac{p}{\sigma}\exp\left\{ -\frac{(x_i-\mu)^2}{2\sigma^2} \right\} \prod_{\substack{j=1 \\ j\neq i}}^{n} (1-p)\exp\left\{ -\frac{(x_j-\mu)^2}{2} \right\};$$

es fácil ver que para $\mu = x_i$ la expresión anterior crece sin límite cuando $\sigma \to 0$. Por tanto, incluso aunque tengamos muchas observaciones y la probabilidad p sea muy pequeña, el problema de inexistencia de un máximo global para la función de verosimilitud puede presentarse. Véase Cox y Hinkley (1974), pág. 291.

En ocasiones, el estimador máximo verosímil existe, pero con pequeñas muestras puede ser de muy pobres resultados. El siguiente ejemplo, algo artificial si se quiere, lo muestra de un modo bastante espectacular.

Ejemplo 6.7 *(un estimador máximo verosímil inadmisible)* Consideremos una variable aleatoria X binaria de parámetro θ. Sabemos que $\theta \in (\frac{1}{3},\frac{2}{3})$, y hemos de estimar dicho parámetro con ayuda de una única observación. La verosimilitud tendría por expresión:

$$f_X(x,\theta) = \theta^x(1-\theta)^{(1-x)} \qquad (x=0,1)$$

Con θ constreñida a estar en el intervalo indicado anteriormente, el estimador máximo verosímil es:

$$\hat{\theta}_{\mathrm{MV},n} = \begin{cases} \frac{1}{3} & \text{si } x=0, \\ \frac{2}{3} & \text{si } x=1 \end{cases}$$

y su error cuadrático medio resulta ser:

$$E[\hat{\theta}_{\mathrm{MV},n} - \theta]^2 = \theta\left(\frac{2}{3} - \theta\right)^2 + (1-\theta)\left(\frac{1}{3} - \theta\right)^2 = \frac{3\theta^2 - 3\theta + 1}{9} \quad (6.18)$$

Consideremos ahora un estimador que ignora el valor tomado por X y atribuye siempre a θ el valor $\frac{1}{2}$. Su error cuadrático medio sería:

$$E\left[\frac{1}{2} - \theta\right]^2 = \theta\left(\frac{1}{2} - \theta\right)^2 + (1-\theta)\left(\frac{1}{2} - \theta\right)^2 = \frac{4\theta^2 - 4\theta + 1}{4} \quad (6.19)$$

Efectuando la diferencia (6.18)-(6.19) vemos que es

$$\frac{-24\theta^2 + 24\theta - 5}{36}.$$

Examinando esta función se comprueba que en el intervalo $(\frac{1}{3}, \frac{2}{3})$ es siempre positiva; el estimador máximo verosímil resulta dominado incluso por uno que, como el propuesto, lejos de hacer uso óptimo de la información muestral, no hace *ningún* uso.

El valor de θ que maximiza la verosimilitud no tiene porqué ser único.

Ejemplo 6.8 Consideremos una distribución uniforme $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, de la que tomamos una muestra $X_1, \ldots, X_n$. Es fácil ver que cualquier valor $\theta \in [X_{(n)} - 1, X_{(1)} + 1]$ da lugar al mismo valor de la verosimilitud (= 1), y por tanto es igualmente válido como estimador máximo verosímil.

Menos simple, pero más frecuente en la práctica, es el caso de múltiples máximos locales y/o globales en la función de verosimilitud. Véase el Ejercicio 6.1.

El estimador máximo verosímil es frecuentemente sesgado en pequeñas muestras, aunque asintóticamente insesgado bajo las condiciones de regularidad que otorgan vigencia al Teorema 6.2.

Ejemplo 6.9 Consideremos el problema de estimar θ en una distribución uniforme, $U(0, \theta)$, con ayuda de una muestra de tamaño n. El estadístico suficiente y estimador máximo verosímil de θ es $X_{(n)}$, mayor de las observaciones (véase el Ejemplo 3.7, pág. 3.7). Es evidente que $X_{(n)} \leq \theta$ y como estimador de θ es por tanto sesgado por defecto.

De nuevo este es un ejemplo algo académico; pero en la práctica pueden encontrarse multitud de otros. Así, el estimador máximo verosímil de la varianza en una distribución normal es $s^2 = n^{-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$. Como en el caso anterior, el sesgo tiende a cero cuando $n \to \infty$.

Quizá la objeción más seria que puede plantearse al uso del estimador máximo verosímil es que obliga a especificar, salvo en los parámetros que se estiman, la forma de las distribuciones: es un requisito previo el fijar la familia de distribuciones que estamos dispuestos a considerar. Esto puede originar estimadores con propiedades no imaginadas. Por ejemplo, el suponer que la distribución originando X es $N(\theta, 1)$ nos llevaría a adoptar $\overline{X}$ como estimador de θ. Si la distribución

fuera de Cauchy, $\mathcal{C}(\theta)$, tal estimador tendría desastrosas propiedades —de hecho, no tendría varianza finita, cualquiera que fuera el tamaño muestral—.

Si la ausencia de robustez frente al incumplimiento de los supuestos distribucionales, la complejidad de cómputo, y el comportamiento, a veces, pobre en pequeñas muestras son inconvenientes, es preciso señalar que el estimador MV tiene todavía mucho en su haber[1]. Requiere no obstante cuidado el hacer uso inteligente de él.

CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

6.1 Examínese la función de verosimilitud de una distribución de Cauchy $\mathcal{C}(\theta)$ (se introdujo en el Ejemplo 6.5, pág. 84) y demuéstrese que tiene en general múltiples máximos relativos.

6.2 Sean $X_1, \ldots, X_n$ v.a. independientes con distribución binaria de parámetro θ. Se comprobó (Ejemplo 3.8, pág. 36) que no existe estimador insesgado de $\gamma(\theta) = \theta(1-\theta)$. ¿Hay estimador máximo verosímil de $\gamma(\theta)$? ¿Es único?

6.3 Si $\hat{\theta}_{\text{MV}}$ es el estimador máximo verosímil de θ y $\delta = \delta(\theta)$ es una función 1-1 de θ, entonces $\hat{\delta}_{\text{MV}} = \delta(\hat{\theta}_{\text{MV}})$. Demuéstrese. Si $\delta(\theta)$ es una función, por ejemplo, convexa, y $\hat{\theta}_{\text{MV}}$ es insesgado ¿qué podemos decir del sesgo de $\hat{\delta}_{\text{MV}}$? (Ayuda: hágase uso de la desigualdad de Jensen (Teorema 4.2, pág. 49).)

[1] Una vehemente opinión contraria al uso de máxima verosimilitud, enérgicamente contestada, puede verse en Berkson (1980). Es también interesante Rao (1962).

Capítulo 7

Estimación máximo verosímil en la práctica.

7.1. Introducción.

Como el Ejemplo 6.5 ponía de manifiesto, la obtención del estimador máximo verosímil puede no ser fácil. Incluso en el caso en que se tiene la certeza de que la verosimilitud tiene un único máximo relativo y es bien comportada, la solución analítica de la ecuación de verosimilitud

$$L'(\theta) = \sum_{j=1}^{n} U_j(\theta) = 0$$

puede ser inabordable. Se hace preciso acudir a métodos numéricos aproximados en muchas ocasiones.

La Sección 7.2 muestra que en la familia exponencial es posible en ocasiones obtener soluciones de las ecuaciones de verosimilitud de modo simple, igualando los valores muestrales de los estadísticos suficientes a sus valores medios. La Sección 7.3 presenta la aplicación del método general de Newton-Raphson a la resolución de la ecuación de verosimilitud. La Sección 7.4 presenta el método conocido como de *scoring*, estrechamente relacionado con el anterior. La Sección 7.5 describe con algún detalle el algoritmo EM, muy utilizado para maximizar verosimilitudes, que presenta la interesante ventaja de permitir trabajar de modo simple con verosimilitudes de datos incompletos.

7.2. Estimación máximo verosímil en la familia exponencial.

Consideremos el logarítmo de la verosimilitud en forma canónica de una distribución en la familia exponencial. Sin pérdida de generalidad, la escríbiremos en términos de sus parámetros naturales:

$$L(\boldsymbol{\theta}, \boldsymbol{x}) = \sum_{i=1}^{n} \left[\sum_{j=1}^{k} \theta_j b_j(x_i) + c(\boldsymbol{\theta}) + d(x_i) \right] \tag{7.1}$$

Como vimos en la Sección 3.5, el vector

$$(T_1, \ldots, T_k) = \left(\sum_{i=1}^{n} b_1(\boldsymbol{x_1}), \ldots, \sum_{i=1}^{n} b_k(\boldsymbol{x_1} \right)$$

proporciona de inmediato los estadísticos mínimos suficientes para el vector $\boldsymbol{\theta}$. Derivando el logaritmo de la verosimilitud respecto de $\theta_1, \ldots, \theta_k$ e igualando a cero para obtener puntos estacionarios de la función de verosimilitud tenemos:

$$\frac{\partial L(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \theta_j} \quad = \quad T_j + \frac{\partial nc(\boldsymbol{\theta})}{\partial \theta_j} = 0 \tag{7.2}$$

Las ecuaciones anteriores podrían proporcionar, si son de fácil solución, valores de $\hat{\theta}_1, \ldots, \hat{\theta}_k$, funciones de los estadísticos suficientes, candidatos a ser estimadores máximo verosímiles. Si recordamos (Lema 5.1) que

$$E_\theta \left[\frac{\partial L(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \theta_j} \right] = 0$$

obtenemos de (7.2) que:

$$E_\theta \left[\frac{\partial L(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \theta_j} \right] \quad = \quad E_\theta \left[T_j \right] + \frac{\partial nc(\boldsymbol{\theta})}{\partial \theta_j} = 0 \tag{7.3}$$

De (7.2)-(7.3) obtenemos entonces que ha de verificarse:

$$T_j - E_\theta[T_j] = 0$$

para $j = 1, \ldots, k$. La regla es pues simple: basta igualar los estadísticos suficientes a sus valores medios (funciones éstos últimos de $\boldsymbol{\theta}$) para obtener soluciones de las ecuaciones de verosimilitud. El ejemplo que sigue lo ilustra.

Ejemplo 7.1 Consideremos el caso de una normal multivariante $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Se desean los estimadores máximo verosímiles de $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

La verosimilitud de una mestra de tamaño n viene dada, por:

$$\prod_{i=1}^{n} \left\{ |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp \left[-\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma^{-1}} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right] \right\}$$

Si tomamos logaritmo nepariano de la expresión anterior y reordenamos sus términos podemos llegar a:

$$L(\boldsymbol{\theta}) = -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x_i}\right) - \frac{1}{2}\text{traza}\left\{\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x_i}\boldsymbol{x_i}'\right)\right\}$$

La expresión anterior exhibe los estadísticos conjuntamente suficientes para $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$: $(T_1, T_2) = (\sum_{i=1}^{n}\boldsymbol{x_i}, \sum_{i=1}^{n}\boldsymbol{x_i}\boldsymbol{x_i}')$ Igualando dichos estadísticos suficientes a sus valores medios, obtenemos:

$$E_\theta[T_1] = n\boldsymbol{\mu} = \sum_{i=1}^{n}\boldsymbol{x_i} \tag{7.4}$$

$$E_\theta[T_2] = n\boldsymbol{\Sigma} + n\boldsymbol{\mu}\boldsymbol{\mu}' = \sum_{i=1}^{n}\boldsymbol{x_i}\boldsymbol{x_i}'; \tag{7.5}$$

la primera ecuación inmediatamente proporciona $\hat{\boldsymbol{\mu}}_{\text{MV}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{x_i} = \overline{\boldsymbol{x}}$, que sustituido en la segunda proporciona $\hat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{x_i}\boldsymbol{x_i}' - \hat{\boldsymbol{\mu}}_{\text{MV}}\hat{\boldsymbol{\mu}}_{\text{MV}}{}'$.

7.3. Método de Newton-Raphson.

7.3.1. Descripción

Sea $\hat{\theta}$ una raíz de la ecuación de verosimilitud y $\hat{\theta}_{(1)}$ una solución inicial aproximada. Desarrollando en serie de potencias en torno a $\hat{\theta}_{(1)}$ hasta términos de segundo orden, obtenemos:

$$L'(\hat{\theta}) = 0 \simeq L'(\hat{\theta}_{(1)}) + L''(\hat{\theta}_{(1)})(\hat{\theta} - \hat{\theta}_{(1)}) \tag{7.6}$$

de donde:

$$\hat{\theta} \simeq \hat{\theta}_{(1)} - \frac{L'(\hat{\theta}_{(1)})}{L''(\hat{\theta}_{(1)})} \tag{7.7}$$

A partir de una aproximación inicial $\hat{\theta}_{(1)}$ la relación anterior proporciona otra. Nada impide emplear esta última como nueva aproximación inicial y repetir el proceso cuantas veces haga falta hasta convergencia, si se produce. Es decir, dada la aproximación $\hat{\theta}_{(n)}$ obtendremos la siguiente, $\hat{\theta}_{(n+1)}$, así:

$$\hat{\theta}_{(n+1)} = \hat{\theta}_{(n)} - \frac{L'(\hat{\theta}_{(n)})}{L''(\hat{\theta}_{(n)})} \tag{7.8}$$

deteniendo la iteración cuando $\hat{\theta}_{(n+1)}$ y $\hat{\theta}_{(n)}$ difieran entre sí en menos de una tolerancia preespecificada.

Es interesante señalar que *una sola iteración* empleando (7.8) basta para producir un estimador consistente y asintóticamente eficiente, siempre que el punto de partida $\hat{\theta}_{(1)}$ sea consistente "a la suficiente velocidad". El siguiente teorema hace precisa la anterior afirmación.

Teorema 7.1 *Supongamos que se verifican las condiciones en el Teorema 6.2, y que $\tilde{\theta}_n$ es un estimador que converge en probabilidad a θ de tal forma*[1] *que $(\tilde{\theta}_n - \theta) = O_p(n^{-\frac{1}{2}})$. Entonces,*

$$\hat{\theta}_n = \tilde{\theta}_n - \frac{L'(\tilde{\theta}_n)}{L''(\tilde{\theta}_n)} \tag{7.9}$$

es asintóticamente eficiente y normal.

La demostración puede encontrarse en Lehmann (1983), pág. 422.

■

La discusión precedente se generaliza fácilmente al caso en que hay un vector de parámetros a estimar, sin más que reemplazar en (7.6) $\hat{\theta}$ por un vector de estimadores y $L'(\theta)$ y $L''(\theta)$ por el vector gradiente $\nabla L(\theta)$ y la matriz de segundas derivadas $\nabla^2 L(\theta)$. La iteración toma entonces la forma:

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \left(\nabla^2 L(\hat{\theta}_n)\right)^{-1} \nabla L(\hat{\theta}_n) \tag{7.10}$$

7.3.2. Propiedades

Con el método de Newton-Raphson la convergencia no está garantizada. No obstante, si la verosimilitud es bien comportada, es un método eficaz y conduce en un número habitualmente pequeño de iteraciones a una raíz de la ecuación $L'(\theta) = 0$.

Definición 7.1 *Sea una ecuación $g(x) = 0$ cuya solución x_* buscamos. Sea x_n la aproximación obtenida mediante un método iterativo en la iteración n-ésima y $e_n = x_n - x_*$ el error de aproximación. Se dice que un método de solución de es de convergencia cuadrática cuando $e_n \propto (e_{n-1})^2$.*

Convergencia cuadrática. El método de Newton-Raphson para aproximar una raíz de $g(x) = 0$, cuando converge, goza de convergencia cuadrática. En efecto, supongamos una aproximación x_n lo suficientemente cercana a x_*. Consideremos $f(x) = x - g(x)/g'(x)$. Entonces,

$$e_n = x_n - x_* = x_n - f(x_*) \tag{7.11}$$
$$= f(x_{n-1}) - f(x_*) \tag{7.12}$$

Si desarrollamos $f(x_{n-1})$ en torno al punto x_*, de la igualdad anterior deducimos:

$$e_n = f(x_*) + f'(x_*)(x_{n-1} - x_*) + \frac{1}{2} f''(z)(x_{n-1} - x_*)^2 - f(x_*) \tag{7.13}$$

[1] Véase en el Apéndice A.4 el significado de la notación $O_p()$.

siendo z un punto entre x_{n-1} y x_*. Como

$$f'(x_*) = 1 - \frac{(g'(x_*))^2}{(g'(x_*))^2} + \frac{g(x_*)g''(x_*)}{(g'(x_*))^2} = 0, \quad (7.14)$$

tenemos

$$e_n = \frac{1}{2}f''(z)(x_{n-1} - x_*)^2,$$

lo que muestra que la iteración de Newton converge —cuando lo hace— cuadráticamente.

No monotonía. Naturalmente, nada garantiza que no podamos alcanzar una solución que sea máximo relativo de la verosimilitud en lugar de máximo global[2]. De hecho, la iteración anterior puede dar lugar a verosimilitudes decrecientes: el aproximarnos a una raíz de $L'(\theta)$ no garantiza que dicha raíz corresponda a un máximo relativo de $L(\theta)$.

Es posible modificar el algoritmo de Newton-Raphson de modo que la verosimilitud crezca monótonamente (lo que garantiza al menos que la convergencia es hacia un máximo relativo). En efecto, en (7.10) el "paso" de θ_n a θ_{n-1} es $\Delta\theta = (-\nabla^2 L(\hat{\theta}_n))^{-1}\nabla L(\hat{\theta}_n) = A\nabla L(\hat{\theta}_n)$, con $A = (-\nabla^2 L(\hat{\theta}_n))^{-1}$ Desarrollando en serie en torno al punto $\hat{\theta}_n$:

$$L(\hat{\theta}_n + \alpha\Delta\hat{\theta}) - L(\hat{\theta}_n) = \alpha[\nabla L(\hat{\theta}_n)]' A[\nabla L(\hat{\theta}_n)] + o(\alpha) \quad (7.15)$$

Para α lo suficientemente pequeño, el signo del lado derecho viene dado por el del primer sumando. Si A es simétrica definida positiva, entonces el signo es positivo y $L(\hat{\theta})$ se incrementa al pasar de $\hat{\theta}_n$ a

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \alpha\Delta\hat{\theta}_n.$$

Si con A definida como se ha indicado la forma cuadrática en la derecha de (7.15) no fuera definida positiva, podríamos definir: $\Delta\theta = A\nabla L(\hat{\theta}_n)$ con *cualquier* A simétrica definida positiva, y el argumento anterior proporcionaría un algoritmo monónotamente creciente en $L(\theta)$. Hay muchas posibles elecciones: con A igual a la matriz unidad, tenemos un algoritmo gradiente convencional. Si hacemos

$$A = E\left[-\nabla^2 L(\hat{\theta}_n)\right]$$

tenemos el algoritmo de *scoring* descrito en la sección que sigue. Otras elecciones y variantes son posibles: puede verse una discusión más completa en Lange (1998).

[2]La distribución de Cauchy, tan fecunda suministradora de contraejemplos, ilustra una vez más esta situación. La verosimilitud de su parámetro de ubicación tiene con gran frecuencia varios extremos relativos, si la muestra es grande.

7.4. Método *scoring* de Fisher.

El algoritmo de *scoring* procede de forma enteramente análoga al de Newton-Raphson. Su rasgo distintivo consiste en sustituir $-\nabla^2 L(\theta)$ por $-E\left[(\nabla L(\theta)\nabla L(\theta)'\right]$. Obsérvese que esta última matriz es, bajo las habituales condiciones de regularidad, definida positiva. A menudo su expresión es también relativamente simple, lo que hace fácil su cálculo en cada iteración. Por contra, el método de *scoring* puede ser acusadamente más lento que el de Newton-Raphson.

7.5. El algoritmo EM.

Aunque utilizable con completa generalidad, el algoritmo EM es preferentemente utilizado en el caso en que hay datos faltantes. La referencia seminal es Dempster et al. (1976), aunque en forma menos general las ideas subyacentes parecen haber existido antes. La descripción a continuación hace uso también de Laird (1993) y Navidi (1997). Una monografía reciente con muchas referencias es G.J.McLachlan y Krishnan (1997).

7.5.1. Notación

Consideraremos, por simplicidad notacional, el caso de un único parámetro θ; el caso multivariante no añade nada esencial. Denotaremos por $f_X(x;\theta)$ la verosimilitud de la muestra completa, si fuera observada: x es un vector o una matriz, no todas cuyas componentes son observadas. Observamos sólo y, y hay una relación $x = \mathcal{X}(y)$ que a cada y hace corresponder muchos posibles x (dicho de otro modo: la sóla observación de y no permite obtener de manera unívoca x).

Como parte de x es no observada, tendrá sentido escribir valores medios como

$$\begin{aligned} Q(\theta|\theta') &\stackrel{\text{def}}{=} E\log\left[f_X(x;\theta)|\theta', y\right] && (7.16)\\ &= \int_{\mathcal{X}(y)} \log f_X(x;\theta) f_{X|Y}(x|y;\theta')dx && (7.17)\\ H(\theta|\theta') &\stackrel{\text{def}}{=} E\log\left[f_{X|Y}(x|y;\theta)|\theta', y\right] && (7.18)\\ &= \int_{\mathcal{X}(y)} \log f_{X|Y}(x|y;\theta) f_{X|Y}(x|y;\theta')dx && (7.19)\end{aligned}$$

Denominemos,

$$L(\theta) \stackrel{\text{def}}{=} \log f_Y(y;\theta). \quad (7.20)$$

Como

$$\log f_{X|Y}(x|y;\theta) = \log f_X(x;\theta) - \log f_Y(y;\theta), \quad (7.21)$$

multiplicando cada término de (7.21) por $f_{X|Y}(x|y;\theta')$ e integrando, obtenemos:

$$Q(\theta|\theta') = L(\theta) + H(\theta|\theta'). \quad (7.22)$$

Estamos interesados en maximizar $L(\theta)$, la verosimilitud calculada con la parte de muestra y que realmente observamos.

7.5.2. La iteración EM

Si observáramos todo x, el problema de estimación máximo verosímil de θ se reduciría a maximizar una función. Como parte de x es inobservable, no podemos acometer directamente la maximización de $\log f_X(x;\theta)$. Una posibilidad sería sustituir la función desconocida por su valor esperado dada la parte de muestra que sí conocemos y bajo el *supuesto* de que $\theta = \theta'$; es decir, reemplazar $\log f_X(x;\theta)$ por $Q(\theta|\theta')$ y maximizar esta última. [3]

Observemos que para tomar el valor medio *necesitamos* el valor del parámetro (y si lo conociéramos, el problema de estimación máximo verosímil ya no tendría objeto). Una posibilidad sería;

1. *(Paso E)* Calcular $Q(\theta|\theta')$ para un valor θ', la mejor aproximación de θ que tengamos.

2. *(Paso M)* Maximizar $Q(\theta|\theta')$ respecto de θ.

3. Iterar los pasos anteriores hasta convergencia, si se produce.

La idea es que al ejecutar por primera vez el paso E (de valor **E**sperado, porque en dicho paso tomamos un valor medio) obtendremos una función no muy similar a la que querríamos maximizar. Por ello, el paso M (de **M**aximizar) no dará el máximo de la función que realmente desearíamos maximizar, sino el de una diferente. Pero este máximo suministra una nueva estimación de θ diferente de la inicial, presumiblemente mejor, que nos permite reiniciar el proceso.

La idea anterior constituye el núcleo del algoritmo EM, cuya iteración básica describimos más formalmente como Algoritmo 1.

Algorithm 1 – Algoritmo EM

1: Fijar valor inicial $\theta^{(0)}$ de θ.
2: Fijar ϵ {Mínima diferencia entre valores sucesivos de θ para seguir iterando.}
3: $i \leftarrow 0$
4: **repeat**
5: $\quad i \leftarrow i + 1$
6: $\quad Q(\theta|\theta^{(i-1)}) \leftarrow E\left[\log f_X(x;\theta)|\theta^{(i-1)}, y\right]$
7: $\quad \theta^{(i)} \leftarrow \arg\max_\theta Q(\theta|\theta^{(i-1)})$
8: **until** $\left\{|\theta^{(i)} - \theta^{(i-1)}| < \epsilon\right\}$
9: $\hat{\theta}_{\text{MV}} \leftarrow \theta^{(i)}$

[3] Nótese que para calcular el valor esperado de $\log f_X(x;\theta)$ *necesitamos* un punto de partida, es decir, un valor inicial θ' de θ; el algoritmo EM suministra una pauta para refinar este valor inicial hasta llegar al estimador máximo verosímil.

Lo que antecede muestra un modo de operar, pero nada garantiza que haya convergencia ni, caso de que la hubiera, que se produzca a un valor de θ maximizando la verosimilitud, siquiera sea localmente. Bosquejaremos ahora este resultado, mostrando que:

1. Cada iteración del Algoritmo 1 incrementa $L(\theta)$.

2. Si la verosimilitud $L(\theta)$ está acotada y

$$Q(\theta^{(i)}|\theta^{(i-1)}) - Q(\theta^{(i-1)}|\theta^{(i-1)}) \geq \lambda(\theta^{(i)} - \theta^{(i-1)})^2$$

   entonces $\theta^{(i)} \rightarrow \theta_*$.

3. Si $\theta^{(i)} \rightarrow \theta_*$ y

$$\left[\frac{\partial Q(\theta|\theta^{(i-1)})}{\partial \theta}\right]_{\theta=\theta^{(i)}} = 0,$$

   entonces

$$\left[\frac{\partial L(\theta)}{\partial \theta}\right]_{\theta=\theta_*} = 0.$$

Obsérvese que los tres resultados anteriores tomados en su conjunto, todavía no garantizan la convergencia del algoritmo EM a $\hat{\theta}_{\text{MV}}$ o a un máximo local. Para ello haría falta mostrar que el valor estacionario de la verosimilitud θ_* corresponde a un máximo y no a un mínimo o punto de silla. Una demostración completa que incluye éste y otros detalles puede encontrarse en Dempster et al. (1976).

Teorema 7.2 *En el Algoritmo 1, la verosimilitud crece monótonamente.*

DEMOSTRACION:

De (7.22) deducimos:

$$L(\theta^{(i)}) = Q(\theta^{(i)}|\theta^{(i-1)}) - H(\theta^{(i)}|\theta^{(i-1)}) \quad (7.23)$$
$$L(\theta^{(i-1)}) = Q(\theta^{(i-1)}|\theta^{(i-1)}) - H(\theta^{(i-1)}|\theta^{(i-1)}). \quad (7.24)$$

Restando (7.24) de (7.23) obtenemos

$$\begin{aligned} L(\theta^{(i)}) - L(\theta^{(i-1)}) &= (Q(\theta^{(i)}|\theta^{(i-1)}) - Q(\theta^{(i-1)}|\theta^{(i-1)})) \\ &+ (H(\theta^{(i-1)}|\theta^{(i-1)}) - H(\theta^{(i)}|\theta^{(i-1)})). \end{aligned} \quad (7.25)$$

El primer miembro de la derecha de (7.25) es no negativo por el modo en que ha sido tomado el paso M de la iteración (se maximiza $Q(\theta|\theta^{(i-1)})$ respecto de θ,

y por tanto necesariamente $Q(\theta^{(i)}|\theta^{(i-1)}) - Q(\theta^{(i-1)}|\theta^{(i-1)}) \geq 0$). El segundo término es necesariamente no negativo[4]. Por tanto, $L(\theta^{(i)}) - L(\theta^{(i-1)}) \geq 0$.

■

Teorema 7.3 *Cuando la verosimilitud está acotada, $L(\theta^{(i)}) \rightarrow L_*$, para algún valor L_*. Si, además,*

$$Q(\theta^{(i)}|\theta^{(i-1)}) - Q(\theta^{(i-1)}|\theta^{(i-1)}) \geq \lambda(\theta^{(i)} - \theta^{(i-1)})^2$$

para todo i, entonces $\theta^{(i)} \rightarrow \theta_*$.

DEMOSTRACION:

Una sucesión monótona acotada necesariamente tiene un límite: esto da cuenta de la existencia de L_*, a la vez que garantiza que los términos de la sucesión $L(\theta^{(i)})$ deben cumplir la condición de Cauchy para sucesiones convergentes. Por tanto, para todo $r > 1$ y $p > p(\epsilon)$

$$\sum_{j=1}^{r}(L(\theta^{(p+j)} - L(\theta^{p+j-1})) \quad = \quad |L(\theta^{(p+r)} - L(\theta^{p})| < \epsilon,$$

y por consiguiente

$$\begin{aligned}
\epsilon &> \sum_{j=1}^{r}(L(\theta^{(p+j)} - L(\theta^{p+j-1})) \\
&\geq \sum_{j=1}^{r}(Q(\theta^{(p+j)}|\theta^{(p+j-1)}) - Q(\theta^{(p+j-1)}|\theta^{(p+j-1)})) \\
&\geq \lambda\sum_{j=1}^{r}(\theta^{(p+j)} - \theta^{(p+j-1)})^2 \\
&\geq \lambda(\theta^{(p+r)} - \theta^{(p)})^2.
\end{aligned}$$

Ello muestra que $\theta^{(p)}$ verifica también una condición de Cauchy y en consecuencia converge a algún θ_*.

■

Establecido que $\theta^{(i)}$ converge, resta por ver que el límite, si es un punto estacionario de $Q(\theta|\theta)$, lo es también de la función de verosimilitud.

[4]Puede verse $H(\theta^{(i)}|\theta^{(i-1)}) - H(\theta^{(i-1)}|\theta^{(i-1)})$ como la distancia de Kullback-Leibler (véase (6.5), pág. 78) entre dos distribuciones de parámetros respectivos $\theta^{(i)}$ y $\theta^{(i-1)}$. Esta distancia se minimiza cuando $\theta^{(i)} = \theta^{(i-1)}$.

Teorema 7.4 *Supongamos que $\theta^{(i)} \to \theta_*$. Entonces, bajo condiciones de regularidad suficientes,*

$$\left[\frac{\partial L(\theta)}{\partial \theta}\right]_{\theta=\theta_*} = 0.$$

DEMOSTRACION:

Derivando en (7.22) obtenemos

$$\left[\frac{\partial L(\theta)}{\partial \theta}\right]_{\theta=\theta^{(i)}} = \left[\frac{\partial Q(\theta|\theta^{(i-1)})}{\partial \theta}\right]_{\theta=\theta^{(i)}} - \left[\frac{\partial H(\theta|\theta^{(i-1)})}{\partial \theta}\right]_{\theta=\theta^{(i)}} \quad (7.26)$$

Es claro que si la iteración converge, $\theta^{(i)}$ y $\theta^{(i-1)}$ en la expresión anterior pueden ambos sustituirse por θ_*. La derivada de $H(\theta|\theta')$ se anula para $\theta = \theta' = \theta_*$. La de $Q(\theta_*|\theta_*)$ también se anula —en cada iteración la función se maximiza, y su derivada por tanto se anula aunque no hayamos aún logrado convergencia—. En consecuencia, el lado izquierdo de (7.26) se anula.

■

7.5.3. Distribuciones de la familia exponencial.

Cuando trabajamos con distribuciones en la familia exponencial, el algoritmo puede en ocasiones simplificarse de modo notable. Consideremos una distribución cuya densidad escrita en términos de su parámetro natural (lo que no conlleva pérdida de generalidad) fuera

$$f_X(x;\theta) = e^{\theta b(x)+c(\theta)+d(x)}.$$

El logaritmo de la función de verosimilitud asociada a una muestra de tamaño n es

$$\begin{aligned}\log f_{\boldsymbol{X}}(\boldsymbol{x};\theta) &= \log \prod_{i=1}^{n}\left[e^{\theta b(x_i)+c(\theta)+d(x_i)}\right] \\ &= \theta\sum_{i=1}^{n} b(x_i) + nc(\theta) + \sum_{i=1}^{n} d(x_i) \\ &= \theta T(\boldsymbol{x}) + C(\theta) + D(\boldsymbol{x}).\end{aligned}$$

Entonces, la expresión (7.16) se convierte en

$$\begin{aligned}Q(\theta|\theta^{(i)}) &= E\left[\log f_X(x;\theta)|\theta^{(i)}, y\right] &(7.27)\\ &= E\left[\theta T(\boldsymbol{x}) + C(\theta) + D(\boldsymbol{x})|\theta^{(i)}, y\right] &(7.28)\\ &= \theta T^{(i)} + C(\theta) + E\left[D(\boldsymbol{x})|\theta^{(i)}, y\right]. &(7.29)\end{aligned}$$

Podemos reemplazar esta expresión de $Q(\theta|\theta^{(i)})$ en el lugar correspondiente del Algoritmo 1. Observemos, adicionalmente, que el último término en (7.29) no depende de θ. Por lo tanto, podemos maximizar respecto de θ sólamente la expresión $\theta T^{(i)} + C(\theta)$. Incorporando estos cambios al Algoritmo 1, obtenemos el Algoritmo 2.

Algorithm 2 – Algoritmo EM para distribuciones en la familia exponencial

1: Fijar valor inicial $\theta^{(0)}$ de θ.
2: Fijar ϵ {Mínima diferencia entre valores sucesivos de θ para seguir iterando.}
3: $i \leftarrow 0$
4: **repeat**
5: $\quad i \leftarrow i + 1$
6: $\quad T^{(i)} \leftarrow E\left[T(\boldsymbol{x})|\theta^{(i-1)}, y\right]$
7: $\quad \theta^{(i)} \leftarrow \arg\max_\theta \left(\theta T^{(i)} + C(\theta)\right)$
8: **until** $\left\{|\theta^{(i)} - \theta^{(i-1)}| < \epsilon\right\}$
9: $\hat{\theta}_{\text{MV}} \leftarrow \theta^{(i)}$

Ejemplo 7.2 El siguiente ejemplo, adaptado de Laird (1993), ilustra el funcionamiento del algoritmo EM en una distribución de la familia exponencial. Supongamos observaciones procedentes de uan distribución trinomial con vector de parámetros $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ (uno redundante, al estar constreñidos a sumar 1). Poseemos una muestra tomada al azar incompletamente clasificada, como recoge la siguiente tabla:

θ_1	θ_2	θ_3	
21	9	20	$n_{1.} = 50$
8		7	$n_{2.} = 15$
$n_{,1}$	$n_{,2}$	$n_{,3}$	

Hay $n_{1.} = 50$ observaciones completamente clasificadas; por el contrario, hay $n_{2.} = 15$ de las que sólo sabemos si pertenecen a la clase tercera o a una de las dos primeras.

Es claro que $n_{,1}, n_{,2}, n_{,3}$ son estadísticos suficientes para $\boldsymbol{\theta}$; pero sólo $n_{,3}$ es conocido. El algoritmo EM procede sustituyendo $n_{,1}$ y $n_{,2}$ por sus respectivos valores esperados para obtener una estimación de $\boldsymbol{\theta}$. Obtenida ésta, se utiliza para recalcular los valores esperados de $n_{,1}$ y $n_{,2}$, y se itera hasta convergencia.

En el caso que nos ocupa, una estimación inicial de $\boldsymbol{\theta}$ podría ser la máximo verosímil con las 50 observaciones completamente clasificadas[5]: $\hat{\theta}^{(0)} = (\frac{21}{50}, \frac{9}{50}, \frac{20}{50})$.

[5]Podríamos comenzar con un vector arbitrario, pero si tenemos alguna aproximación razonable, como en este caso, ello acelera la convergencia.

Tenemos ahora que los valores esperados de los estadísticos suficientes $n_{,1}, n_{,2}$ y $n_{,3}$ dado $\theta = \hat{\theta}^{(0)}$ son:

$$\begin{aligned} n_{,1}^{(1)} &= 21 + 8 \times \frac{\hat{\theta}_1^{(0)}}{\hat{\theta}_1^{(0)} + \hat{\theta}_2^{(0)}} \simeq 26{,}6 \\ n_{,2}^{(1)} &= 9 + 8 \times \frac{\hat{\theta}_2^{(0)}}{\hat{\theta}_1^{(0)} + \hat{\theta}_2^{(0)}} \simeq 11{,}4 \\ n_{,3}^{(1)} &= 27. \end{aligned}$$

En esencia, hemos "repartido" las 8 observaciones cuya adscripción no consta entre las clases primera y segunda sobre la base de la mejor información disponible acerca de $\boldsymbol{\theta}$. Con los valores esperados (de $n_{,1}$ y $n_{,2}$) u observados (de $n_{,3}$) de los estadísticos suficientes podemos ahora obtener una estimación refinada del vector de parámetros, $\hat{\theta}^{(1)} = (\frac{26{,}6}{65}, \frac{11{,}4}{65}, \frac{27}{65})$, con la que recalcular los valores medios de los estadísticos suficientes que lo precisan, y así hasta convergencia.

Capítulo 8

Contraste de Hipótesis.

8.1. Introducción.

Examinaremos en lo que sigue el caso en que existen dos posibles estados de la naturaleza, asociados a sendos conjuntos de valores de un cierto parámetro: así, un estado corresponde a $\theta \in \Theta_0$ y otro a $\theta \in \Theta_a$. Un *contraste de hipótesis* es un procedimiento estadístico $\delta(\boldsymbol{X})$ para escoger entre ambos estados (inobservables) sobre la base de la información muestral proporcionada por una variable aleatoria $\boldsymbol{X}$ con densidad (o cuantía) $f_{\boldsymbol{X}|\theta}(\boldsymbol{x}\,|\theta)$. El procedimiento $\delta(\boldsymbol{X})$ puede proporcionar una de dos decisiones: d_0 (= "el estado es Θ_0") y d_a (= "el estado es Θ_a").

Frecuentemente, ésta es una elección bastante artificial, entre dos alternativas ninguna de las cuales tiene visos de ser "exactamente" cierta. Esto es particularmente cierto cuando se contrastan hipótesis que especifican un único y preciso valor para algún parámetro (como $H_0 : \theta = \theta_0$). Sin embargo, como hace notar Garthwaite et al. (1995), pág. 2, el contraste de hipótesis

> " . . .es a menudo un modo conveniente de actuar y subyace a una parte importante de la investigación científica."

De que esto es así da testimonio el uso continuo e intenso que se hace del contraste de hipótesis en muchas ramas del saber. Que la metodología habitualmente utilizada para contrastar hipótesis no siempre se emplea debidamente, es también un hecho. Véase al respecto la crítica enérgica y virulenta que del contraste de hipótesis se hace en Wang (1993).

Se dice que una clase de distribuciones es *simple* si contiene una única distribución. Es compuesta en caso contrario. Un contraste de hipótesis será simple si tanto Θ_0 como Θ_a especifican una única distribución.

Si disponemos de una función de pérdida completamente especificada, emplearemos la teoría examinada en capítulos anteriores para seleccionar un procedimiento adecuado: procedimiento de Bayes (si disponemos además de una distribución *a priori* para θ), minimax, etc.

Es frecuente, sin embargo, que no haya una función de pérdida bien especificada. El contraste se efectúa entonces de manera convencional minimizando la probabilidad de error, que puede ser de dos clases: el error de tipo I (o de tipo α) consiste en seleccionar d_a cuando $\theta \in \Theta_0$, mientras que el error de tipo II (o de tipo β) consiste en seleccionar d_0 cuando $\theta \in \Theta_a$. Denominamos *nivel de significación* de un contraste (a veces también llamado *tamaño* del contraste) al supremo de la probabilidad de error de tipo I:

$$\alpha \stackrel{\text{def}}{=} \sup_{\theta \in \Theta_0} \text{Prob}\{\delta(\boldsymbol{X}) = d_a\}$$

y *potencia* $\Pi(\theta)$ al complemento a uno de la probabilidad de error de tipo II:

$$\Pi(\theta) \stackrel{\text{def}}{=} 1 - \beta(\theta) \stackrel{\text{def}}{=} 1 - \text{Prob}\{\delta(\boldsymbol{X}) = d_0; \theta \in \Theta_a\}$$

Siempre es preciso establecer un compromiso entre ambos tipos de error. Es habitual fijar el nivel de significación α en un valor convencional como 0.01, 0.05 ó 0.10 y tratar de encontrar el contraste que minimiza $\beta(\theta)$ (o, lo que es lo mismo, que maximiza la potencia) de entre todos los que tienen el nivel de significación prefijado.

En su forma más sencilla, un contraste de hipótesis puede verse como particionando el espacio muestral en dos regiones. Una de ellas, llamada *región crítica*, S, agrupa los resultados muestrales $\boldsymbol{X}$ cuya observación daría lugar a $\delta(\boldsymbol{X}) = d_a$, en tanto la otra región S^c agrupa los resultados cuya observación daría lugar a $\delta(\boldsymbol{X}) = d_0$. Alternativamente, un contraste quedaría completamente especificado mediante su *función crítica* $\lambda(\boldsymbol{x})$, definida así:

$$\lambda(\boldsymbol{x}) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{si } \boldsymbol{x} \in S, \\ 0 & \text{si } \boldsymbol{x} \notin S. \end{cases} \tag{8.1}$$

Si insistimos en obtener un contraste con un α prefijado, puede ser preciso complicar ligeramente las cosas. El siguiente ejemplo muestra un caso muy simple en que no existe una región crítica proporcionando un $\alpha = 0{,}07$ (naturalmente, no hay ninguna razón especial por la que en la práctica no hubiéramos de contentarnos con $\alpha = 0{,}05$ ó $\alpha = 0{,}08$, que sí son accesibles; el ejemplo tiene finalidad exclusivamente ilustrativa).

Ejemplo 8.1 Consideremos el caso en que hemos de contrastar $H_0 : \theta = \theta_0$ frente a la alternativa $H_a : \theta = \theta_a$. Las distribuciones asociadas a cada valor del parámetro son las especificadas en la tabla siguiente:

x	0	1	2	3	4	5
Prob $\{x; \theta_0\}$	0.60	0.26	0.05	0.04	0.04	0.01
Prob $\{x; \theta_a\}$	0.10	0.15	0.10	0.25	0.30	0.10

Si tomamos como estadístico de contraste una única observación X y como región crítica $S = \{4, 5\}$ ó $S = \{3, 5\}$, el nivel de significación es $\alpha = 0{,}05$. Podemos tomar otros puntos en otras combinaciones para obtener $\alpha = 0{,}06$, $\alpha = 0{,}08$ y $\alpha = 0{,}09$, pero no $\alpha = 0{,}07$

El problema se presenta en el ejemplo anterior debido al carácter discreto de la distribución: no podemos incrementar con la suficiente finura la probabilidad bajo θ_0 de la región crítica. Tal problema puede sin embargo resolverse recurriendo a procedimientos aleatorizados.

Ejemplo 8.2 Supongamos que, en el ejemplo anterior, estamos dispuestos a considerar procedimientos aleatorizados. Entonces podríamos obtener un nivel de significación exacto de 0.07. Podríamos, por ejemplo, tomar una región crítica $S = \{4, 5\}$, que totaliza $\alpha = 0{,}05$ y añadir "parte" del punto $x = 3$. Para "despiezar" dicho punto, podemos construir una lotería que con probabilidad $\frac{1}{2}$ proporcione rechazo de H_0 y con probabilidad $\frac{1}{2}$ aceptación de H_0. Si adoptamos la regla de rechazar H_0 siempre que obtengamos $X = 4$ ó $X = 5$ y de jugar a la lotería indicada cuando obtengamos $X = 3$, la probabilidad total de rechazo cuando $\theta = \theta_0$ es:

$$\alpha = 0{,}04 + 0{,}01 + \frac{1}{2} \times \text{Prob}\{X = 3; \theta_0\} = 0{,}07$$

Para recoger el caso en que nos vemos obligados a realizar contrastes aleatorizados debemos considerar funciones críticas algo más complejas que la descrita en (8.1). Un contraste general vendrá así especificado por una función crítica como:

$$\lambda(\boldsymbol{x}) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{si } \boldsymbol{x} \in S^+, \\ \gamma & \text{si } \boldsymbol{x} \in S^= \stackrel{\text{def}}{=} (S^+ \cup S^-)^c, \\ 0 & \text{si } \boldsymbol{x} \in S^-. \end{cases} \tag{8.2}$$

S^+ es la región crítica, y S^- la región no crítica. El conjunto de puntos muestrales que no pertenecen ni a una ni a otra da lugar al rechazo con probabilidad γ. En el Ejemplo 8.2, $S^+ = \{4, 5\}$, $S^- = \{0, 1, 2\}$ y $(S^+ \cup S^-)^c = \{3\}$.

Observemos finalmente que en términos de la función crítica:

$$\text{Potencia } = \Pi(\theta) = 1 - \beta(\theta) = E_\theta(\lambda(\boldsymbol{X}))$$

y para contrastes con nivel de significación α ha de verificarse:

$$E_\theta \lambda(\boldsymbol{X}) \leq \alpha \qquad \forall \theta \in \Theta_0$$

8.2. El Teorema de Neyman–Pearson.

La construcción de regiones críticas para el contraste de una hipótesis simple $\theta = \theta_0$ frente a una alternativa también simple $\theta = \theta_a$ resulta sumamente fácil (al menos conceptualmente) gracias al siguiente resultado.

Teorema 8.1 *Sea un problema de decisión consistente en escoger entre dos posibles estados de la naturaleza, θ_0 y θ_a. Para cualquier $\alpha \in [0,1]$, existe un contraste $\lambda(\boldsymbol{x})$ y una constante $k > 0$ verificando:*

(i)

$$\lambda(\boldsymbol{x}) = \begin{cases} 1 & \text{cuando } f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a) > k f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0), \\ \gamma & \text{cuando } f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a) = k f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0), \\ 0 & \text{cuando } f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a) < k f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0). \end{cases} \tag{8.3}$$

$$E_{\theta_0}\lambda(\boldsymbol{X}) = \alpha \tag{8.4}$$

(ii) Las condiciones (8.3)–(8.4) son suficientes para garantizar que el contraste $\lambda(\boldsymbol{x})$ es el más potente para la hipótesis θ_0 frente a θ_a al nivel α.

(iii) Recíprocamente, si $\lambda(\boldsymbol{x})$ es el contraste más potente para el par de hipótesis citadas, entonces verifica (8.3)–(8.4) para algún valor k, a menos que exista un contraste de tamaño menor que α y potencia 1.

DEMOSTRACION:

Bosquejamos a continuación la demostración. Un mayor detalle puede encontrarse en Lehmann (1959), p. 65.

Para $\alpha = 0$ ó $\alpha = 1$ el teorema es trivial. Sea:

$$\alpha(c) \stackrel{\text{def}}{=} \text{Prob}\left\{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a) > c f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)|\theta_0\right\}$$

Como $\alpha(c)$ es una probabilidad computada cuando $\theta = \theta_0$, podemos desentendernos de los puntos $\boldsymbol{x}$ en que $f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0) = 0$, y escribir:

$$\begin{aligned} \alpha(c) &= \text{Prob}\left\{\frac{f_{\boldsymbol{X}}(\boldsymbol{X};\theta_a)}{f_{\boldsymbol{X}}(\boldsymbol{X};\theta_0)} > c|\theta_0\right\} \\ &= 1 - \text{Prob}\left\{\frac{f_{\boldsymbol{X}}(\boldsymbol{X};\theta_a)}{f_{\boldsymbol{X}}(\boldsymbol{X};\theta_0)} \leq c|\theta_0\right\} \\ \Longrightarrow 1 - \alpha(c) &= \text{Prob}\left\{\frac{f_{\boldsymbol{X}}(\boldsymbol{X};\theta_a)}{f_{\boldsymbol{X}}(\boldsymbol{X};\theta_0)} \leq c|\theta_0\right\} \end{aligned}$$

Por tanto, $1-\alpha(c)$ es una función de distribución, no decreciente y continua por la derecha, y $\alpha(c)$ es no creciente y continua por la derecha, verificando $\alpha(-\infty) = 1$ y $\alpha(\infty) = 0$. Para cualquier $\alpha \in [0,1]$ existirá por tanto un c_0 verificando:

$$\alpha(c_0) \leq \alpha \leq \alpha(c_0^-)$$

Sea el contraste:

$$\lambda(\boldsymbol{x}) = \begin{cases} 1 & \text{cuando } f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a) > c_0 f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0), \\ \dfrac{\alpha - \alpha(c_0)}{\alpha(c_0^-) - \alpha(c_0)} & \text{cuando } f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a) = c_0 f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0), \\ 0 & \text{cuando } f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a) < c_0 f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0). \end{cases} \tag{8.5}$$

Es fácil ver que no hay problemas de anulación del denominador en el quebrado que aparece en la definición, pues el conjunto de puntos en que éste se anula tiene probabilidad cero. En consecuencia, (8.5) define casi en todo punto (con respecto a $f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)$) el contraste $\lambda(\boldsymbol{x})$. El tamaño de dicho contraste es:

$$\begin{aligned} E_{\theta_0}\left[\lambda(\boldsymbol{X})\right] &= \text{Prob}\left\{\frac{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a)}{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)} > c_0|\theta_0\right\} \\ &\quad + \quad \frac{\alpha-\alpha(c_0)}{\alpha(c_0^-)-\alpha(c_0)}\text{Prob}\left\{\frac{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a)}{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)} = c_0|\theta_0\right\} \\ &= \alpha \end{aligned}$$

Esto da cuenta de la existencia. Comprobemos ahora (ii). Sea $\lambda(\boldsymbol{x})$ el contraste definido en (8.5) y $\lambda^*(\boldsymbol{x})$ cualquier otro, de tamaño no mayor que α: $E_{\theta_0}\lambda^*(\boldsymbol{X}) \leq \alpha$. Sean S^+, $S^=$, y S^- las tres regiones del espacio muestral en que se verifican, respectivamente, cada una de las tres condiciones expresadas en (8.5). Puede verse que sobre cualquiera de dichas regiones:

$$\int (\lambda(\boldsymbol{x})-\lambda^*(\boldsymbol{x}))(f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a)-c_0 f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0))d\boldsymbol{x} \geq 0 \tag{8.6}$$

En efecto: cuando $(f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a)-c_0 f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)) > 0$, $\lambda(\boldsymbol{x}) = 1$, y por tanto $(\lambda(\boldsymbol{x}) - \lambda^*(\boldsymbol{x})) \geq 0$; el integrando es por consiguiente no negativo. Cuando $(f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a) - c_0 f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)) < 0$, $\lambda(\boldsymbol{x}) = 0$, $(\lambda(\boldsymbol{x}) - \lambda^*(\boldsymbol{x})) \leq 0$, y el integrando es de nuevo no negativo. Por consiguiente, la integral (8.6) extendida a todo S es no negativa, y realizando el producto en el integrando obtenemos:

$$\int_S (\lambda(\boldsymbol{x})-\lambda^*(\boldsymbol{x}))f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a)d\boldsymbol{x} - \int_S (\lambda(\boldsymbol{x})-\lambda^*(\boldsymbol{x}))c_0 f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0))d\boldsymbol{x} \geq 0 \tag{8.7}$$

$$\text{Potencia}(\lambda(\boldsymbol{X})) - \text{Potencia}(\lambda^*(\boldsymbol{X})) - c_0 \underbrace{(\alpha - E_{\theta_0}\lambda^*(\boldsymbol{X}))}_{\geq 0} \geq 0 \tag{8.8}$$

Por tanto:

$$\text{Potencia}(\lambda(\boldsymbol{X})) \geq \text{Potencia}(\lambda^*(\boldsymbol{X}))$$

Comprobemos finalmente (iii). Sea $\lambda^*(\boldsymbol{x})$ el contraste más potente de tamaño α para θ_0 frente a θ_a. Sea por otra parte $\lambda(\boldsymbol{x})$ el contraste verificando (8.3)-(8.4). Denominemos $\mathcal{C}$ al conjunto de puntos muestrales verificando:

$$\mathcal{C} = \{\boldsymbol{x} : [\lambda^*(\boldsymbol{x}) \neq \lambda(\boldsymbol{x})] \wedge [f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a) \neq k f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)]\}$$

Vamos a ver que $\mathcal{C}$ tiene medida cero, y por tanto ambos contrastes son esencialmente el mismo. Como ya se ha visto en el apartado (ii):

$$\int_S (\lambda(\boldsymbol{x})-\lambda^*(\boldsymbol{x}))(f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a)-k f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0))d\boldsymbol{x} \geq 0$$

Pero basta que integremos en $\mathcal{C}$ (pues fuera de $\mathcal{C}$ el integrando se anula). Por tanto:

$$\begin{aligned}\int_{\mathcal{C}}(\lambda(\boldsymbol{x})-\lambda^*(\boldsymbol{x}))f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a)d\boldsymbol{x} &> k\int_{\mathcal{C}}(\lambda(\boldsymbol{x})-\lambda^*(\boldsymbol{x}))f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)d\boldsymbol{x}\\ &= k\left(\alpha - E_{\theta_0}\lambda^*(\boldsymbol{X})\right)\end{aligned}$$

La integral del lado izquierdo es la diferencia de potencias, y el lado derecho — si $\lambda^*(\boldsymbol{x})$ está constreñido a tener nivel de significación no mayor que α— es no negativo. Por tanto, $\lambda(\boldsymbol{x})$ sería más potente que $\lambda^*(\boldsymbol{x})$, contra la hipótesis, a menos que $\mathcal{C}$ sea un conjunto de probabilidad cero cuando $\theta=\theta_0$.

■

Observación 8.1 Los contrastes pueden diferir en $\{\boldsymbol{x} : f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a) = k f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)\}$. La definición de cualquiera de ambos contrastes en dicha región "frontera" no afecta a sus respectivas potencias, y es por tanto arbitraria.

Observación 8.2 La decisión a tomar depende de la muestra sólo a través de $f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a)/f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)$. No es extraño que esto suceda. Vimos (Ejemplo 3.10) que la razón de verosimilitudes es un estadístico suficiente, y (Sección 3.6) que los procedimientos de Bayes pueden siempre hacerse depender de estadísticos suficientes. El empleo del teorema de Neyman-Pearson proporciona pues acceso a todos los procedimientos de Bayes. Como se vio en la Sección 1.10, tal clase completada con sus límites incluye en general la totalidad de los procedimientos que deseamos considerar (admisibles). La relación entre el teorema de Neyman-Pearson y la Teoría de la Decisión esbozada en el Capítulo 1 resulta adicionalmente clarificada en la Sección 8.3.

Observación 8.3 Del contenido de la Sección anterior se desprende que la potencia de un contraste varía de acuerdo con la alternativa considerada. De hecho, se ha definido potencia (en (8.1)) como *una función* de θ. Es claro pues que, en general, el contraste de tamaño α más potente de θ_0 frente a θ_1 no coincidirá con el de igual tamaño y máxima potencia de θ_0 frente a θ_2. Hay casos, sin embargo, en que un mismo contraste es el más potente frente a una clase compuesta de alternativas Θ_a. Se dice que es *uniformemente más potente (UMP)* para dicha clase de alternativas. Volveremos sobre esto en la Sección 8.4.

8.3. Teorema de Neyman-Pearson y procedimientos de Bayes.

Sea el problema de contrastar una hipótesis simple $H_0 : \theta = \theta_0$ frente a una alternativa también simple, $H_a : \theta = \theta_a$. Supongamos que hay una distribución *a priori* definida sobre θ, que atribuye probabilidades ξ_0 y ξ_a respectivamente a θ_0 y θ_a.

Designemos por c_0 y c_a los costes respectivos de tomar equivocadamente las decisiones $d_0 : \theta = \theta_0$ y $d_a : \theta = \theta_a$.

Estudiemos el problema de construir un contraste $\lambda(\boldsymbol{x})$ cuyo riesgo de Bayes $R_\xi(\lambda)$ sea mínimo. Tenemos que:

$$\begin{aligned} R_\xi(\lambda) &= \int_{R^n} c_a\xi_0\lambda(\boldsymbol{x})f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)d\boldsymbol{x} + \int_{R^n} c_0\xi_a(1-\lambda(\boldsymbol{x}))f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a)d\boldsymbol{x} \\ &= \int_{R^n} \lambda(\boldsymbol{x})\left[c_a\xi_0 f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0) - c_0\xi_a f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a)\right]d\boldsymbol{x} \\ &\qquad + \int_{R^n} c_0\xi_a f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a)d\boldsymbol{x} \end{aligned} \tag{8.9}$$

Como quiera que el segundo sumando de (8.9) no depende de $\lambda(\boldsymbol{x})$, basta minimizar el primero; y es claro que para ello debemos tomar:

$$\begin{aligned} \lambda(\boldsymbol{x}) &= 1 \text{ cuando } c_0\xi_a f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a) - c_a\xi_0 f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0) > 0 \\ \lambda(\boldsymbol{x}) &= 0 \text{ cuando } c_0\xi_a f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a) - c_a\xi_0 f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0) < 0 \end{aligned}$$

Es decir, $\lambda(\boldsymbol{x}) = 1$ si:

$$\frac{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)}{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a)} < \frac{c_0\xi_a}{c_a\xi_0} \tag{8.10}$$

que es precisamente la condición que establece el teorema de Neyman-Pearson para rechazar θ_0 en beneficio de θ_a. Hay una diferencia, no obstante: el enfoque basado en la Teoría de la Decisión fija el valor que debe tener el umbral a superar por la razón de verosimilitudes para que se produzca el rechazo de θ_0; analizando (8.10) vemos además que dicho umbral depende de la forma intuitivamente esperable de los parámetros c_0, c_a, ξ_0 y ξ_a.

El enfoque basado en el Teorema de Neyman-Pearson proporciona una familia de contrastes idéntica, pero el umbral a superar por la razón de verosimilitudes se fija estableciendo (habitualmente de modo un tanto arbitrario) el nivel de significación deseado. Cuando se disponga de una función de pérdida especificada y de una distribución *a priori* sobre las dos posibles hipótesis competidoras, el uso de (8.10) parece lo indicado. En caso contrario, habrá de hacerse uso del Teorema de Neyman-Pearson, con la precaución de especificar un nivel de significación tanto más pequeño (= un rechazo tanto más difícil) cuanto más grave sea la adopción injustificada de θ_a, o más fuerte sea la creencia de encontrarnos ante θ_0.

8.4. Contrastes uniformemente más potentes (UMP).

Se ha indicado ya que, en general, el contraste más potente proporcionado por el Teorema de Neyman-Pearson depende tanto de la hipótesis nula como de la alternativa. En algunas circunstancias, no obstante, dada una hipótesis nula H_0, el

mismo contraste $\lambda(\boldsymbol{x})$ es el más potente de tamaño α para todas las alternativas en una cierta clase. Se dice que es *uniformemente más potente* (UMP) en dicha clase.

Ejemplo 8.3 Consideremos una muestra procedente de una población con distribución exponencial $f_X(x,\theta) = \theta^{-1}e^{-x/\theta}$, $\theta > 0$, con ayuda de la cual queremos contrastar $H_0 : \theta = \theta_0$ frente a la alternativa (compuesta) $H_a : \theta > \theta_0$. Para *cualquier* $\theta_a > \theta_0$, el teorema de Neyman-Pearson prescribe tomar como región crítica la formada por los $\boldsymbol{x}$ verificando

$$\frac{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_a)}{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)} = \left(\frac{\theta_0}{\theta_a}\right)^n \exp\left\{-\sum_{i=1}^n x_i\left(\frac{1}{\theta_a}-\frac{1}{\theta_0}\right)\right\} \geq c,$$

o equivalentemente

$$\exp\left\{-\sum_{i=1}^n x_i\left(\frac{\theta_0-\theta_a}{\theta_0\theta_a}\right)\right\} > c\left(\frac{\theta_a}{\theta_0}\right)^n$$

$$\sum_{i=1}^n x_i > \left[\log_e c - n\log\left(\frac{\theta_0}{\theta_a}\right)\right]\left(\frac{\theta_a-\theta_0}{\theta_0\theta_a}\right)^{-1} \quad (8.11)$$

Por consiguiente, todo se reduce a calcular el valor del estadístico $\sum_{i=1}^n x_i$ y compararlo con la constante, k, dada por el lado derecho de (8.11). Dicha k se calcula de modo que $\sum_{i=1}^n X_i > k$ bajo H_0 con la probabilidad α que hayamos prefijado. En el caso que nos ocupa, $\sum_{i=1}^n X_i$ sigue bajo H_0 una distribución $\gamma(\theta_0^{-1}, n)$, y k resulta de resolver

$$\int_k^\infty \frac{1}{\Gamma(n)\theta_0^n} e^{-x/\theta_0} x^{n-1} dx = \alpha.$$

Por tanto, k no depende de cuál sea θ_a (con tal de que $\theta_a > \theta_0$) y el contraste es uniformemente más potente en la clase indicada.

Hay una caracterización simple que permite detectar la existencia de contrastes UMP cuando existen. Requiere la siguiente definición.

Definición 8.1 *Sea X una v.a. con distribución $\{F_x(x;\theta), \theta \in \Theta\}$. Sea $f_{\boldsymbol{X}|\theta}(\boldsymbol{x}\,|\theta)$ la función de verosimilitud asociada a una muestra $\boldsymbol{x} = (x_1,\ldots,x_n)$. Se dice que $\{F_x(x;\theta), \theta \in \Theta\}$ tiene* razón de verosimilitud monótona *si para algún estadístico $T(\boldsymbol{x})$ y cualquier $\boldsymbol{x}$ se verifica*

$$\frac{f_{\boldsymbol{X}|\theta}(\boldsymbol{x}\,|\theta)}{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)} = g(T(\boldsymbol{x})), \quad (8.12)$$

siendo $g(.)$ una función monótona no decreciente y θ_0, θ valores cualesquiera en Θ con $\theta > \theta_0$.

Ejemplo 8.4 El Ejemplo 8.3 muestra una familia de distribuciones con una razón de verosimilitud monótona. Si hacemos $T(\boldsymbol{x}) = \sum_{i=1}^n x_i$, tenemos que

$$\frac{f_{\boldsymbol{X}|\theta}(\boldsymbol{x}\,|\theta)}{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)} \propto \exp\left\{T(\boldsymbol{x})\left(\frac{\theta-\theta_0}{\theta_0\theta}\right)\right\},$$

que es una función creciente de $T(\boldsymbol{x})$ para cualesquiera $\theta, \theta_0 \in \Theta$ con $\theta > \theta_0$.

Se deduce con facilidad de (8.12) que si una familia de distribuciones tiene razón de verosimilitud monótona,

$$\frac{f_{\boldsymbol{X}\,|\theta}(\boldsymbol{x}\,|\theta)}{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)} \geq c \iff g(T(\boldsymbol{x})) \geq c \Longleftrightarrow T(\boldsymbol{x}) \geq g^{-1}(c).$$

Por tanto, el contraste más potente que proporciona el Teorema de Neyman–Pearson es independiente de la alternativa dentro de la familia considerada: es UMP y puede construirse haciendo uso del estadístico $T(\boldsymbol{x})$.

Por otra parte, es fácil identificar $T(\boldsymbol{x})$ en las distribuciones de la familia exponencial cuando existe un contraste UMP. En efecto, sea $\theta > \theta_0$; para cualquier distribución en la familia exponencial,

$$\begin{aligned}\frac{f_{\boldsymbol{X}\,|\theta}(\boldsymbol{x}\,|\theta)}{f_{\boldsymbol{X}}(\boldsymbol{x};\theta_0)} &= \frac{\exp\left\{a(\theta)\sum_{i=1}^n b(x_i) + c(\theta) + \sum_{i=1}^n d(x_i)\right\}}{\exp\left\{a(\theta_0)\sum_{i=1}^n b(x_i) + c(\theta_0) + \sum_{i=1}^n d(x_i)\right\}} \\ &= \exp\left\{(a(\theta) - a(\theta_0))\sum_{i=1}^n b(x_i) + (c(\theta) - c(\theta_0))\right\}.\end{aligned}$$

Por consiguiente, si $a(\theta)$ es función no decreciente de θ, la distribución considerada tiene razón de verosimilitud monótona, y admite un contraste UMP que puede expresarse en función del estadístico suficiente $T(\boldsymbol{x}) = \sum_{i=1}^n b(x_i)$.

8.5. Contrastes razón de verosimilitudes generalizada.

Con frecuencia tenemos hipótesis anidadas, del tipo: $H_0 : \theta \in \Theta_0$ versus $H_a : \theta \in \Theta_a$, en que $\Theta_a = \Theta - \Theta_0$; es decir, la hipótesis nula prescribe que θ toma valores en un subconjunto propio de Θ. Típicamente, H_0 constriñe θ a un subconjunto de dimensión menor que la de Θ.

Cuando esto ocurre, bajo condiciones de regularidad que hagan el estimador MV de θ asintóticamente insesgado y normal, el resultado a continuación permite construir contrastes que son en ocasiones los únicos disponibles.

Teorema 8.2 *Sea el contraste $H_0 : \theta \in \Theta_0$ versus $H_a : \theta \in \Theta_a$, en que $\Theta_a = \Theta - \Theta_0$, y supongamos que dim$(\Theta_a) = r$. Bajo condiciones de regularidad como las requeridas en el Teorema 6.2, pág, 81,*

$$\Lambda = -2\log_e\left(\frac{\sup_{\theta\in\Theta_0} f_{\boldsymbol{X}\,|\theta}(\boldsymbol{x}\,|\theta)}{\sup_{\theta\in\Theta} f_{\boldsymbol{X}\,|\theta}(\boldsymbol{x}\,|\theta)}\right) \sim \chi^2_r. \tag{8.13}$$

DEMOSTRACION:

Presentamos, por simplicidad, la demostración para el caso unidimensional en que la hipótesis nula es simple, $H_0 : \theta = \theta_0$, en tanto la alternativa es $H_a : \theta \in \Theta$ con $\dim(\Theta) = 1$ (y, por tanto, $r = \dim(\Theta) - \dim(\theta_0) = 1$). Sean

$$\hat{\theta} = \sup_{\theta \in \Theta} f_{\boldsymbol{X}|\theta}(\boldsymbol{x}\,|\theta), \tag{8.14}$$

$$U_i(\theta) = \frac{\partial \log_e f_X(X_i, \theta)}{\partial \theta}, \tag{8.15}$$

Tenemos que

$$\Lambda = 2\left[\log_e f_{\boldsymbol{X}}(\boldsymbol{X}; \hat{\theta}) - \log_e f_{\boldsymbol{X}}(\boldsymbol{X}; \theta_0)\right]. \tag{8.16}$$

Desarrollando en serie el segundo sumando de la derecha de (8.16) en torno al punto $\hat{\theta}$ obtenemos

$$\begin{aligned} \log_e f_{\boldsymbol{X}}(\boldsymbol{X}; \theta_0) &= \log_e f_{\boldsymbol{X}}(\boldsymbol{X}; \hat{\theta}) + \left[\frac{\partial \log_e f_{\boldsymbol{X}}(\boldsymbol{X}; \theta)}{\partial \theta}\right]_{\theta=\hat{\theta}} (\theta_0 - \hat{\theta}) \\ &\quad + \frac{1}{2!}\left[\frac{\partial^2 \log_e f_{\boldsymbol{X}}(\boldsymbol{X}; \theta)}{\partial \theta^2}\right]_{\theta=\tilde{\theta}} (\theta_0 - \hat{\theta})^2 \end{aligned} \tag{8.17}$$

en que $\tilde{\theta}$ es un punto entre θ_0 y $\hat{\theta}$, es decir, $|\tilde{\theta} - \theta_0| < |\hat{\theta} - \theta_0|$. Sustituyendo (8.17) en (8.16) obtenemos

$$\begin{aligned} \Lambda &= -2\left[\frac{\partial \log_e f_{\boldsymbol{X}}(\boldsymbol{X}; \theta)}{\partial \theta}\right]_{\theta=\hat{\theta}} (\theta_0 - \hat{\theta}) \\ &\quad - \left[\frac{\partial^2 \log_e f_{\boldsymbol{X}}(\boldsymbol{X}; \theta)}{\partial \theta^2}\right]_{\theta=\tilde{\theta}} (\hat{\theta} - \theta_0)^2 \qquad (8.18) \\ &= -2(\theta_0 - \hat{\theta}) \sum_{i=1}^{n} U_i(\hat{\theta}) - (\hat{\theta} - \theta_0)^2 \sum_{i=1}^{n} U_i'(\tilde{\theta}) \qquad (8.19) \end{aligned}$$

Ahora bien, bajo las condiciones de regularidad impuestas, el estimador máximo verosímil anula la primera derivada de la función de verosimilitud, y

$$\sum_{i=1}^{n} U_i(\hat{\theta}) = \left[\frac{\partial \log_e f_{\boldsymbol{X}}(\boldsymbol{X}; \theta)}{\partial \theta}\right]_{\theta=\hat{\theta}} = 0;$$

por tanto, (8.19) queda reducida a

$$\Lambda = -(\hat{\theta} - \theta_0)^2 \left(\sum_{i=1}^{n} U_i'(\tilde{\theta})\right) = n(\hat{\theta} - \theta_0)^2 \left(-\frac{\sum_{i=1}^{n} U_i'(\tilde{\theta})}{n}\right). \tag{8.20}$$

En virtud del Teorema 6.2,

$$n(\hat{\theta} - \theta_0)^2 \xrightarrow{\mathcal{L}} I(\theta_0)^{-1} \times \chi_1^2. \tag{8.21}$$

Por otra parte, $\tilde{\theta} \xrightarrow{c.s.} \theta_0$ (ya que $\hat{\theta} \xrightarrow{c.s.} \theta_0$ y $|\tilde{\theta} - \theta_0| < |\hat{\theta} - \theta_0|$), y por consiguiente

$$-n^{-1} \sum_{j=1}^{n} U_j'(\tilde{\theta}) \quad \xrightarrow{p} \quad -n^{-1} \sum_{j=1}^{n} U_j'(\theta_0). \tag{8.22}$$

La expresión (8.22) converge en probabilidad al valor medio de cada uno de los sumando promediados, $E_{\theta_0}[-U_j'(\theta_0)] = I(\theta_0)$, en virtud de la ley débil de los grandes números (Teorema A.2, pág. 148):

$$-\frac{\sum_{i=1}^{n} U_i'(\tilde{\theta})}{n} \quad \xrightarrow{p} \quad I(\theta_0). \tag{8.23}$$

Haciendo uso de (8.21) y (8.23) vemos que la expresión (8.20) converge en distribución a una χ_1^2.

■

Observación 8.4 *(criterio AIC y verosimilitudes penalizadas)* Incidentalmente, hay una conexión interesante entre el contraste razón de verosimilitudes generalizada y el criterio conocido como AIC (An Information Criterion, o Akaike's Information Criterion).

Supongamos que deseamos comparar modelos con diferente número de parámetros. Consideremos, por ejemplo, uno cuyo vector de parámetros $\boldsymbol{\theta}$ pertenece a Θ, y otro competidor tal que $\boldsymbol{\theta} \in \Theta_0$ con $\Theta_0 \subset \Theta$ y $\dim(\Theta) - \dim(\Theta_0) = r$. Del Teorema 8.2 deducimos que, bajo H_0,

$$2 \log_e \left(\frac{\sup_{\theta \in \Theta} f_{\boldsymbol{X}|\theta}(\boldsymbol{x}\,|\theta)}{\sup_{\theta \in \Theta_0} f_{\boldsymbol{X}|\theta}(\boldsymbol{x}\,|\theta)} \right) \sim \chi_r^2. \tag{8.24}$$

Numerador y denominador de (8.24) son las verosimilitudes maximizadas bajo H_a y bajo H_0 respectivamente. Dado que $\Theta_0 \subset \Theta$, es claro que la verosimilitud bajo H_0 nunca será mayor: no tiene pues sentido una comparación directa de ambas verosimilitudes para escoger entre ambos modelos. Si tomamos valor medio en (8.24) y dividimos entre dos vemos que, bajo H_0,

$$E\left[\log_e \sup_{\theta \in \Theta} f_{\boldsymbol{X}|\theta}(\boldsymbol{x}\,|\theta) - \log_e \sup_{\theta \in \Theta_0} f_{\boldsymbol{X}|\theta}(\boldsymbol{x}\,|\theta) \right] = \frac{r}{2}. \tag{8.25}$$

Es decir, *incluso cuando H_0 es cierta y no tiene objeto seleccionar el modelo alternativo con $\theta \in \Theta - \Theta_0$*, la verosimilitud de dicho modelo alternativo será en promedio $\frac{r}{2}$ unidades mayor, siendo r la diferencia de dimensión entre Θ y Θ_0 (normalmente coincidente con la diferencia en el número de parámetros ajustados). Podría parecer adecuado corregir las verosimilitudes correspondientes a modelos diferentes, restando al logaritmo de cada una la mitad del número de parámetros utilizado, $\frac{r}{2}$. Ello las pondría "en pie de igualdad", rectificando en valor medio el incremento de verosimilitud que se produce por el mero hecho de ajustar un mayor número de parámetros.

Así, en lugar de logaritmos de verosimilitudes, compararíamos logaritmos de verosimilitudes corregidos en valor medio como

$$\log_e f_{\boldsymbol{X}}(\boldsymbol{x}, \hat{\theta}_{\text{MV}} \in \Theta_0) \quad - \quad \frac{r_1}{2} \tag{8.26}$$

$$\log_e f_{\boldsymbol{X}}(\boldsymbol{x}, \hat{\theta}_{\text{MV}} \in \Theta) \quad - \quad \frac{r_2}{2}. \tag{8.27}$$

No obstante, preferir el segundo modelo al primero sobre la base de que

$$\log_e f_{\boldsymbol{X}}(\boldsymbol{x}, \hat{\theta}_{\text{MV}} \in \Theta) - \frac{r_1}{2} \quad > \quad \log_e f_{\boldsymbol{X}}(\boldsymbol{x}, \hat{\theta}_{\text{MV}} \in \Theta_0) - \frac{r_2}{2},$$

o, equivalentemente,

$$2\left(\log_e \frac{f_{\boldsymbol{X}}(\boldsymbol{x}, \hat{\theta}_{\text{MV}} \in \Theta)}{f_{\boldsymbol{X}}(\boldsymbol{x}, \hat{\theta}_{\text{MV}} \in \Theta_0)}\right) \quad > \quad (r_1 - r_2),$$

es tanto como hacer un contraste de hipótesis de uno frente a otro tomando como valor crítico de una $\chi^2_{r_1-r_2}$ su valor medio. Ello daría lugar a un α (error de tipo I) inaceptablemente grande. Parece que se impone una penalización mayor del número de parámetros.

La expresión,

$$2\log_e f_{\boldsymbol{X}}(\boldsymbol{x}, \hat{\theta}_{\text{MV}}) - 2r$$

siendo r el número de parámetros libres en θ que hemos ajustado se conoce como criterio AIC y fue propuesto en Akaike (1972), haciendo uso de un argumento diferente. Obsérvese que penaliza adicionalmente la verosimilitud respecto de la propuesta en (8.26)–(8.27). Discrimina con ello más a favor de modelos "simples." Es sólo una de las muchas manifestaciones de una idea bastante más general: la de penalizar las verosimilitudes de modo que se tome en consideración su diferente "complejidad", medida de ordinario por el número de parámetros ajustados o alguna función del número de parámetros y el tamaño de la muestra. Sobre esta cuestión volvemos en el Capítulo 9.

8.6. Contrastes de significación puros

8.6.1. Caso de hipótesis simples

En ocasiones, deseamos contrastar una hipótesis sin especificar una alternativa. Típicamente, la hipótesis H_0 que se desea contrastar puede describirse como una "hipótesis *statu quo*" o comúnmente aceptada, que queremos poner a prueba. no tenemos una idea clara de cuales puedan ser las alternativas competidoras. Deseamos simplemente examinar si la evidencia muestral es compatible con H_0.

Los ingredientes necesarios para un contraste de esta naturaleza son:

- La hipótesis nula de interés, H_0.
- Un estadístico $T(\boldsymbol{X})$ cuya distribución bajo H_0 es conocida, y sobre el que adoptaremos la convención de que valores mayores suponen un mayor alejamiento de la muestra del comportamiento esperable bajo H_0.

Procederemos entonces del modo habitual:

1. Realizaremos el muestreo, obteniendo $\boldsymbol{x}$.

2. Calcularemos el valor del estadístico de contraste, $T(\boldsymbol{X})$, correspondiente a la muestra $\boldsymbol{x}$. Sea dicho valor $t_{\text{obs}} = T(\boldsymbol{x})$.

3. Calcularemos,
$$p_{\text{obs}} = \text{Prob}\left\{T(\boldsymbol{X}) \geq t_{\text{obs}} | H_0\right\}, \tag{8.28}$$
nivel de significación empírico o *p-value*. Para un nivel de significación (probabilidad de error de tipo I) prefijado, α, rechazaremos H_0 si $p_{\text{obs}} < \alpha$, y no rechazaremos en caso contrario.

Podemos interpretar p_{obs} como la probabilidad cuando H_0 es cierta de obtener una muestra tan o más "rara" que la obtenida. En efecto, valores crecientes de $T(\boldsymbol{x})$ reflejan discrepancias crecientes de la muestra con el comportamiento previsible bajo H_0. La lógica del contraste de significación consiste pues en rechazar H_0 cuando lo que observamos sería "excesivamente raro" en una situación en que H_0 prevaleciera.

Ejemplo 8.5 El contraste de ajuste χ^2 es posiblemente el de más uso (y abuso) de entre todos los contrastes de significación puros. Si particionamos los valores obtenibles de la variable aleatoria en k clases,
$$T(\boldsymbol{X}) = \frac{\sum_{i=1}^{k}(n_i - e_i)^2}{e_i}, \tag{8.29}$$
siendo n_i el número de observaciones en la clase i-ésima, y e_i el número de observaciones que esperaríamos obtener en dicha clase bajo H_0 (véase por ej. Trocóniz (1987), p. 245). Valores grandes de $T(\boldsymbol{X})$ corresponden a discrepancias notables en una o varias clases entre el número de observaciones esperado y el que se ha presentado en la muestra.

Si H_0 especifica por completo una distribución, $T(\boldsymbol{X})$ se distribuye aproximadamente (para muestras grandes y clases no muy despobladas) como una χ^2_{k-1}.

Obsérvese que estamos contrastando acuerdo de la muestra con H_0 sin especificar ninguna alternativa, es decir, sin precisar en qué modo habría de presentarse, de existir, la discrepancia entre la muestra y la distribución prescrita por H_0.

Otros muchos ejemplos pueden darse de contrastes de significación puros: el contraste de ajuste de Kolmogorov-Smirnov (véase Trocóniz (1987), p. 255), contrastes de independencia, etc.

8.6.2. Caso de hipótesis compuestas

El problema se hace un poco más complejo cuando la hipótesis de interés no es simple sino compuesta; es decir, H_0 no especifica por completo la distribución de la que supuestamente procede la muestra.

Ejemplo 8.6 La hipótesis de normalidad sería compuesta: no hay una única distribución normal, sino una familia de ellas.

Cuando esto ocurre, el modo tan simple de operar descrito más arriba ya no es de aplicación. Podemos quizá encontrar todavía un estadístico $T(\boldsymbol{X})$ que sea buen indicador de la discrepancia entre el comportamiento de la muestra y el esperable bajo H_0. El cálculo de p_{obs} ya no es en general, sin embargo, tan simple como el mostrado en (8.28). Puede ocurrir que la probabilidad en el lado derecho de (8.28 sea diferente, dependiendo de la distribución concreta que consideremos de entre todas las que componen H_0.

En general, las hipótesis compuestas suelen prescribir una familia de distribuciones indeterminadas en el valor de uno o varios *parámetros de ruido*. Así, en el Ejemplo 8.6, H_0 prescribía para la muestra una distribución $N(\mu, \sigma^2)$ para valores indeterminados de μ y σ. Cuando esto ocurre, hay varias soluciones que podemos adoptar para realizar el contraste de significación deseado.

1. Estimar el o los parámetros de ruido. Esto es tanto como convertir la hipótesis compuesta en una simple "similar", individualizando una única distribución de entre todas las que componen H_0.

   Ejemplo 8.7 Supongamos que deseamos contrastar la hipótesis de que una determinada muestra procede del muestreo de una distribución de Poisson, $\mathcal{P}(\lambda)$. Podríamos estimar λ por $\hat{\lambda} = N^{-1}\sum_i X_i$ y contrastar la hipótesis simple resultante.

   Hay que tener presente que, al estimar el o los parámetros haciendo uso de la muestra, estamos seleccionando de entre todas las distribuciones que componen H_0 una particularmente "cercana" a los datos analizados. Este efecto deberá de ordinario tenerse en cuenta en la obtención de la distribución del estadístico de contraste $T(\boldsymbol{X})$. Si hacemos uso de un contraste χ^2 como el descrito en el Ejemplo 8.5, deberemos ahora comparar el valor t_{obs} con los cuantiles de una χ^2_{k-2}; el grado de libertad perdido en la χ^2 recoge el hecho de que la distribución $\mathcal{P}(\hat{\lambda})$ es "la más cercana" a los datos de entre todas las $\mathcal{P}(\lambda)$, y por este motivo debemos esperar que el valor de $T(\boldsymbol{X})$ sea en promedio menor que si λ fuera un valor previamente fijado sin hacer uso de la muestra.

   Observación 8.5 Puede formalizarse la expresión "la más cercana" empleada en el ejemplo anterior. Si el procedimiento de estimación del o los parámetros de ruido es el de máxima verosimilitud, la distribución seleccionada de entre la familia que componen H_0 es la que está a mínima distancia de Kullback-Leibler de la distribución empírica de la muestra.

   Ejemplo 8.8 *(contraste de normalidad)* Para hacer un contraste de normalidad —sin especificar la distribución normal concreta—, podríamos estimar μ y σ y emplear un contraste de ajuste de Kolmogorov-Smirnov. Compararíamos así la distribución empírica de la muestra con la de una $N(\hat{\mu}, \hat{\sigma})$. Siendo el de Kolmogorov-Smirnov un contraste de

naturaleza asintótica, que se realiza con muestras de tamaño bastante grande, podríamos en general prescindir del hecho de que hemos estimado dos parámetros.

Lo que antecede es una ilustración y no un modo aconsejado de operar: hay contrastes especializados como el de d'Agostino (véase D'Agostino (1971)) o el de Shapiro-Wilk (véase Shapiro y Francia (1972) por ejemplo).

2. Podemos en algunos casos convertir la hipótesis compuesta en simple de un modo *ad hoc*, como ilustra el ejemplo siguiente.

   Ejemplo 8.9 Consideremos el caso en que $X \sim N(\mu, \sigma_0)$ y deseamos contrastar $H_0 : \mu \leq \mu_0$ con σ_0 conocida. Un estadístico adecuado sería $T(\boldsymbol{X}) = \overline{X}$, conduciendo al rechazo de H_0 valores convenientemente "grandes".

   Necesitamos individualizar una entre todas las distribuciones en $\{N(\mu, \sigma_0)\}$ para hacer el cálculo de p_{obs}:

   $$p_{\text{obs}} = \text{Prob}\left\{T(\boldsymbol{X}) \geq t_{\text{obs}} | H_0\right\}; \tag{8.30}$$

   tiene sentido entonces calcular p_{obs} así:

   $$p_{\text{obs}} = \text{Prob}\left\{T(\boldsymbol{X}) \geq t_{\text{obs}} | N(\mu_0, \sigma_0)\right\}. \tag{8.31}$$

   Hemos escogido la distribución en la familia H_0 más extrema. La lógica de hacerlo así es que el p_{obs} calculado bajo dicha distribución es el máximo de los que calcularíamos bajo cualquiera de las que componen H_0. Estamos así actuando de manera conservadora. La probabilidad de obtener bajo H_0 una muestra tan o más "rara" que la observada será *como máximo* p_{obs}. Si p_{obs} es convenientemente pequeño, podemos rechazar confiadamente H_0.

3. Hay una tercera opción, que cuando es factible es frecuentemente la preferida. En lugar de estimar los parámetros de ruido, podemos eliminarlos considerando la distribución condicional sobre un estadístico suficiente para los mismos. El ejemplo que sigue ilustra el modo de operar.

   Ejemplo 8.10 Estamos interesados en contrastar ajuste a una distribución de Poisson $\mathcal{P}(\lambda)$, sin precisar λ. Disponemos de una muestra $\boldsymbol{X} = (X_1, \ldots, X_n)$. Sabemos (ver Ejemplo 3.8, p. 36) que $S = \sum_{i=1}^n X_i$ es un estadístico suficiente para λ, y que la distribución condicionada es

   $$f_{\boldsymbol{X}|S}(\boldsymbol{x}|s) = \frac{s!}{n^s \prod_{i=1}^n x_i!}. \tag{8.32}$$

   Por consiguiente, condicionalmente en el valor observado s del estadístico suficiente, una muestra como la obtenido tiene una probabilidad dada por el lado derecho de (8.32; llamémosle π. Podemos computar p_{obs} como la probabilidad de encontrar, dado $S = s$, una muestra tan o más rara que la obtenida:

   $$p_{\text{obs}} = \sum_{\boldsymbol{x} \in \mathcal{C}(s)} \frac{s!}{n^s \prod_{i=1}^n x_i!}, \tag{8.33}$$

siendo

$$\mathcal{C}(s) = \left\{ \boldsymbol{x} : \frac{s!}{n^s \prod_{i=1}^n x_i!} \leq \pi \right\}.$$

El problema de contrastar si la muestra dada procede de una $\mathcal{P}(\lambda)$ con λ indeterminado, ha quedado convertido en el problema de contrastar si es plausible que la muestra obtenida $\boldsymbol{x}$ proceda de una distribución multinomial de parámetros $(\frac{1}{n}, \ldots, \frac{1}{n})$.

Ejemplo 8.11 *(contraste exacto de Fisher)* Un caso de gran aplicación (y que ya fue discutido por Fisher) es aquél en que estamos interesados en contrastar la independencia entre dos caracteres. Por ejemplo, si deseáramos contrastar la efectividad de un cierto tratamiento preventivo, podríamos administrarlo a un grupo de pacientes en tanto otros homogéneos reciben un placebo. Tras un periodo de tiempo, podríamos ver cuantos enfermaron de uno y otro grupo y compilar una tabla como la siguiente (c_1, c_2, r_1, r_2 son los totales de filas y columnas respectivamente):

	Sano	Enfermo	
Placebo	n_{11}	n_{12}	r_1
Tratamiento	n_{21}	n_{22}	r_2
	c_1	c_2	

A la vista de la misma, desearíamos contrastar independencia entre los sucesos "Tomar el tratamiento" y "Mantenerse sano".

Bajo la hipótesis de independencia entre ambos caracteres, la probabilidad de estar en la casilla ij es $p_{ij} = p_{i.}p_{.j}$, siendo $p_{i.}$ y $p_{.j}$ las probabilidades marginales de estar en la fila i y en la columna j. Las probabilidades de cada casilla bajo la hipótesis de independencia dependen exclusivamente de las probabilidades marginales y c_1, c_2, r_1, r_2 son estadísticos suficientes para las mismas (se comprueba fácilmente). La distribución condicionada sobre c_1, c_2, r_1, r_2 de un resultado como el recogido en la tabla es, bajo independencia, independiente de los parámetros: puede comprobarse (ver el desarrollo en, por ejemplo, Garín y Tusell (1991), ejercicio 6.16) que dicha probabilidad es

$$p' = \frac{\binom{c_1}{n_{11}}\binom{c_2}{n_{12}}}{\binom{n}{r_1}}.$$

Podemos ahora considerar la clase Δ formada por todas las tablas t que pueden construirse respetando los márgenes c_1, c_2, r_1, r_2 y tienen una probabilidad condicional menor que p', y obtener el nivel de significación empírico así: $p_{\text{obs}} = \sum_{t \in \Delta} \text{Prob}\,\{t\}$.

8.6.3. Hay que tener en cuenta que...

Los contrastes de significación tienen algunas peculiaridades que es preciso considerar.

1. Los contrastes de significación evalúan el acuerdo entre una muestra y una determinada hipótesis nula, H_0. No se explicita la alternativa, y ello puede dar lugar a resultados absurdos por falta de cuidado al interpretar los resultados. En particular, una muestra puede ser extremadamente "rara" bajo H_0, y aún serlo más bajo cualquiera de las situaciones que podamos considerar como alternativas. En este caso, es necesario tomar en cuenta explícitamente estas alternativas en el proceso de decisión.

   Ejemplo 8.12 Si hubiéramos de contrastar la hipótesis $H_0 : X \sim N(0, \sigma^2 = 1)$ frente a toda alternativa, y contamos con 100 observaciones, parece sensato computar como estadístico de contraste $\overline{X}$ y rechazar H_0 cuando $\overline{X}$ no esté incluido en el intervalo $(-1{,}96/\sqrt{100}, 1{,}96/\sqrt{100})$; esto daría lugar a una prueba con un $\alpha = 0{,}05$. Si, sin embargo, la naturaleza del problema sugiriera que las únicas alternativas posibles son distribuciones normales con varianza unitaria y media mayor que 5, sería claramente inadecuado rechazar H_0 con un valor, por ejemplo, de $\overline{X} = 2$. Tal valor sería extremadamente raro bajo H_0 —estaría a veinte desviaciones típicas de la media—, y sugeriría su rechazo; ¡pero aún sería más raro bajo cualquiera de las alternativas! *Aún cuando un contraste de significación no requiera la fijación de alternativas, debemos estar vigilantes ante situaciones como la descrita, que sugieren una insuficiente consideración de los estados de naturaleza posibles.*

2. En el caso de contrastes de significación es particularmente importante distinguir entre *significación estadística* y relevancia práctica de la discrepancia con H_0 que el contraste pone de manifiesto. Sobre esta cuestión puede verse Wang (1993), Cap. 1. El siguiente ejemplo ilustra la naturaleza del problema.

   Ejemplo 8.13 Consideremos de nuevo la situación en el Ejemplo 8.12. A efectos prácticos, puede acontecer que sea indiferente el que la media sea $\epsilon = 10^{-8}$ en lugar de exactamente cero. No obstante, incluso una diferencia tan minúscula sería declarada significativa con probabilidad tan cercana a uno como deseáramos si el tamaño muestral crece lo suficiente. En efecto, si adoptamos una región crítica como $(-t_{\alpha/2}/\sqrt{n}, +t_{\alpha/2}/\sqrt{n})^c$, un n lo suficientemente grande hará que $|t_{\alpha/2}/\sqrt{n}| < \epsilon$, conduciendo por tanto al rechazo de H_0 al nivel de significación α.

   Pensemos ahora que todo modelo es, en la práctica, una aproximación útil, pero no exacta. ¡Si fuéramos estrictos en rechazar un modelo al obtener un resultado estadísticamente significativo contra él, *todo* modelo sucumbiría ante una acumulación suficiente de evidencia! Esto es claramente absurdo. Deberíamos más bien preguntarnos si una media de ϵ representa a efectos prácticos una desviación suficiente de una media cero como para justificar el rechazo de esta última hipótesis. Sólo en caso de que la respuesta sea afirmativa estaría indicado un contraste estadístico.

3. Una peculiaridad de los contrastes de significación es que la misma evidencia puede dar lugar a interpretaciones diferentes según el procedimiento de muestreo. El siguiente ejemplo lo ilustra.

   Ejemplo 8.14 Consideremos una moneda cuya regularidad (H_0 : Prob $\{\text{Cara}\}$ = Prob $\{\text{Cruz}\}$) deseamos contrastar. Podemos lanzar cinco veces una moneda y contar el número de "caras" (Experimento 1) o lanzar la moneda hasta obtener una "cruz" y examinar el número total de lanzamientos (Experimento 2). Imaginemos dos experimentadores, haciendo el primero el Experimento 1 y el segundo el Experimento 2. Imaginemos que ambos obtienen cuatro "caras" al comienzo y una "cruz" en el quinto lanzamiento.

   Tanto uno como otro se inclinarían a considerar el resultado como evidencia de mayor probabilidad de "cara", pero aquí acabaría el acuerdo. El primero, computaría p_{obs} —la probabilidad de obtener un resultado tanto o más extremo que el obtenido así:

$$\begin{aligned} p_{\text{obs}} &= \text{Prob}\{4 \text{ caras}\} + \text{Prob}\{5 \text{ caras}\} \\ &= \binom{5}{4}\left(\frac{1}{2}\right)^5 + \binom{5}{5}\left(\frac{1}{2}\right)^5 \\ &= \frac{3}{16}. \end{aligned}$$

   El segundo, en cambio, calcularía:

$$\begin{aligned} p_{\text{obs}} &= \text{Prob}\{\text{Primera "cruz" en lugar quinto o posterior}\} \\ &= \left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^1 + \left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right) + \ldots + \left(\frac{1}{2}\right)^{n-1}\left(\frac{1}{2}\right) + \ldots \\ &= \frac{1}{32}\frac{1}{1-\frac{1}{2}} = \frac{1}{16}. \end{aligned}$$

   Ambos experimentadores han obtenido el mismo resultado, y sin embargo uno le otorga más peso que el otro. Es molesto que la interpretación que se hace de una misma evidencia dependa de *cosas que podrían haber ocurrido, pero no lo han hecho*.

   Observación 8.6 Relacionado con el ejemplo precedente: parecería sensato el cálculo anterior de p_{obs} si existiera alguna razón para suponer que la desviación de la regularidad de la moneda, de producirse, lo ha de ser hacía una mayor probabilidad de "cara". De no ser así, el experimentador que hace uso del Experimento 1 debería doblar su p_{obs}: hay también resultados más "raros" que el obtenido a causa de un anormalmente pequeño número de caras. *No es legítimo esperar a ver el resultado para decidir sobre qué tipo de desviaciones de H_0 queremos considerar,* y en consecuencia sobre el modo en que vamos a computar p_{obs}.

4. En ocasiones, se realizan varios contrastes de significación sobre la misma hipótesis, con muestras distintas y arrojando resultados p_{obs} que pueden verse como variables aleatorias independientes. Supongamos dos experimentos

que han arrojado sendos p^*_{obs} y p^{**}_{obs}. Siendo interpretables como probabilidades (de obtener una muestra tanto o más "rara" que la obtenida, cuando H_0 es cierta), podría pensarse en $p_{\text{obs}} = p^*_{\text{obs}} \times p^{**}_{\text{obs}}$ como un nivel de significación empírico sumarizando toda la evidencia disponible. Esto es incorrecto: véase Cox y Hinkley (1974), Cap. 4 y Garín y Tusell (1991), ejercicio 9.12.

8.7. Contrastes localmente más potentes

En ocasiones, la hipótesis alternativa es compuesta y no hay un contraste uniformemente mas potente. Una táctica que parece sensata podría ser maximizar la potencia frente a una alternativa "próxima". Por ejemplo, si tenemos $H_0 : \theta = \theta_0$ vs. $H_a : \theta > \theta_0$, podríamos plantearnos escoger el contrate que permitiera discriminar óptimamente entre H_0 y la alternativa simple "local" $H_{a'} : \theta = \theta + \delta$ para un δ pequeño.

De acuerdo con el teorema de Neyman-Pearson, la región crítica que da lugar al contraste más potente para un α prefijado, sería:

$$RC = \left\{ \boldsymbol{x} : \frac{f_{\boldsymbol{X}}(\boldsymbol{x}; \theta_0 + \delta)}{f_{\boldsymbol{X}}(\boldsymbol{x}; \theta_0)} \geq k_\alpha \right\}, \tag{8.34}$$

para algún k_α; o, equivalentemente,

$$RC = \{ \boldsymbol{x} : \log f_{\boldsymbol{X}}(\boldsymbol{x}; \theta_0 + \delta) - \log f_{\boldsymbol{X}}(\boldsymbol{x}; \theta_0) \geq c_\alpha \}. \tag{8.35}$$

Consideremos la variable aleatoria

$$\log f_{\boldsymbol{X}}(\boldsymbol{X}; \theta_0 + \delta) - \log f_{\boldsymbol{X}}(\boldsymbol{X}; \theta_0) \tag{8.36}$$

y desarrollemos en serie en torno al punto θ_0. Tenemos entonces que

$$\begin{aligned} \log f_{\boldsymbol{X}}(\boldsymbol{X}; \theta_0 + \delta) &- \log f_{\boldsymbol{X}}(\boldsymbol{X}; \theta_0) \\ &\cong \log f_{\boldsymbol{X}}(\boldsymbol{X}; \theta_0) + \delta \left(\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X}; \theta)}{\partial \theta} \right)_{\theta = \theta_0} - \log f_{\boldsymbol{X}}(\boldsymbol{X}; \theta_0) \\ &= \delta \left(\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X}; \theta)}{\partial \theta} \right)_{\theta = \theta_0}; \end{aligned}$$

los términos despreciados en el desarrollo en serie son de orden δ^2 y superior, y por tanto despreciables frente al único incluido cuando δ es muy pequeño. Cuando la hipótesis nula es cierta, tenemos (en virtud del Lema 5.1 y (5.5) que

$$E_{\theta_0} \left[\delta \left(\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X}; \theta)}{\partial \theta} \right)_{\theta = \theta_0} \right] = 0 \tag{8.37}$$

$$\text{Var} \left[\delta \left(\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X}; \theta)}{\partial \theta} \right)_{\theta = \theta_0} \right] = \delta^2 E_{\theta_0} \left(\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X}; \theta)}{\partial \theta} \right)^2_{\theta = \theta_0} \tag{8.38}$$

$$= n\delta^2 I(\theta_0). \tag{8.39}$$

Por consiguiente,

$$\left| \frac{\delta \left(\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial \theta} \right)_{\theta=\theta_0}}{\sqrt{n\delta^2 I(\theta_0)}} \right| = \left| (nI(\theta_0))^{-\frac{1}{2}} \left(\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial \theta} \right)_{\theta=\theta_0} \right| \tag{8.40}$$

es una variable aleatoria tipificada que podemos emplear como estadístico de contraste si conocemos su distribución. Esta última puede ser desconocida, pero para n grande, teniendo en cuenta que cuando tenemos observaciones independientes e idénticamente distribuidas

$$\log f_{\boldsymbol{X}}(\boldsymbol{X};\theta) = \log \prod_{i=1}^{n} f_X(X_i;\theta) = \sum_{i=1}^{n} \log f_X(X_i;\theta), \tag{8.41}$$

cabrá esperar un fuerte efecto teorema central del límite, y una distribución de (8.40) aproximadamente normal. Rechazaremos pues la hipótesis nula si

$$\left| (nI(\theta_0))^{-\frac{1}{2}} \left(\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial \theta} \right)_{\theta=\theta_0} \right| > z_{\alpha/2}, \tag{8.42}$$

siendo $z_{\alpha/2}$ el cuantil adecuado de una distribución $N(0,1)$. Alternativamente podríamos comparar el cuadrado de (8.42) con el cuantil $\chi^2_{1;\alpha}$.

En el caso en que hay varios parámetros, hemos de sustituir θ por $\boldsymbol{\theta}$ y modificar consecuentemente el desarrollo anterior; las ideas son las mismas. El resultado es también similar: si hay k parámetros libres en $\boldsymbol{\theta}$, tenemos que bajo H_0, asintóticamente

$$U(\boldsymbol{\theta_0})'(nI(\boldsymbol{\theta_0}))^{-1}U(\boldsymbol{\theta_0}) \sim \chi^2_k, \tag{8.43}$$

en que

$$U(\boldsymbol{\theta_0})' = \left(\frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial \boldsymbol{\theta_{(1)}}}, \ldots, \frac{\partial \log f_{\boldsymbol{X}}(\boldsymbol{X};\theta)}{\partial \boldsymbol{\theta_{(k)}}} \right) \tag{8.44}$$

y $\theta_{(i)}$ es la i-ésima componente de $\boldsymbol{\theta}$. Se conoce a este contraste como *score test*, o también como contraste multiplicador de Lagrange.

A la vista de (8.37) y (8.39) podríamos pensar también en contrastes haciendo uso de:

$$(\hat{\theta} - \boldsymbol{\theta_0})'(nI(\theta_0))^{-1}(\hat{\theta} - \boldsymbol{\theta_0}) \overset{H_0}{\sim} \chi^2_k \tag{8.45}$$

$$(\hat{\theta} - \boldsymbol{\theta_0})'(nI(\hat{\theta}))^{-1}(\hat{\theta} - \boldsymbol{\theta_0}) \overset{H_0}{\sim} \chi^2_k; \tag{8.46}$$

ambas son versiones asintóticamente equivalentes del contraste de Wald. Véase Garthwaite et al. (1995), p. 89.

Capítulo 9

Máxima verosimilitud, complejidad y selección de modelos

9.1. Introducción

William de Ockham (1290?–1349?) propuso como criterio para seleccionar lo que hoy llamaríamos modelos el prescindir de complicaciones innecesarias; el «no multiplicar las entidades sin necesidad.» Entre dos posibles explicaciones de un mismo fenómeno, Ockham sugería así que retuviéramos la más simple. Un principio que se ha popularizado como «la navaja de Ockham.»

Es difícil —tal vez imposible— justificar tal recomendación si pretendemos hacerlo con rigor. Se puede ver como una regla de economía intelectual. Pero ha de ser la adecuación entre modelo[1] y realidad lo que guíe nuestro esfuerzo, si somos realistas; no nuestra comodidad intelectual. ¿Por qué hemos de preferir explicaciones simples si el mundo real, en muchas de sus manifestaciones, parece extremadamente complejo?

Quizá la mejor línea de defensa argumental de la recomendación de Ockham pueda basarse en su extraordinario éxito. La búsqueda de explicaciones «simples» ha sido un criterio que ha guiado la perspicacia de los científicos casi invariablemente hacia «buenos» modelos: modelos con relativa gran capacidad explicativa

[1] Siendo acaso muy impreciso con el lenguaje, utilizo «modelo» para designar un mecanismo formalizable en ecuaciones matemáticas que suponemos «explica» un fenómeno.

que frecuentemente se funden armoniosamente con otros en unificaciones progresivamente mejores. Esto ha sucedido en Física y también en otras disciplinas.

Pero ¿qué es simple? Porque para seguir el consejo de Ockham necesitamos saber cuando uno de dos modelos es más simple que otro.

Hay casos en los que hay poca duda. Entre dos modelos que proporcionen predicciones igualmente buenas, si uno hace uso de todos los supuestos de otro y alguno adicional, preferiremos el primero. Hablaremos en tal caso de modelos anidados.

Pero esto es la excepción y no la regla. Más bien se nos presenta con frecuencia el caso de modelos «solapados» o incluso aparentemente «disjuntos.» Se hace mucho más difícil en este caso decidir cuál es el más simple. Y el problema sólo puede complicarse cuando tenemos modelos estadísticos que ofrecen un grado diferente de explicación o ajuste de la evidencia empírica. ¿Qué debemos preferir: un modelo muy simple, que sólo imprecisamente parece dar cuenta del fenómeno de interés, u otro que logra gran precisión al coste de una complejidad mucho mayor?

¿Qué precio debemos pagar por la simplicidad en términos de adecuación de los resultados proporcionados por nuestro modelo a los datos reales? O, alternativamente, ¿qué complejidad adicional está justificada por un mejor ajuste a la evidencia?

Ejemplo 9.1 Consideremos el caso en que tratamos de establecer un modelo de regresión relacionando la talla y el peso de un colectivo de personas. Imaginemos N pares de valores $(\text{Talla}_i, \text{Peso}_i)$. Cabría imaginar una relación lineal entre ambos, o una relación polinómica (que, a la luz de la naturaleza de los datos, presupondríamos fácilmente cúbica). Es decir, podemos pensar, entre otras, en las siguientes dos relaciones entre Talla y Peso:

$$\text{Peso}_i = \beta_0 + \beta_1 \text{Talla}_i + \epsilon \quad (9.1)$$

$$\text{Peso}_i = \beta_0 + \beta_1 \text{Talla}_i + \beta_2 (\text{Talla}_i)^2 + \beta_3 (\text{Talla}_i)^3 + \epsilon. \quad (9.2)$$

Los β_i son parámetros y ϵ es una perturbación aleatoria inobservable que diluye la relación entre las dos magnitudes objeto de estudio: dos personas de la misma talla no necesariamente tienen el mismo peso. Es claro que (9.2) es un modelo más complejo que (9.1), que puede verse como un caso particular de aquél.

No sólo podríamos pensar en dos relaciones como las citadas (la segunda de las cuales siempre proporcionará un mejor ajuste que la primera, si nos dejan escoger los parámetros). Podríamos pensar en una relación funcional ajustando *perfectamente* los datos. Por ejemplo, un polinomio de grado $N-1$ (suponemos que no hay abscisas Talla_i repetidas). Intuitivamente, parece que tal relación funcional es mucho más compleja, y aunque el ajuste *a los* N *puntos muestrales* fuera perfecto, seríamos bastante reticentes a aceptar un polinomio de grado muy elevado como modelo adecuado de una relación subyacente entre talla y peso.

El ejemplo anterior sugiere que el número de parámetros de un modelo es un candidato a medir su complejidad. También que, a mayor numero de parámetros —si trabajamos con modelos anidados—, mejor ajuste del modelo a los datos muestrales. Sin embargo, en una situación como la anterior podríamos acaso preferir una

relación cúbica a una lineal —la mejora de ajuste quizá «vale» los dos parámetros adicionales de «complejidad»—, pero seríamos reticentes a admitir como modelo un polinomio de grado $N-1$.

Este tipo de planteamiento se ha hecho desde largo tiempo, y hay un sin número de criterios de bondad de ajuste que dan orientaciones para dirimir el conflicto ajuste–simplicidad. Volveremos sobre ellos más tarde tras considerar brevemente las ideas de Kolmogorov, Chaitin y Solomonoff. A la luz de su contribución —y a la de la precedente y fundamental de Shannon— se puede ver el trabajo estadístico desde una nueva óptica, que ha encontrado un enérgico y brillante valedor en Rissanen (véase Rissanen (1989)).

9.2. La lógica máximo-verosímil y la elección de modelos

9.2.1. Criterio máximo verosímil y modelos con diferente número de parámetros

Es interesante ver el parentesco del principio de máxima verosimilitud con la «navaja de Ockham.» No es la misma cosa, pero sí muestra cierta similitud: evitar el pensar en sucesos infrecuentes cuando hay alternativas más plausibles que dan cuenta de lo que observamos es un modo de buscar simplicidad.

Es preciso enfatizar que mientras el método máximo-verosímil no ofrece problemas en la estimación de los parámetros de un modelo, no es utilizable tal cual para *escoger entre modelos con diferente número de parámetros*: los modelos más parametrizados tenderán a dar valores mayores de la función de verosimilitud, sin que ello suponga que sean mejores. El siguiente ejemplo es ilustrativo.

Ejemplo 9.2 Supongamos cien monedas, aparentemente idénticas, cada una de ellas con dos caras que denotamos por «cara» (C) y «cruz» (+). Imaginemos que cada una de ellas tiene probabilidad θ de proporcionar C en un lanzamiento[2] y correlativa probabilidad $1-\theta$ de proporcionar '+'.

Lanzamos las cien monedas y obtenemos el resultado $\boldsymbol{x} = (x_1, \ldots, x_{100})$ con sesenta 'C' y cuarenta '+'. La Teoría de la Probabilidad indica que si la probabilidad de 'C' es θ, la probabilidad del suceso considerado[3] viene dada por,

$$P(\boldsymbol{x}|\theta) \quad = \quad \theta^{60}(1-\theta)^{40}; \tag{9.3}$$

un sencillo cálculo muestra que el estimador máximo verosímil de θ (que hace máxima (9.3)) es $\hat{\theta} = \frac{6}{10}$. El correspondiente valor de $P(\boldsymbol{x}|\theta)$ es $\approx 5{,}9085 \times 10^{-30}$. Llamamos verosimilitud de la muestra $\boldsymbol{x} = (x_1, \ldots, x_{100})$ a la expresión (9.3) vista como función de θ. El maximizar dicha expresión respecto de θ supone entonces escoger el valor del parámetro (estado de la Naturaleza) que hace más probable un suceso como el observado.

[2] Con lo cual, para simplificar, queremos decir que imaginamos que en una sucesión muy larga de lanzamientos tenderíamos a observar un 100θde 'C' y el resto de '+'.

[3] Es decir, sesenta «caras» y cuarenta «cruces» *precisamente* en el orden en que han aparecido; si prescindiéramos de considerar el orden, la cifra dada habría de multiplicarse por $\binom{100}{60}$.

Una alternativa sería imaginar que cada moneda, pese a ser aparentemente idéntica a las restantes, tiene su propia probabilidad de proporcionar 'C' ó '+'. La expresión (9.3) se transformaría entonces en

$$P(\boldsymbol{x}|\theta) \quad = \quad \prod_i \theta_i \prod_j (1-\theta_j), \tag{9.4}$$

en que el primer producto consta de sesenta términos y el segundo de cuarenta. Siendo $0 \leq \theta \leq 1$, (9.4) se maximiza dando a θ_k, $k = 1, \ldots, 100$, valor 1 ó 0, según la moneda correspondiente haya proporcionado cara o cruz. El valor máximo de (9.4) es así 1.

Es poco natural atribuir a cada moneda una probabilidad θ_i de «cara» diferente, habida cuenta de que parecen iguales. Obviamente, al hacerlo maximizamos la probabilidad de observar algo como lo acontecido: ¡con la elección referida de los cien parámetros $\theta_1, \ldots, \theta_{100}$ el suceso observado pasaría a tener probabilidad 1, lo que hace el suceso casi seguro! Sin embargo, aparte de poco atractivo intuitivamente, el modelo es claramente más complejo que el que usa sólo un parámetro, y difícilmente sería adoptado por nadie. Y ello a pesar de que tendría óptima capacidad generadora de un resultado como el observado.

Observación 9.1 Un fenómeno similar al que el ejemplo anterior muestra en un caso un tanto artificial y extremo se presenta cuando tratamos de seleccionar un modelo de regresión lineal. En presencia de normalidad en las perturbaciones, es fácil ver que el valor de la verosimilitud decrece monótonamente al crecer la suma de cuadrados de los residuos (SSE). Seleccionar el modelo dando lugar al máximo valor de la verosimilitud, sería equivalente a tomar aquél con mínima suma de cuadrados. Esto a su vez implica favorecer los modelos excesivamente parametrizados, porque la inclusión de un nuevo regresor siempre hace disminuir (o por lo menos no aumentar) SSE.

Como conclusión provisional de lo anterior, el criterio máximo verosímil es intuitivamente atrayente, aparte de tener propiedades muy deseables en grandes muestras (véase por ejemplo, Lehmann (1983); Cox y Hinkley (1974)); pero no puede tomarse en consideración para comparar modelos cuya complejidad —en un sentido aún por determinar, pero que parece tener mucho que ver con el número de parámetros— es muy disimilar.

9.2.2. El criterio AIC

Akaike propuso (ver Akaike (1972), Akaike (1974) reimpreso en Akaike (1991)) un criterio de selección de modelos que toma en cuenta el número de parámetros ajustados en cada uno: busca con ello corregir la tendencia del criterio máximo verosímil a favorecer los modelos más parametrizados. El criterio AIC enlaza con trabajo anterior del mismo autor (ver Akaike (1969), Akaike (1970)) y fue la primera de una larga serie de propuestas similares. Examinaremos en lo que sigue su fundamento siguiendo los trabajos Akaike (1991) y de Leeuw (2000).

Consideramos el caso en que con una muestra de tamaño N hemos de seleccionar uno entre m modelos. Cada uno de ellos se caracteriza por pertenecer su vector de parámetros $\boldsymbol{\theta}$ a un diferente espacio paramétrico, Θ_k. Se verifica

$$\ldots \Theta_k \subset \Theta_{k+1} \subset \ldots \Theta_m; \tag{9.5}$$

denotamos $\boldsymbol{\theta_k} \in \Theta_k$ al vector de parámetros correspondiente al modelo k-ésimo, y $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}$ a su estimador máximo verosímil.

Ejemplo 9.3 Consideremos modelos autorregresivos de órdenes crecientes,

$$X_t = \theta_1 X_{t-1} + \ldots + \theta_k X_{t-k} + \epsilon; \tag{9.6}$$

tenemos que $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ y los vectores de parámetros de los diferentes modelos toman valores en espacios anidados.

Para contrastar la hipótesis $H_0 : \boldsymbol{\theta} \in \Theta_k$ frente a $H_a : \boldsymbol{\theta} \in \Theta_\ell$, $\ell > k$, podemos recurrir al estadístico razón generalizada de verosimilitudes (Sección 8.5, pág. 109). En efecto, bajo H_0 tenemos que

$$-2 \log_e \left(\frac{\text{máx}_{\boldsymbol{\theta} \in \Theta_k} f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta})}{\text{máx}_{\boldsymbol{\theta} \in \Theta_\ell} f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta})} \right) \sim \chi^2_{\ell - k} \tag{9.7}$$

y rechazaremos H_0 si el estadístico en el lado izquierdo excede el valor crítico $\chi^2_{\ell-k;\alpha}$. No habría ningún problema si dejáramos ℓ fijo. El problema se presenta cuando al crecer el tamaño muestral N, crecen también k y ℓ. En tal caso, $\text{máx}_{\boldsymbol{\theta} \in \Theta_\ell} f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta})$ puede llegar a ser una estimación completamente distorsionada —optimista— debido al gran número de parámetros ajustados. El criterio AIC da una respuesta a este problema. Consideremos la expresión:

$$E_{\boldsymbol{Y}} \left[\int f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta_0}) \log_e \left(\frac{f_{\boldsymbol{X}}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}(\boldsymbol{Y}))}{f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta_0})} \right) d\boldsymbol{x} \right]. \tag{9.8}$$

Observemos que, para un cierto $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{Y})$, la expresión en el corchete es (con signo opuesto) la distancia de Kullback-Leibler entre las densidades $f_{\boldsymbol{X}}(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$ y $f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta_0})$. Maximizar dicho corchete equivaldría a maximizar

$$\int f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta_0}) \log_e f_{\boldsymbol{X}}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) d\boldsymbol{x} \approx \frac{1}{n} \sum_{i=1}^{n} \log_e f_X(z_i, \hat{\boldsymbol{\theta}}), \tag{9.9}$$

lo que muestra que $\hat{\boldsymbol{\theta}}$ debería ser aproximadamente el estimador máximo verosímil. Limitarse a maximizar el corchete estaría sujeto a los problemas derivados de tomar como modelo el que maximiza la verosimilitud (Ejemplo 9.2 y Observación 9.1 más arriba). Pero la propuesta de Akaike es diferente: propone maximizar *toda* la expresión (9.8).

Para convencernos de la razonabilidad de (9.8) como expresión a maximizar podemos reescribirla en términos de Teoría de la Decisión. Tenemos que

$$L(\boldsymbol{\theta_0}, \hat{\boldsymbol{\theta}}) = -\int f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta_0}) \log_e \left(\frac{f_{\boldsymbol{X}}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}(\boldsymbol{Y}))}{f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta_0})} \right) d\boldsymbol{x} \tag{9.10}$$

es una medida razonable de la pérdida derivada de seleccionar el modelo correspondiente a $\hat{\boldsymbol{\theta}}$ en lugar del "correcto", correspondiente a $\boldsymbol{\theta_0}$. El riesgo

$$r_{\boldsymbol{\theta_0}}(\hat{\boldsymbol{\theta}}) = E_{\boldsymbol{Y}} \left[L(\boldsymbol{\theta_0}, \hat{\boldsymbol{\theta}}(\boldsymbol{Y})) \right] \tag{9.11}$$

coincide entonces (salvo en el signo) con la expresión propuesta por Akaike, de manera que maximizar (9.8) es equivalente a minimizar el riesgo (9.11).

La expresión (9.11) depende de $\boldsymbol{\theta_0}$, y no es por ello directamente minimizable. Pero

$$\begin{aligned} 2r_{\boldsymbol{\theta_0}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) &= E_{\boldsymbol{Y},\boldsymbol{X}} \left[-2 \log_e \left(\frac{f_{\boldsymbol{X}}(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}(\boldsymbol{Y}))}{f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta_0})} \right) \right] && (9.12) \\ &\approx -\frac{2}{n} \sum_{i=1}^{n} \log_e \left(\frac{f_X(x_i; \hat{\boldsymbol{\theta}}_{\boldsymbol{k}})}{f_X(x_i, \boldsymbol{\theta_0})} \right) && (9.13) \\ &\stackrel{\text{def}}{=} D_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}, \boldsymbol{\theta_0}). && (9.14) \end{aligned}$$

Dado que $D_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}, \boldsymbol{\theta_0})$ no es evaluable (depende de $\boldsymbol{\theta_0}$), podemos tratar de estimar $2r_{\boldsymbol{\theta_0}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})$ por $D_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\ell}})$; si la parametrización "correcta" $\boldsymbol{\theta_0}$ se encuentra entre las consideradas, entonces, al ajustar el modelo más parametrizado $\hat{\boldsymbol{\theta}}_{\boldsymbol{\ell}} \xrightarrow{p} \hat{\boldsymbol{\theta}}_{\mathbf{0}}$ y podríamos esperar que $D_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\ell}}) \xrightarrow{p} D_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}, \boldsymbol{\theta_0})$. Este no tiene por qué ser el caso si $\ell \to \infty$ cuando $n \to \infty$: en tal caso, $D_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\ell}})$ será una estimación optimista de $D_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}, \boldsymbol{\theta_0})$, debido al gran número de parámetros empleado en su denominador. El criterio AIC busca corregir este sesgo optimista obteniendo una estimación aproximadamente insesgada de $D_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}, \boldsymbol{\theta_0})$.

En lugar de utilizar la función de pérdida directamente nos serviremos de aproximaciones de segundo orden como

$$L(\boldsymbol{\theta_0}, \boldsymbol{\theta}) \approx L(\boldsymbol{\theta_0}, \boldsymbol{\theta_0}) + \left[L'(\boldsymbol{\theta_0}, \boldsymbol{\theta}) \right]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}} (\boldsymbol{\theta} - \boldsymbol{\theta_0}) + (\boldsymbol{\theta} - \boldsymbol{\theta_0})' \left[L''(\boldsymbol{\theta_0}, \boldsymbol{\theta}) \right]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}} (\boldsymbol{\theta} - \boldsymbol{\theta_0});$$

bajo suficientes condiciones de regularidad,

$$\begin{aligned}
\left[L'(\boldsymbol{\theta_0},\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}} &= \left[\frac{\partial}{\partial\boldsymbol{\theta}}\int -f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta_0})\log_e\left(\frac{f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})}{f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta_0})}\right)d\boldsymbol{x}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}} \\
&= \int -f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta_0})\left[\frac{\partial\log_e f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}} d\boldsymbol{x} \\
&= \int -f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta_0})\frac{1}{f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta_0})}\left[\frac{\partial}{\partial\boldsymbol{\theta}}f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}} d\boldsymbol{x} \\
&= \int \left[-\frac{\partial}{\partial\boldsymbol{\theta}}f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}} d\boldsymbol{x} \\
&= \left[-\frac{\partial}{\partial\boldsymbol{\theta}}\int f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})d\boldsymbol{x}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta_0}} \\
&= 0.
\end{aligned}$$

En consecuencia,

$$L(\boldsymbol{\theta_0},\boldsymbol{\theta}) \approx (\boldsymbol{\theta}-\boldsymbol{\theta_0})'L''(\boldsymbol{\theta_0},\boldsymbol{\theta})(\boldsymbol{\theta}-\boldsymbol{\theta_0}). \tag{9.15}$$

Como (véase la Definición 5.1, pág. 62)

$$\left[L''(\boldsymbol{\theta_0},\boldsymbol{\theta})\right]_{(\boldsymbol{\theta}-\boldsymbol{\theta_0})} = I(\boldsymbol{\theta_0}), \tag{9.16}$$

en que $I(\boldsymbol{\theta_0})$ es la información de Fisher contenida en $\boldsymbol{X}$, tenemos que

$$L(\boldsymbol{\theta_0},\boldsymbol{\theta}) \approx (\boldsymbol{\theta}-\boldsymbol{\theta_0})'I(\boldsymbol{\theta_0})(\boldsymbol{\theta}-\boldsymbol{\theta_0}). \tag{9.17}$$

Definamos $\langle .,. \rangle_{I(\boldsymbol{\theta_0})}$ así:

$$\langle \boldsymbol{a},\boldsymbol{b}\rangle_{I(\boldsymbol{\theta_0})} = \boldsymbol{a}'I(\boldsymbol{\theta_0})\boldsymbol{b}, \tag{9.18}$$

y consiguientemente $\|\boldsymbol{a}\|^2_{I(\boldsymbol{\theta_0})} = \boldsymbol{a}'I(\boldsymbol{\theta_0})\boldsymbol{a}$. Sea

$$\boldsymbol{\theta_{0|k}} \stackrel{\text{def}}{=} \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta_0}-\boldsymbol{\theta}\|^2_{I(\boldsymbol{\theta_0})}, \tag{9.19}$$

es decir, la proyección de $\boldsymbol{\theta_0}$ sobre Θ_k en la métrica inducida por $\langle .,. \rangle_{I(\boldsymbol{\theta_0})}$. Tenemos entonces que:

$$\begin{aligned}
L(\boldsymbol{\theta_0},\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) &\approx (\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}-\boldsymbol{\theta_0})'I(\boldsymbol{\theta_0})(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}-\boldsymbol{\theta_0}) \\
&= \|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}-\boldsymbol{\theta_0}\|^2_{I(\boldsymbol{\theta_0})} \\
&= \|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}-\boldsymbol{\theta_{0|k}}\|^2_{I(\boldsymbol{\theta_0})} + \|\boldsymbol{\theta_{0|k}}-\boldsymbol{\theta_0}\|^2_{I(\boldsymbol{\theta_0})} \\
&\quad +\langle \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}-\boldsymbol{\theta_{0|k}},\boldsymbol{\theta_{0|k}}-\boldsymbol{\theta_0}\rangle_{I(\boldsymbol{\theta_0})}.
\end{aligned} \tag{9.20}$$

Consideremos ahora

$$\begin{aligned}
n\hat{D}_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{0}},\boldsymbol{\theta_{0|k}}) &\approx n(\hat{\boldsymbol{\theta}}_{\boldsymbol{0}}-\boldsymbol{\theta_{0|k}})'I(\boldsymbol{\theta_0})(\hat{\boldsymbol{\theta}}_{\boldsymbol{0}}-\boldsymbol{\theta_{0|k}}) \\
n\hat{D}_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}},\boldsymbol{\theta_{0|k}}) &\approx n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}-\boldsymbol{\theta_{0|k}})'I(\boldsymbol{\theta_0})(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}-\boldsymbol{\theta_{0|k}}).
\end{aligned}$$

Cuando $n \to \infty$, $\hat{\boldsymbol{\theta}}_{\mathbf{0}} \to \boldsymbol{\theta}_{\mathbf{0}}$ y $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} \to \boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}}$. Supongamos que $k \to \infty$ de modo que $\boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}} \to \boldsymbol{\theta}$ a la velocidad suficiente (basta que $n^{\frac{1}{2}}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}}) \not\to \infty$.) Entonces,

$$\begin{aligned} n\hat{D}_n(\hat{\boldsymbol{\theta}}_{\mathbf{0}}, \boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}}) &\approx n\|(\hat{\boldsymbol{\theta}}_{\mathbf{0}} - \boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}})\|^2_{I(\boldsymbol{\theta}_{\mathbf{0}})} \\ n\hat{D}_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}, \boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}}) &\approx n\|(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}})\|^2_{I(\boldsymbol{\theta}_{\mathbf{0}})} \end{aligned}$$

y tomando la diferencia de ambas expresiones,

$$\begin{aligned} n\hat{D}_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}, \hat{\boldsymbol{\theta}}_{\mathbf{0}}) &\approx n\|(\hat{\boldsymbol{\theta}}_{\mathbf{0}} - \boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}})\|^2_{I(\boldsymbol{\theta}_{\mathbf{0}})} - n\|(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}})\|^2_{I(\boldsymbol{\theta}_{\mathbf{0}})} \\ &= n\|\boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}} - \boldsymbol{\theta}\|^2_{I(\boldsymbol{\theta}_{\mathbf{0}})} + n\|\hat{\boldsymbol{\theta}}_{\mathbf{0}} - \boldsymbol{\theta}\|^2_{I(\boldsymbol{\theta}_{\mathbf{0}})} \\ &\quad - 2n\langle \hat{\boldsymbol{\theta}}_{\mathbf{0}} - \boldsymbol{\theta}, \boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}} - \boldsymbol{\theta}_{\mathbf{0}} \rangle_{I(\boldsymbol{\theta})} - n\|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}}\|^2_{I(\boldsymbol{\theta}_{\mathbf{0}})} \end{aligned} \tag{9.21}$$

Haciendo uso de (9.20) y (9.21) y tomando valor medio, los productos internos son aproximadamente cero en comparación con los otros términos y tenemos:

$$\begin{aligned} E\left[nL(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) - n\hat{D}_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}, \hat{\boldsymbol{\theta}}_{\mathbf{0}})\right] = E\Big[& n\|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}}\|^2_{I(\boldsymbol{\theta}_{\mathbf{0}})} + n\|\boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}} - \boldsymbol{\theta}\|^2_{I(\boldsymbol{\theta}_{\mathbf{0}})} \\ & - 2n\langle \hat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}}, \boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}} - \boldsymbol{\theta} \rangle_{I(\boldsymbol{\theta})} - n\|\boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}} - \boldsymbol{\theta}\|^2_{I(\boldsymbol{\theta}_{\mathbf{0}})} \\ & n\|\hat{\boldsymbol{\theta}}_{\mathbf{0}} - \boldsymbol{\theta}\|^2_{I(\boldsymbol{\theta}_{\mathbf{0}})} + 2n\langle \hat{\boldsymbol{\theta}}_{\mathbf{0}} - \boldsymbol{\theta}, \boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}} - \boldsymbol{\theta} \rangle_{I(\boldsymbol{\theta})} \\ & + n\|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}}\|^2_{I(\boldsymbol{\theta}_{\mathbf{0}})}\Big]. \end{aligned} \tag{9.22}$$

Cancelando términos de signo opuesto nos queda:

$$E\left[nL(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) - n\hat{D}_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}, \hat{\boldsymbol{\theta}}_{\mathbf{0}})\right] = 2n\|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \boldsymbol{\theta}_{\mathbf{0}|\boldsymbol{k}}\|^2_{I(\boldsymbol{\theta}_{\mathbf{0}})} - n\|\hat{\boldsymbol{\theta}}_{\mathbf{0}} - \boldsymbol{\theta}\|^2_{I(\boldsymbol{\theta}_{\mathbf{0}})}. \tag{9.23}$$

Por lo tanto, el sesgo en que incurrimos al aproximar $E[nL(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}})]$, que es lo que desearíamos utilizar, por $E[n\hat{D}_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}, \hat{\boldsymbol{\theta}}_{\mathbf{0}})]$, que es lo que podemos utilizar, es la suma de los dos términos en (9.23). El último de ellos es independiente de k, y podemos prescindir de él. El primero tiene valor medio $2k$. Por consiguiente, adoptaremos como modelo el que corresponda a $\boldsymbol{\theta}_{\boldsymbol{k}}$ minimizando

$$n\hat{D}_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}, \hat{\boldsymbol{\theta}}_{\mathbf{0}}) + 2k, \tag{9.24}$$

lo que a la vista de la definición de $\hat{D}_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}, \hat{\boldsymbol{\theta}}_{\mathbf{0}})$ en (9.12) equivale a minimizar

$$\text{AIC}(\boldsymbol{\theta}_{\boldsymbol{k}}) = -\frac{2}{n}\sum_{i=1}^{n} \log_e f_X(x_i; \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) + 2k, \tag{9.25}$$

expresión habitualmente utilizada como definición del criterio AIC.

9.3. Teoría de la información

Precisamos de un último ingrediente antes de introducir la noción de complejidad según Kolmogorov-Chaitin-Solomonoff, y su aplicación, entre otras, estadística. Es la Teoría de la Información, para la que Shannon (1948) (reimpreso en Shannon y Weaver (1949)) continúa siendo una referencia fundamental además de fácilmente accesible a no matemáticos. Otros textos introductorios son Abramson (1966) y Cullman et al. (1967).

Supongamos una fuente aleatoria de símbolos $a_1, \ldots, a_k$ que genera una sucesión de los mismos con probabilidades respectivas $p_1, \ldots, p_k$. Supongamos que símbolos sucesivos se generan de modo independiente[4]. Nos planteamos el problema de codificar (por ejemplo, binariamente) el flujo de símbolos, de tal modo que la transmisión de los mismos pueda hacerse con el mínimo número de dígitos binarios en promedio.

La solución es bastante obvia, y no se separa de la que Samuel Morse adoptó sobre base intuitiva al diseñar el código que lleva su nombre: reservaremos palabras de código (dígitos binarios, o combinaciones de ellos) «cortas» a los símbolos que se presenten con gran probabilidad, y asignaremos las de mayor longitud a los símbolos más improbables. De este modo, gran parte del tiempo estaremos transmitiendo palabras de código cortas[5].

Shannon dio base matemática a esta intuición, obteniendo algunos resultados de gran interés. En lo que sigue, sólo se proporcionan versiones simplificadas de algunos de ellos, que no obstante retienen bastante de su interés y evitan complicaciones formales. Pero bastantes enunciados podrían ser más generales[6].

Central a la Teoría de la Información es el concepto de *entropía*. Si tenemos una fuente aleatoria como la aludida al comienzo de la sección, generando k símbolos independientemente unos de otros con probabilidades respectivas $(p_1, \ldots, p_k)$, la entropía de la fuente (o de la distribución asociada a ella) viene dada por

$$H(p) \stackrel{\text{def}}{=} -\sum_{i=1}^{k} p_i \log_2 p_i,$$

con el convenio de que $p \log_2 p = 0$ si $p = 0$. La función $H(p)$ tiene bastantes propiedades interesantes. Una de ellas, inmediata, es que se anula cuando la distribución de símbolos se hace causal —es decir, cuando un símbolo se genera con probabilidad 1 y el resto con probabilidad cero—. Alcanza su máximo cuando la distribución es lo más difusa posible —en el caso de una distribución discreta que puede dar lugar a k símbolos, cuando cada uno de ellos tiene probabilidad $\frac{1}{k}$ de aparecer—.

[4] Es decir, que la fuente es de memoria nula. Se puede extender la teoría a fuentes markovianas en que este supuesto está ausente.

[5] Morse reservó el `.` para la letra `e`, muy frecuente en inglés, reservando para símbolos bastante más infrecuentes los códigos más largos (por ejemplo el cero, `0`, codificado mediante `-----`).

[6] En particular, las distribuciones utilizadas podrían ser continuas en vez de discretas, y los logaritmos en cualquier base, en lugar de binarios.

Cuadro 9.1: Ejemplo de construcción de código de Fano-Shannon.

Símbolo	p_i	$P_i = \sum_{j<i} p_j$	P_i	$L(i) = \lceil -\log_2 p_i \rceil$	Código
a_1	0,500	0	0.**0**00000...	1	0
a_2	0,250	0,500	0.**10**0000...	2	10
a_3	0,125	0,750	0.**110**000...	3	110
a_4	0,125	0,875	0.**111**000...	3	111

Un resultado muy fácil de demostrar[7] es el siguiente:

Teorema 9.1 *Para cualesquiera distribuciones discretas asignando respectivamente probabilidades* $(p_1, \ldots, p_k)$ *y* $(q_1, \ldots, q_k)$ *a* k *símbolos* $(a_1, \ldots, a_k)$*, se tiene:*

$$-\sum_{i=1}^{k} p_i \log_2 q_i \quad \geq \quad -\sum_{i=1}^{k} p_i \log_2 p_i. \tag{9.26}$$

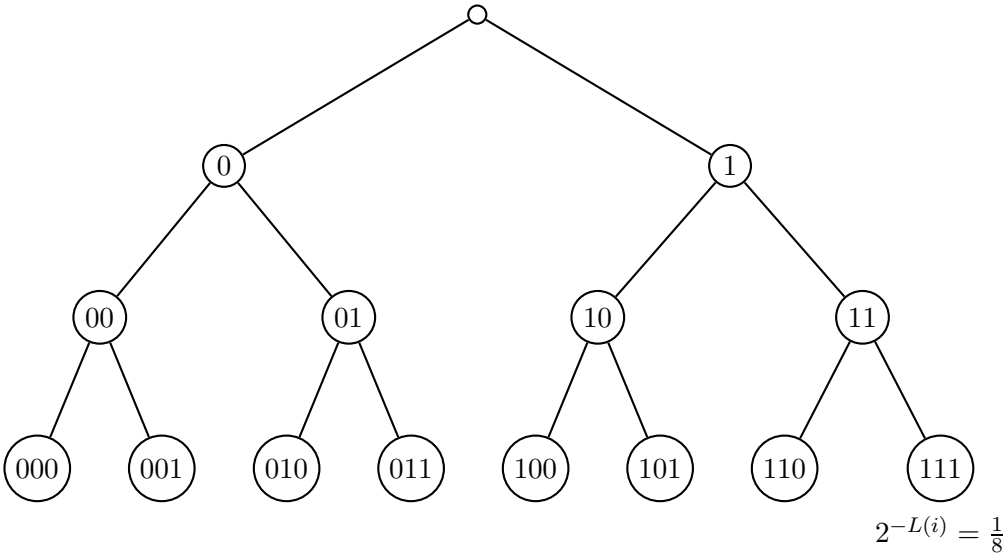
Hay otros interesantes hechos en los que la entropía juega un papel central. Por ejemplo, la mejor codificación que podemos hacer de los símbolos $(a_1, \ldots, a_k)$ requiere en promedio un número de dígitos binarios por símbolo acotado inferiormente por $H(p)$. Esto es intuitivamente coherente con la interpretación ya aludida de la entropía: $H(p)$ muy baja, significaría distribución de las probabilidades de los símbolos muy concentrada (dando gran probabilidad a uno o unos pocos símbolos, y poca al resto). Ello permitiría codificar los pocos símbolos muy probables con palabras de código muy cortas, y sólo raramente hacer uso de palabras más largas (para los símbolos más improbables).

Ejemplo 9.4 *(código de Fano-Shannon)* Veamos un modo de hacerlo. Supongamos una fuente generando cuatro símbolos a_1, a_2, a_3, a_4 ordenados de acuerdo a sus probabilidades respectivas p_1, p_2, p_3, p_4. Supongamos que éstas son las que se recogen en la segunda columna del Cuadro 9.1. Sea $P_i = \sum_{j<i} p_i$ como se indica en el Cuadro 9.1. Las palabras de código se asignan tomando una parte de la expresión binaria de P_i de longitud $L(i)$ igual a $-\log_2 p_i$ redondeado a la unidad superior. Intuitivamente, es fácil ver que el código anterior es razonable: asigna palabras cortas a los símbolos más probables —que ocupan las primeras posiciones en la tabla— y progresivamente más largas al resto.

El código de Fano-Shannon comparte con otros una propiedad que se deriva fácilmente del proceso constructivo que hemos seguido (véase por ejemplo Li y Vitányi (1993), p. 63) y que es aparente en la última columna del Cuadro 9.1: ninguna palabra de código es prefijo de otra de longitud mayor. Por ejemplo, a_2 se

[7] Véase por ejemplo Abramson (1966), p. 30.

Figura 9.1: Arbol binario completo de profundidad tres



codifica por `10` que no es comienzo de ninguna de las dos palabras de código de longitud tres (`110` y `111`). Esta propiedad —la de ser un código *libre de prefijos* o *instantáneo* permite decodificar «al vuelo». Cuando observamos `10`, sabemos que hemos llegado al final de una palabra, que podemos decodificar como a_2; esto no ocurriría si nuestro código incluyera palabras como `101`.

Los códigos libres de prefijos tienen longitudes de palabra $L(i)$ verificando la llamada *desigualdad de Kraft*, recogida en el siguiente

Teorema 9.2 *La condición necesaria y suficiente para que exista un código libre de prefijos con longitudes de palabra $L(1), \ldots, L(k)$ es que*

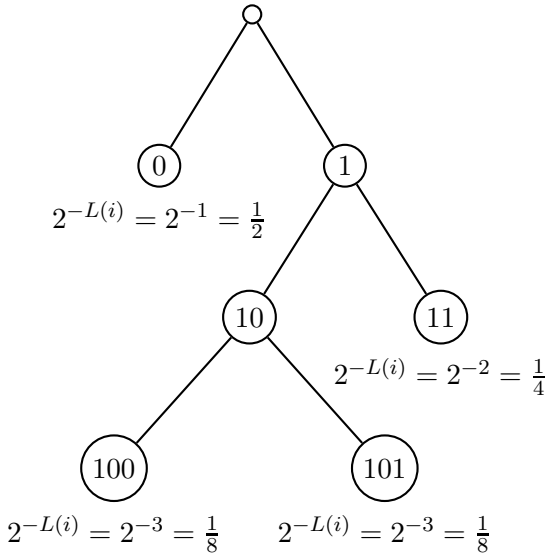
$$\sum_i 2^{-L(i)} \leq 1 \tag{9.27}$$

DEMOSTRACIÓN*:

La demostración es muy simple. Pensemos en todas las posibles palabras de todas las longitudes dispuestas en un árbol binario como el recogido en el Gráfico 9.1 (truncado a la profundidad 3). Si utilizáramos como palabras de código todas las de longitud 3, tendríamos $L(i) = 3$ y $2^{-l(i)} = \frac{1}{8}$ para $i = 1, \ldots, 8$ y la inecuación (9.27) se verificaría con igualdad.

Si escogemos una de las palabras de longitud inferior (uno de los nodos que no son «hojas» en el Gráfico 9.1), el requerimiento de ausencia de prefijos nos obliga a prescindir de todas las palabras correspondientes a nodos «hijos». El Gráfico 9.2

Figura 9.2: Arbol binario truncado



representa un árbol truncado con cuatro nodos terminales u «hojas», junto a las que se ha escrito $2^{-L(i)}$. Vemos que el tomar en `0` obliga a prescindir de `01`, `00`, y todos sus descendientes; pero 2^{-1} —contribución de `0` al lado izquierdo de (9.27)— es igual a la suma de las contribuciones a dicha expresión de todos los descendientes de los que hemos de prescindir.

Por tanto, trunquemos como trunquemos el árbol binario, la suma de $2^{-L(i)}$ extendida a sus «hojas» o nodos terminales será siempre 1. La desigualdad (9.27) sólo es estricta cuando despreciamos algún nodo terminal al construir nuestro código.

■

Podemos ya bosquejar la demostración del siguiente resultado:

Teorema 9.3 *Dada una fuente aleatoria con entropía* $H(p) = -\sum_i p_i \log_2 p_i$ *cualquier código instantáneo precisa un promedio de al menos* $H(p)$ *dígitos binarios de código por símbolo. Es decir, si la palabra codificando* a_i *tiene longitud* $L(i)$*, se verifica:*

$$\sum_i p_i L(i) \geq -\sum_i p_i \log_2 p_i \tag{9.28}$$

DEMOSTRACIÓN:
Definamos

$$q_i = \frac{2^{-L(i)}}{\sum_i 2^{-L(i)}}, \tag{9.29}$$

con lo que

$$\log_2 q_i \quad = \quad -L(i) - \log_2 \left(\sum_i 2^{-L(i)} \right) \geq -L(i). \tag{9.30}$$

La desigualdad anterior junto con el Teorema 9.1 proporcionan entonces de inmediato (9.28).

■

Obsérvese que el código de Fano-Shannon hacía $L(i) \approx -\log_2 p_i$ (redondeaba a la unidad superior): aproximadamente lo correcto. Verificaría (9.28) con igualdad si $-\log_2 p_i$ ($i = 1, \ldots, k$) resultaran ser siempre números enteros. En cualquier caso, el resultado que nos interesa es que para codificar un evento de probabilidad p_i, el código libre de prefijos óptimo requiere del orden de $-\log_2 p_i$ dígitos binarios.

9.4. Complejidad en el sentido de Kolmogorov

9.4.1. Información y complejidad

Estamos ya en condiciones de abordar la noción de complejidad según Kolmogorov-Chaitin-Solomonoff.

De cuanto se ha visto en la Sección 9.3 se deduce que $\log_2 p_i$ mide aproximadamente la información contenida en a_i. Se da sin embargo una paradoja, ya puesta de manifiesto por Laplace (véase por ejemplo Cover et al. (1989)), que sugiere emplear como medida de la complejidad de a_i algo diferente (aunque íntimamente relacionado con lo anterior).

Imaginemos las dos siguientes cadenas de dígitos binarios:

0000000000000000000000000000000

0011010001011101010001010111011

Ambas tienen el mismo número de dígitos binarios, 31. Si imaginamos el conjunto de todas las cadenas de 31 dígitos binarios —hay 2^{31} diferentes— y tomamos de ellas una al azar, cualquiera de las dos exhibidas tiene la misma probabilidad de aparecer: 2^{-31}. Sin embargo, desearíamos asignar a la primera una complejidad menor que a la segunda. Un modo de racionalizar esto es que podemos transmitir la primera a un tercero mediante una descripción muy parca: «treinta y un ceros.» La segunda requiere una descripción más verbosa, que a duras penas podría ser más escueta que la cadena misma[8].

[8] Esto es lo que caracteriza a las cadenas binarias «típicas»; véase por ejemplo Li y Vitányi (1993).

9.4.2. Complejidad de Kolmogorov*

Una idea prometedora en línea con la discusión anterior fue propuesta en los años sesenta por Solomonoff, Kolmogorov y Chaitin, de manera independiente unos de otros y con ligeras variantes[9]. La *complejidad de Kolmogorov* de una cadena binaria x es la longitud del mínimo programa p capaz de computarla. Formalmente,

$$C_f(x) \quad = \quad \text{mín}\left\{l(p) : f(p) = x\right\}. \tag{9.31}$$

Por razones técnicas, f en (9.31) debe ser una función recursiva —el tipo de función que puede computar una máquina de Turing—. Naturalmente, el «programa» p que, al ser ejecutado por el «computador» f, produce la cadena x, depende de f. Sea cual fuere x, podríamos imaginar un «computador» especializado que tan pronto se pone en marcha imprime x —es decir, que requiere un programa de longitud $l(p) = 0$ para computar x. ¿Implicaría esto que la complejidad de x es cero?

No. La complejidad de x relativa a la máquina de Turing que computa f vendría dada por (9.31). Relativa a otra máquina de Turing computando la función g sería $C_g(x)$, definida análogamente a $C_f(x)$. Definiremos la complejidad de Kolmogorov en relación a una máquina de Turing universal —una máquina que con el programa adecuado puede emular cualquier otra—. No hay una única máquina universal, pero para dos máquinas universales de Turing computando las funciones u y v y para cualquier cadena x se verifica

$$|C_u(x) - C_v(x)| \leq c_{u,v}, \tag{9.32}$$

en que $c_{u,v}$ es una constante que depende de u y de v, pero *no* de x.

> **Ejemplo 9.5** En Li y Vitányi (1993) se propone una ilustración de lo anterior que ayuda a la intuición a ver el sentido de (9.32). Hay lenguajes de alto nivel especializados en cálculo numérico y en cálculo simbólico: `FORTRAN` y `LISP` serían dos buenos ejemplos. Cierto tipo de problemas pueden programarse muy fácilmente en `FORTRAN` y son considerablemente más farragosos en `LISP`; en otros ocurre lo contrario. Pero podríamos imaginar programar en `FORTRAN` un intérprete de `LISP` (requiriendo un programa de c_1 bits de longitud) y en `LISP` uno de `FORTRAN` (requiriendo a su vez una longitud de c_2 bits). Entonces, la diferencia de longitudes de programa para resolver un mismo problema en `FORTRAN` o `LISP` nunca excedería de $c_{F,L} = \text{máx}\, c_1, c_2$; $C_{F,L}$ sería el máximo «precio» a pagar para implementar el lenguage más favorable al problema a mano en el otro lenguaje. Este precio es independiente del programa que se desea ejecutar: una vez programado en `FORTRAN` un interprete de `LISP` podemos emplear éste para ejecutar programas en `LISP` de cualquier longitud.

[9] La precedencia en el tiempo parece corresponder a Solomonoff: como en tantas otras ocasiones, la escena estaba preparadas en los años cincuenta para que investigadores trabajando de modo independiente llegarán a resultados similares. Véase una historia somera en Li y Vitányi (1993), Sección 1.6.

Todas las máquinas de Turing universales (o, alternativamente, las funciones recursivas que computan) se agrupan en clases de equivalencia en que cada pareja de funciones verifica (9.32), para una constante que sólo depende de la pareja considerada. Se puede demostrar que existe una «clase mínima», en el sentido de que (9.32) no se verifica para ninguna constante $c_{u,v}$ si u pertenece a la clase mínima y v no. Entonces, $C_u(x)$ define (salvo una constante) la complejidad de una cadena binaria x.

9.4.3. $C_u(x)$ no es computable*

El desarrollo anterior es útil por su poder clarificador, pero no directamente aplicable para computar un número que sea complejidad de una cierta cadena binaria. No existe un algoritmo con garantía de término que, al ser ejecutado por una máquina de Turing y alimentado con una cadena binaria, proporcione su complejidad.

No este el lugar para una discusión detallada de la no computabilidad de la complejidad de Kolmogorov, pero si puede intentarse una percepción intuitiva del motivo[10].

Imaginemos una cadena binaria x de n bits. Su complejidad no puede exceder mucho de n bits, ya que x es una descripción de sí misma. El programa más corto generando x no puede ser más largo que «print x», o su equivalente en la máquina de Turing de referencia que estemos empleando. Supongamos que la longitud de dicho programa es $(n + c)$ bits.

Podríamos ingenuamente pensar en formar una tabla con las cadenas binarias de longitud menor o igual que $(n + c)$, y ejecutarlas sucesivamente como programas en nuestra máquina de Turing, anotando si el resultado es x o no. Cada vez que obtuviéramos x, anotaríamos la longitud de la cadena binaria que hubiera servido como programa. Al final, la menor de las longitudes así anotadas, sería la complejidad de x.

Pero nada garantiza que haya final, porque nada garantiza que la máquina de Turing que empleamos se detenga al ejecutar como programa una cualquiera de las cadenas que le pasamos; mucho menos que lo haga con todas. La no computabilidad de $C_u(x)$ deriva del *halting problem*, o imposibilidad de determinar anticipadamente si una máquina de Turing se detendrá o proseguirá indefinidamente ejecutando un programa determinado. Sobre la no computabilidad de $C_u(x)$, y su relación con el teorema de Gödel y la indecidibilidad de proposiciones puede verse Li y Vitányi (1993) y Chaitin (1987).

[10] Que sigue el razonamiento en el último capítulo de Ruelle (1991), una introducción muy legible y diáfana al tratar esta cuestión, aunque sólo lo haga tangencialmente al final.

9.5. De la complejidad de Kolmogorov a la Longitud de Descripción Mínima (MDL)

Si bien no podemos hacer uso directamente de la complejidad de Kolmogorov para escoger entre distintos modelos, las ideas expuestas son de forma limitada aplicables. Veremos el modo de hacerlo sobre un ejemplo que, aunque artificialmente simple, ilustra la aproximación propuesta por Rissanen (véase Rissanen (1989)),

> **Ejemplo 9.6** *(continuación del 9.2)* Regresemos al Ejemplo 9.2. Describir llanamente el resultado de un experimento como el allí realizado al lanzar cien monedas al aire requiere 100 bits, si aceptamos el convenio de utilizar el dígito binario 0 para codificar el resultado '+' y el 1 para codificar el resultado 'C'. Obsérvese que 100 bits es exactamente la cantidad de información necesaria para singularizar una cadena binaria de longitud 100 de entre las 2^{100} posibles cuando no hay nada que haga unas de ellas más plausibles que otras.
>
> ¿Lo podemos hacer mejor? Quizá sí. En lo que sigue veremos cómo..

En lo que sigue formalizaremos algo esta idea.

9.5.1. Modelos como generadores de códigos

Consideremos una fuente aleatoria que ha generado $\boldsymbol{x}$. Si tenemos un modelo probabilístico, en general dependiente de parámetros $\boldsymbol{\theta}$, que describe el modo en que se genera $\boldsymbol{x}$, podemos calcular $P(\boldsymbol{x}|\boldsymbol{\theta})$ para los distintos resultados experimentales. Resultados con $P(\boldsymbol{x}|\boldsymbol{\theta})$ «grande» corresponderán a resultados esperables, que desearíamos claramente codificar mediante palabras de código cortas. Lo contrario ocurre con aquéllos en que $P(\boldsymbol{x}|\boldsymbol{\theta})$ es pequeño.

Estamos pensando como si θ fuera fijo y conocido, pero no lo es: lo hemos de escoger (estimar). Si lo hacemos maximizando $P(\boldsymbol{x}|\boldsymbol{\theta})$ (aplicando por tanto el principio de máxima verosimilitud), estamos atribuyendo al resultado $\boldsymbol{x}$ observado la máxima probabilidad. Pero no debemos olvidar que, para que sea posible la decodificación, hemos de facilitar también el valor $\boldsymbol{\theta}$ codificado (y la forma de nuestro modelo). El uso de máxima verosimilitud minimiza $\lceil -\log_2 P(\boldsymbol{x}|\boldsymbol{\theta})\rceil$, pero hace caso omiso de la longitud de código necesaria para $\boldsymbol{\theta}$.

9.5.2. Descripción de longitud mínima (MDL)

El agregar a $\lceil -\log_2 P(\boldsymbol{x}|\theta)\rceil$ el número de bits necesario para codificar los parámetros da lugar a la versión más cruda del llamado criterio MDL o de «mínima longitud de descripción.»

A efectos de codificar los parámetros hemos de considerar dos cosas. En primer lugar, podemos tener información *a priori* sobre los mismos, de cualquier procedencia, traducible a una distribución a priori sobre los mismos con densidad $\pi(\boldsymbol{\theta})$.

En segundo lugar, típicamente $\boldsymbol{\theta}$ es un número real que requeriría infinitos bits fijar con exactitud. Por ello trabajaremos con una versión truncada de él.

Si para el parámetro θ deseamos utilizar q dígitos binarios, llamaremos precisión a $\delta = 2^{-q}$. Suponiendo una densidad *a priori* $\pi(\theta)$, tendríamos los posibles valores de θ clasificados en intervalos de probabilidad aproximada $\pi(\theta)\delta$, especificar uno de los cuales requiere aproximadamente $-\log_2 \pi(\theta)\delta$ bits. Si hay k parámetros, se tiene la generalización inmediata,

$$-\log_2 \pi(\boldsymbol{\theta}) \prod_{i=1}^{k} \delta_i. \tag{9.33}$$

El criterio MDL propone tomar el modelo que minimiza la longitud total de código, la necesaria para los datos $\boldsymbol{x}$ más la necesaria para los parámetros:

$$\begin{aligned} MDL &= -\log_2 P(\boldsymbol{x}|\boldsymbol{\theta}) + l(\boldsymbol{\theta}) && (9.34)\\ &= -\log_2 P(\boldsymbol{x}|\boldsymbol{\theta}) - \log_2 \pi(\boldsymbol{\theta}) - \sum_{i=1}^{k} \log_2 \delta_i. && (9.35) \end{aligned}$$

en que $l(\theta)$ es la longitud de código necesaria para transmitir el o los parámetros empleados. Un ejemplo, de nuevo artificialmente simple, ilustra esto.

Ejemplo 9.7 *(continuación del Ejemplo 9.2)* Imaginemos que decidimos truncar el valor de θ en el Ejemplo 9.2 a 8 bits —por tanto sólo consideramos valores con una resolución de $\delta = 2^{-8} \approx 0{,}003906$—. Llamemos Θ_* al conjunto de valores que puede adoptar el parámetro así truncado. Imaginemos también que tenemos una distribución a priori uniforme $\pi(\theta)$ sobre los valores de θ; como $0 \leq \theta \leq 1$, $\pi(\theta) = 1$.

El criterio MDL para el modelo considerado en el Ejemplo 9.2 tomaría el valor:

$$MDL = \min_{\theta \in \Theta_*} \left\{ -\log_2 \theta^{60}(1-\theta)^{40} - \log_2 \pi(\theta) - \log_2 \delta \right\} \tag{9.36}$$

Si suponemos δ constante, sólo nos hemos de preocupar de minimizar el primer término. De poder escoger θ libremente, tomaríamos $\theta = 0{,}60$. Como estamos truncando los valores, 0.60 no es alcanzable, pero sí lo son $(153 + \frac{1}{2})/256 = 0{,}599609$ y $(154 + \frac{1}{2})/256 = 0{,}603516$, puntos medios de intervalos de longitud $1/256$ en que se subdivide $[0, 1]$ cuando se emplea precisión $\delta = 2^{-8} = 1/256$. El primero de ellos proporciona el mínimo valor de $-\log_2 P(\boldsymbol{x}|\theta)$, que resulta ser $97{,}0951$. Requerimos un total de $97{,}0951 + 8 = 105{,}0951$ bits como longitud de descripción.

Una alternativa (tal y como se discutió a continuación del Ejemplo 9.2) sería considerar cien parámetros, uno para cada moneda. Ello haría «casi seguro» el suceso observado, y el primer sumando de (9.36) sería cero —especificados los parámetros, no haría falta ningún código para especificar el resultado—. Pero el tercer sumando sería, para la misma precisión, mucho mayor: ¡800 bits! Aunque el modelo binomial haciendo uso de cien parámetros hace casi seguro el resultado observado, es inferior al que sólo hace uso de sólo un parámetro, debido al coste de codificar noventa y nueve parámetros adicionales.

Cuadro 9.2: Longitud de descripción para diferentes valores de δ.

q	δ	$\hat{\theta}_{MV}$	$\hat{\theta}$	$\hat{\theta}^{90}(1-\hat{\theta})^{10}$	$-\log_2 \hat{\theta}^{90}(1-\hat{\theta})^{10}$	MDL
1	0.50000	0.90	0.75	$5{,}4314 \times 10^{-18}$	57.35	58.35
2	0.25000	0.90	0.875	$5{,}6211 \times 10^{-15}$	47.34	$49{,}34^*$
3	0.12500	0.90	0.9375	$2{,}7303 \times 10^{-15}$	48.38	51.38
4	0.06250	0.90	0.90625	$7{,}447911 \times 10^{-15}$	46.93	50.93

El ejemplo anterior suponía δ fijo a efectos puramente ilustrativos: pero en la práctica se minimiza MDL en (9.35) sobre θ y sobre δ. Es fácil ver que mientras disminuir la precisión (incrementar δ) disminuye el tercer sumando, hace en general crecer el primero (el «mejor» θ en Θ_* estará en general más lejos del óptimo θ cuanto más tosca sea la discretización de θ).

Un último ejemplo permitirá ver el efecto de optimizar la longitud de descripción sobre δ, precisión del parámetro.

Ejemplo 9.8 *(continuación de los Ejemplos 9.2, 9.6 y 9.7)* Consideremos la misma situación del Ejemplo 9.2, pero supongamos —para mostrar un caso en que se obtiene una reducción apreciable de la longitud de descripción— que se han obtenido noventa «caras» 'C' y diez '+'. Optimizaremos sobre $\delta = 2^{-q}$ dejando variar q sobre los enteros. El estimador máximo verosímil de θ es $\hat{\theta}_{MV} = 0{,}9$. El Cuadro 9.2 muestra el valor de θ entre los posibles que minimiza MDL para cada q. Con un asterisco se señala la descripción más escueta de los datos a que se llega. Obsérvese que cuando consideramos una precisión de $\delta = 2^{-q}$ estamos dividiendo $[0,1]$ en 2^q intervalos del la forma $[n2^{-q}, (n+1)2^{-q})$ ($n = 0, 2^q - 1$), cuyo punto medio es $n2^{-q} + 2^{-q-1}$; éstos son los valores que se recogen en la columna $\hat{\theta}$.

Obsérvese que aquí la longitud de descripción es acusadamente menor que los 100 bits que requeriría describir el resultado de nuestro experimento. Al ser uno de los resultados ('C') considerablemente más frecuente, podemos diseñar un código que tenga esto en consideración. No ocurría lo mismo en el Ejemplo 9.7, en que la ligera mayor probabilidad de 'C' dejaba poco margen a la optimización del código; como se vio, la ventaja obtenida no alcanzaba a «pagar» la especificación del parámetro necesario.

9.5.3. De la MDL a la complejidad estocástica*

La discusión en el apartado anterior no hace sino introducir algunas ideas esenciales; pero en modo alguno hace justicia a la potencia del método.

La mínima longitud de descripción (MDL), en cierto sentido, es *más* de lo que buscábamos. Deseábamos una codificación compacta de $\boldsymbol{x}$ y hemos acabado con una codificación de $\boldsymbol{x}$ y *adicionalmente* de $\boldsymbol{\theta}$. La complejidad estocástica se obtiene integrando $P(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ sobre los parámetros. En otras palabras, tenemos una distribución $P(\boldsymbol{x}|\boldsymbol{\theta})$ de los datos dados los parámetros y el modelo, y una

densidad *a priori* $\pi(\boldsymbol{\theta})$ sobre los parámetros. La complejidad estocástica de los datos $\boldsymbol{x}$ relativa al modelo considerado se define como

$$I(\vec{x}) \quad = \quad \int_{\Theta} P(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{9.37}$$

(véase Rissanen (1989) para más detalles). Además, en el caso de que no tengamos una distribución *a priori* sobre los parámetros, podemos emplear la distribución *a priori* universal. Supongamos que deseamos una codificación que asigne una palabra de código a todos los números naturales n, sobre los que hay definida una distribución $P(n)$. Bajo condiciones muy generales, existe una codificación asignando longitud de palabra $L^*(n)$ a n y que verifica

$$\lim_{N\to\infty} \frac{\sum_{n=0}^{N} P(n)L^*(n)}{\sum_{n=0}^{N} P(n)\log_2 n} \quad = \quad 1 \tag{9.38}$$

Merece la pena examinar la igualdad anterior: ¡hay una codificación que es asintóticamente óptima sobre los enteros y que es «todo terreno»! ¡Vale sea cual fuere la distribución definida sobre ellos, con tal de que sea monótona decreciente a partir de algún n dado! La función $L^*(n)$ viene dada aproximadamente por

$$L^*(n) \quad = \quad \log_2 c + \log_2 \log_2 n + \log_2 \log_2 \log_2 n + \ldots; \tag{9.39}$$

con $c = 2{,}865$, verifica la desigualdad de Kraft y a partir de ella puede obtenerse una distribución *a priori* universal: $P(n) = 2^{-L^*(i)}$. Esta es la que Rissanen propone utilizar en la definición de complejidad estocástica[11]. En el caso en que tenemos parámetros que no toman valores enteros, se puede también definir una distribución *a priori* universal del modo descrito en Rissanen (1983).

9.5.4. Ideas relacionadas y conexas

Aunque en el Ejemplo 9.8 se ha buscado la longitud de descripción minimizando explícitamente sobre la precisión (en el Cuadro 9.2), en la práctica no es preciso recorrer un camino similar con cada modelo que se prueba. Argumentos de tipo asintótico dan un resultado similar en forma mucho más simple. Habitualmente sólo se requiere computar una función que da aproximadamente la longitud de descripción, y que típicamente consta de una parte que disminuye al mejorar el ajuste a los datos (término de fidelidad o ajuste) y otra que crece con el número de parámetros (término de penalización de la complejidad del modelo). Por ejemplo, de modo bastante general (véase Rissanen (1989) para las condiciones necesarias) la mínima longitud de descripción de $\boldsymbol{x} = (x_1, \ldots, x_N)$ utilizando un modelo con p parámetros viene dada por:

$$\text{MDL}(p) \quad = \quad -\log\left(P(\boldsymbol{x}|\hat{\theta})\pi(\hat{\theta})\right) + \frac{p}{2}\log N + O(p). \tag{9.40}$$

[11] En el Ejemplo 9.7 hemos empleado una densidad $\pi(\theta)$ uniforme por simplicidad.

Puede verse un primer término que disminuye al mejorar el ajuste y un segundo término (la penalización) que crece con el número de parámetros p y está dominado por $\frac{p}{2}\log N$.

A la vista de una expresión como (9.40) es forzoso pensar en los muchos criterios que se han propuesto para evaluar la adecuación de un modelo, muchas veces sobre bases puramente heurísticas. En el caso de modelos de regresión lineal tenemos por ejemplo el estadístico conocido como C_p de Mallows,

$$C_p \quad = \quad \frac{\sum_{i=1}^{N} \hat{\epsilon}^2}{\hat{\sigma}^2} + 2p \tag{9.41}$$

en que $\hat{\epsilon}$ son los residuos de la regresión y σ^2 la varianza del término de error: véase Mallows (1973). El primer término de (9.41) disminuye al mejorar el ajuste o fidelidad del modelo a los datos; el segundo, crece con el número de parámetros. El criterio de información de Akaike introducido en la Sección 9.2.2 y definido por

$$AIC(p) \quad = \quad -2\log_e(\text{Prob}\left\{\boldsymbol{x}|\hat{\theta}_{\text{MV}}\right\}) + 2p, \tag{9.42}$$

también de la misma forma que (9.40), aunque penalizando asintóticamente menos la introducción de parámetros. Los ejemplos podrían multiplicarse; una recopilación reciente de trabajos incorporando ideas como las mencionadas a múltiples campos es Dowe et al. (1996).

La búsqueda de longitudes de descripción mínimas o mínimas complejidades no se separa pues, por lo menos asintóticamente, de algunos criterios que han sido utilizados con asiduidad. La novedad está más bien en la justificación de resultados antes obtenidos para problemas concretos y de forma bastante *ad-hoc* desde una perspectiva unificadora.

9.6. ¿Tiene sentido esto?

Se han esbozado ideas que basan la elección de modelos en un criterio de simplificación de la información. Apoyándose en el trabajo pionero que sobre la noción de complejidad y sobre Teoría de la Información se realizó en los años cincuenta y sesenta, estas ideas pueden verse como una navaja de Ockham sofisticada, de posible utilización en el trabajo estadístico. Importa ahora no obstante regresar al origen y preguntarse sobre el alcance, pertinencia y solidez de este modo de actuar.

¿Es la noción de complejidad de Kolmogorov —o versiones menos ambiciosas de la misma idea, como la de Rissanen— el anclaje al que deseamos asirnos para hacer inferencia? No parece evidente. Es un planteamiento no exento de belleza, y que, como se ha indicado, da en su aplicación práctica resultados satisfactorios.

¿Debemos entender por complejidad sólo esto, o algo más? ¿Es la longitud de descripción tal como la hemos presentado una buena medida de la complejidad de un modelo más los datos, haciendo abstracción —por ejemplo— del coste de

llegar a obtenerlo? Murray Gell-Mann (véase Gell-Mann (1994), p. 117) menciona, haciéndose eco de trabajo de Charles Bennet, que la complejidad tiene facetas como la *profundidad* y *cripticidad*. En relación a esta última, por ejemplo, una serie muy larga de números pseudo-aleatorios generados en un ordenador mediante el conocido método multiplicativo, puede tener una complejidad muy baja: se puede describir dando la semilla o valor inicial y los valores de tan sólo dos números. Sin embargo, adivinar cuáles son estos números es muy costoso. ¿Diríamos que esta serie es de baja complejidad?

Un modelo es un modo de especificar regularidades. Decimos que «explica» la realidad cuando lo que observamos se adecúa a las predicciones que obtendríamos con ayuda de dicho modelo. En el caso de un modelo estadístico, ni siquiera exigimos una concordancia perfecta entre predicciones y observaciones, porque la esencia de un modelo de tal naturaleza es no fijar unívocamente las relaciones entre observables.

Es precisamente la existencia de regularidad en la evidencia lo que permite su descripción escueta. Servirse de un criterio como el de mínima longitud de descripción es aceptar como buena la «explicación» que más regularidades encuentra en nuestros datos —o mejor las explota—. Tiene al menos la ventaja sobre la modelización usual de que explicita el coste a pagar por la complejidad añadida. Queda a medio camino entre la inferencia bayesiana y la convencional, y sortea algunos de los aspectos más criticables en esta última —la fijación arbitraria de niveles de significación, por ejemplo—.

Pero, en su raíz, el minimizar la complejidad es un criterio que prioriza la reducción de los datos observados. ¿Es esto sensato? ¿Válido como criterio de inferencia?

B. Russell (véase Russell (1912), p. 35) obliga a responder que no. Un pollo que observara al granjero llevarle grano todos los días —dice Russell—, podría llegar a la conclusión de que el granjero le ama y busca su bien. Tal «modelo» explicaría las repetidas visitas al corral del granjero y su solicitud con el animal. Pero esta «explicación», tan repetidamente apoyada por la evidencia durante la vida del pollo, se ve bruscamente sin valor el día que el granjero decide que el pollo está lo suficientemente gordo como para retorcerle el pescuezo.

Enfrentados al mundo, querríamos saber *porqué*, y ni tan solo sabemos si nuestra noción de causalidad tiene sentido; si cabe hablar de un porqué. Querríamos conocer el fin último, si lo hay, de las idas y venidas del granjero: conformarnos con la explicación menos compleja de su conducta nos coloca en situación no mejor que la del pollo.

Sin embargo, frecuentemente no podemos hacer más. Enfrentados a este hecho, nuestra pertinaz tentativa de entender encuentra en el criterio de minimizar la longitud de descripción un sucedáneo útil: la vieja navaja de Ockham con un nuevo filo. El éxito que alcancemos con su empleo no debiera hacernos olvidar lo endeble de nuestra posición. Quizá el mayor valor de las ideas expuestas más arriba no esté en las respuestas que proporcionan sino en las preguntas que suscitan.

Apéndice A

Convergencias estocásticas

A.1. Sucesiones de variables aleatorias

Podemos considerar una sucesión aleatoria como la generalización del concepto de variable aleatoria. Una v.a. real es una aplicación $X : \Omega \longrightarrow R$ (ó $X : \Omega \longrightarrow R^n$ si se trata de una v.a. multivariante)[1]. Una sucesión aleatoria real es una aplicación $X : \Omega \longrightarrow R^\infty$, que a cada $\omega \in \Omega$ hace corresponder una sucesión de números reales $\{X_n\}$. Es importante notar que, fijado ω, $\{X_n\}$ es una sucesión ordinaria de números reales; la aleatoriedad radica precisamente en la dependencia de ω.

Ejemplo A.1 Las sucesiones aleatorias aparecen de modo natural en multitud de contextos. Imaginemos el caso en que deseamos estimar la probabilidad de que una determinada moneda produzca "cara" al efectuar un lanzamiento. Podríamos, al menos conceptualmente, realizar infinidad de lanzamientos. Si el i-ésimo lanzamiento produce el resultado $X_i(\omega) = 1$ ("cara")

[1] Véase cualquier texto introductorio de Probabilidad y Estadística, por ejemplo Trocóniz (1987), Cap. 5, para una definición precisa. Se requiere que X sea una función medible de Borel, lo que daremos por supuesto. En lo que sigue obviamos también detalles técnicos de similar naturaleza.

ó $X_i(\omega) = 0$ ("cruz"), tendríamos la siguiente sucesión de estimadores:

$$\begin{aligned}
\overline{X}_1(\omega) &= X_1(\omega) \\
\overline{X}_2(\omega) &= \frac{X_1(\omega) + X_2(\omega)}{2} \\
\overline{X}_3(\omega) &= \frac{X_1(\omega) + X_2(\omega) + X_3(\omega)}{3} \\
&\vdots \\
\overline{X}_n(\omega) &= \frac{X_1(\omega) + X_2(\omega) + \ldots + X_n(\omega)}{n} \\
&\vdots
\end{aligned}$$

Podemos ver $\{\overline{X}_n(\omega)\}$ como una sucesión de variables aleatorias. Su estudio cuando $n \to \infty$ proporcionará información sobre el comportamiento esperable de nuestro estimador al dejar crecer sin límite el número de lanzamientos.

Nos interesarán dos cuestiones al estudiar una sucesión aleatoria:

- ¿Se "aproxima" a alguna *distribución* concreta la de $X_n(\omega)$ cuando $n \to \infty$?
- ¿Se "aproxima" $X_n(\omega)$ a alguna *variable aleatoria* cuando $n \to \infty$?

Para responder a ambas necesitamos nociones adecuadas de "aproximación".

A.2. Convergencia en ley

Definición A.1 *La sucesión de funciones de distribución $F_{X_n}(x)$ converge* en distribución *(o* en ley*) a la función de distribución $F_X(x)$ si $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ en todo punto de continuidad de ésta última.*

Por extensión, diremos que la sucesión de v.a. $\{X_n\}$ converge a X, y lo denotaremos así: $X_n \xrightarrow{\mathcal{L}} X$.

Observación A.1 Esta notación, sin embargo, no debe crear la falsa impresión de que X_n "se aproxima" a X (en el sentido de tomar valores muy próximos con elevada probabilidad). Nada más lejos de la verdad. Por ejemplo, podríamos tener una sucesión aleatoria $\{X_n\}$ todos cuyos términos fueran idénticos entre sí, e iguales a una v.a. X con distribución uniforme $U(0,1)$. Entonces, $X_n \xrightarrow{\mathcal{L}} Y = 1 - X$. La distribución de X (y por tanto de cualquier X_n) es igual que la de Y (si $X \sim U(0,1)$, entonces $Y = (1 - X)$ también se distribuye como $U(0,1)$). Sin embargo, *el valor* de X_n no hay razón para esperar que esté en las cercanías del de Y.

A.3. Convergencias en probabilidad, media cuadrática y casi segura

La intuición sugiere que en el Ejemplo A.1 $\overline{X}_n$ se aproxima a la probabilidad p de "cara". En Análisis Matemático, decimos que $a_n \longrightarrow a$ si, prefijado un número $\epsilon > 0$, es posible encontrar $N(\epsilon)$ tal que para $n > N(\epsilon)$ se verifica *necesariamente* que: $|a_n - a| < \epsilon$.

No podemos decir que $\overline{X}_n$ en el Ejemplo A.1 converja a p en este sentido: sea cual fuere n, podría ocurrir que todos los lanzamientos hubieran proporcionado "cara" (o todos "cruz"). No podemos asegurar, para ningún n, que $\overline{X}_n$ estará a distancia menor de p que un $\epsilon > 0$ prefijado.

Sin embargo, en el ejemplo citado, existe elevada probabilidad de que $\overline{X}_n \simeq p$. Ello sugiere el modo de formalizar la percepción intuitiva de que $\overline{X}_n$ "tenderá" a p diciendo que $\overline{X}_n$ converge *en probabilidad* a p. La definición precisa de convergencia en probabilidad es la siguiente:

Definición A.2 *La sucesión* $\{X_n\}$ *converge* en probabilidad *a la variable aleatoria* X *si* $\forall \epsilon > 0$ *y* $\forall \delta > 0$, $\exists N(\epsilon, \delta)$ *tal que* $n > N(\epsilon, \delta)$ *implica*

$$Prob\{\omega : |X_n(\omega) - X(\omega)| < \epsilon\} \geq 1 - \delta \tag{A.1}$$

o, equivalentemente, si para cualquier $\epsilon > 0$ *prefijado*

$$\lim_{n \to \infty} Prob\{\omega : |X_n(\omega) - X(\omega)| < \epsilon\} = 1. \tag{A.2}$$

Es decir, si podemos lograr que X_n esté en un entorno de X de radio $\epsilon > 0$ prefijado con probabilidad tan cercana a 1 como deseemos, tomando n lo suficientemente grande. Denotaremos la convergencia en probabilidad mediante $X_n \xrightarrow{p} X$ o plim$X_n = X$.

Es fácil ver que es equivalente escribir $X_n \xrightarrow{p} X$ ó $(X_n - X) \xrightarrow{p} 0$.

Ejemplo A.2 Definamos una sucesión de variables aleatorias así:

$$X_n = \begin{cases} a & \text{con probabilidad } 1 - \frac{1}{n} \\ bn & \text{con probabilidad } \frac{1}{n}. \end{cases}$$

Es inmediato comprobar que converge en probabilidad a a. Observemos, sin embargo, que $\lim E[X_n] = (a + b) \neq a$. Una variable puede converger en probabilidad a otra (en este caso, una variable degenerada o causal), que siempre toma el valor a y por tanto tiene valor medio a. Los momentos, sin embargo, no necesitan converger.

En ocasiones, X_n converge a X de un modo aún más estricto, *con probabilidad 1* ó *casi seguramente*.

Definición A.3 *La sucesión* $\{X_n\}$ *converge* casi seguramente *a la variable aleatoria* X *si:*

$$Prob\left\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\} = 1 \tag{A.3}$$

Fácilmente se comprueba que $X_n \xrightarrow{c.s.} X \Rightarrow X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{\mathcal{L}} X$. Es útil examinar ejemplos en que se presenta un tipo de convergencia y no otro, para adquirir intuición sobre su naturaleza y respectivas implicaciones; pueden verse, entre otros muchos, Billingsley (1986), Garín y Tusell (1991), Romano y Siegel (1986).

La comparación de las expresiones (A.1) y (A.3) muestra de inmediato que $X_n \xrightarrow{c.s.} X \Rightarrow X_n \xrightarrow{p} X$. La implicación recíproca, por el contrario, no se verifica, como el siguiente ejemplo pone de manifiesto.

Ejemplo A.3 Ejemplo ondas cuadradas.

Definición A.4 *Decimos que* $\{X_n\}$ *converge en* media r *a la variable aleatoria* X *si:*

$$\lim_{n\to\infty} E\left|X_n - X\right|^r = 0. \tag{A.4}$$

Con diferencia, el caso más común es el de $r = 2$*; cuando una sucesión verifica (A.4) con* $r = 2$ *se dice que converge* en media cuadrática *a* X.

Es fácil comprobar (véase Ejercicio A.2) que la convergencia en media cuadrática implica la convergencia en probabilidad. No hay, en cambio, relación entre la convergencia en media cuadrática y casi segura: ninguna implica la otra.

Teorema A.1 *Si* $X_n \xrightarrow{\mathcal{L}} X$ *y* A_n, B_n *son sucesiones aleatorias convergen en probabilidad a (respectivamente)* a, b *se verifica:*

$$A_n X_n + B_n \xrightarrow{\mathcal{L}} aX + b$$

A.4. Ordenes de convergencia en probabilidad

En Análisis Matemático, se distinguen órdenes de convergencia. Por ejemplo, cuando $n \to \infty$ se dice que $a_n = n^2(1/n)$ tiende a infinito con orden $O(n)$, o "es $O(n)$". Ello significa que existe alguna constante $M > 0$ para la cuál

$$\lim_{n\to\infty} a_n = Mn$$

("la sucesión $\{a_n\}$ va a infinito a la misma velocidad que n"). Una sucesión sería de orden $o(n)$ si en la expresión anterior M fuera 0. En general podemos emplear cualquier función $f(n)$ conveniente como patrón de comparación y decir que una sucesión es $O(f(n))$ o $o(f(n))$.

Esto puede generalizarse al caso de sucesiones aleatorias del siguiente modo: decimos que X_n es $O_p\left(f(n)\right)$ si para todo $\epsilon > 0$ existe $M_\epsilon < \infty$ tal que,

$$\text{Prob}\left\{|X_n| \leq M_\epsilon f(n)\right\} \geq 1 - \epsilon \tag{A.5}$$

("tomando términos lo suficientemente avanzados de la sucesión, la probabilidad de que queden acotados por $M_\epsilon f(n)$ puede hacerse tan cercana a uno como deseemos".)

De manera análoga se define que $\{X_n\}$ es $o_p(f(n))$ si

$$\text{plim}_{n\to\infty}\frac{X_n}{f(n)} = 0. \tag{A.6}$$

Ejemplo A.4 Sea $\{X_n\}$ una sucesión de observaciones independientes e idénticamente distribuidas, procedentes de una distribución con media m y varianza σ^2. Construyamos la sucesión $\{Z_n\}$ de medias aritméticas, $Z_n = (X_1 + \cdots + X_n)/n$. Entonces, $E[Z_n] = m$ y Var$(Z_n) = n^{-1}\sigma^2$. De acuerdo con la desigualdad de Tchebichev,

$$\text{Prob}\left\{|Z_n - m| < k\sigma n^{-\frac{1}{2}}\right\} \geq 1 - \frac{1}{k^2}. \tag{A.7}$$

Es decir, con probabilidad tan grande como queramos —k es arbitraria— la variable aleatoria $(Z_n - m)$ queda acotada superiormente por el producto de una constante ($k\sigma$, jugando el papel de M_ϵ en (A.5)) y una función ($n^{-\frac{1}{2}}$, jugando el papel de $f(n)$). Podemos decir entonces que $(Z_n - m)$ es $O_p(n^{-\frac{1}{2}})$.

Observese que si una sucesion $\{X_n\}$ es $O_p(n^k)$, también es $O_p(n^{k+\delta})$ para todo $\delta > 0$. La función $f(n)$ en la definición (A.5) es una función que, multiplicada por la constante, M_ϵ basta para acotar con probabilidad $1 - \epsilon$. No se requiere que $f(n)$ en (A.5) sea la más ajustada de las posibles.

Ejemplo A.5 Sea una sucesión $\{X_n\}$ que converge en probabilidad a X. Entonces la sucesión aleatoria cuyo término general es $(X_n - X)$ es $o_p(1)$. En efecto,

$$\text{plim}X_n = X \Longleftrightarrow \text{plim}\frac{(X_n - X)}{1} = 0 \Longleftrightarrow (X_n - X) = o_p(1)$$

Obsérvese que todas las sucesiones que convergen en probabilidad son cuando menos $o_p(1)$, pero algunas tendrán un orden de convergencia más rápido. En el ejemplo anterior vimos que en la situación habitual de una distribución que posee momentos de primer y segundo orden, la media aritmética de un número creciente de observaciones converge en probabilidad a la media poblacional y $(Z_n - m)$ converge en probabilidad a cero. Vimos que $(Z_n - m)$ es $O_p(n^{-\frac{1}{2}})$. No es en cambio $o_p(n^{-\frac{1}{2}})$; Es fácil ver que $(Z_n - m)$ es $o_p(n^{-\frac{1}{2}+\delta})$ para cualquier δ positivo. Esta es la situación habitual con sucesiones estimadoras paramétricas; se denominan por ello $\sqrt{n}$-*consistentes*. Ocasionalmente se presentan convergencias más rápidas. En estimación no paramétrica, en cambio, son la regla convergencias más lentas.

Las notaciones $O_p()$ y $o_p()$ funcionan de modo enteramente similar a sus correspondientes $O()$ y $o()$ no aleatorias. Por ejemplo, si dos sucesiones aleatorias son respectivamente de órdenes $o_p(n^{-1})$ y $O_p(n^{\frac{1}{2}})$, la sucesión obtenida multiplicando ambas elemento a elemento sería $o_p(n^{-\frac{1}{2}})$.

Análogamente, si $g()$ es una función continua y $\{X_n\} \xrightarrow{p} X$ de suerte que $(X_n - X)$ es $o_p(f(n))$, entonces $(g(X_n) - g(X))$ es $o_p(f(n))$. Pueden verse los resultados al respecto y más detalles en Mann y Wald (1943).

A.5. Leyes de grandes números

Dada una sucesión $\{X_n\}$ de v.a., no necesariamente equidistribuídas, pero con media común, las leyes de grandes números prescriben, bajo diferentes conjuntos de condiciones, la convergencia de $\overline{X}_n$ definida como en el Ejemplo A.1 a la media común $m = E[X_i]$. Esta convergencia puede ser de varios tipos: en probabilidad —y entonces decimos hallarnos ante una *ley débil de grandes números*— o casi seguramente —y entonces hablamos de una *ley fuerte de grandes números*[2]—. Enunciaremos en lo que sigue varios teoremas que establecen convergencias fuertes y débiles en diferentes circunstancias.

A.5.1. Leyes débiles de grandes números.

Una de las versiones más simples (y también más frecuentemente utilizadas) de ley débil de grandes números es la siguiente:

Teorema A.2 *Si la sucesión $\{X_n\}$ esta formada por v.a. independientes e idénticamente distribuídas, con media común m y varianza común σ^2, entonces:*

$$\overline{X}_n \xrightarrow{p} m$$

DEMOSTRACION:

Sea,

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n}$$

Entonces:

$$\begin{aligned} E[\overline{X}_n] &= m \\ \sigma^2_{\overline{X}_n} &= \frac{\sigma^2}{n} \end{aligned}$$

y de acuerdo con la desigualdad de Tchebychev:

$$\text{Prob}\left\{ |\overline{X}_n - m| < k\frac{\sigma}{\sqrt{n}} \right\} \geq 1 - \frac{1}{k^2}$$

[2]También se considera a veces convergencia en media cuadrática, que no hemos examinado aquí. Véase cualquiera de los textos citados más arriba.

Fácilmente se ve que la anterior desigualdad implica (A.1) para $\epsilon > 0, \delta > 0$ prefijados. Basta tomar $k > \delta^{-1/2}$, y $N(\epsilon, \delta)$ lo suficientemente grande como para que:

$$k\frac{\sigma}{\sqrt{N(\epsilon,\delta)}} < \epsilon$$

Las condiciones anteriores pueden ser considerablemente relajadas; no es imprescindible que las v.a. en la sucesión sean independientes, ni que tengan la misma varianza (sería suficiente que se verificase $\lim_{n\to\infty} n^{-2} \sum_{i=1}^{n} \sigma_{X_i}^2 < \infty$).

A.5.2. Leyes fuertes de grandes números

No sólo las condiciones en el Teorema A.2 pueden relajarse, sino que la conclusión puede a su vez reforzarse, dando lugar a una ley fuerte de grandes números. Antes de enunciarla, demostraremos algunos resultados que precisamos[3].

Teorema A.3 (primera desigualdad de Kolmogorov) *Sea $\{X_n\}$ una sucesión de v.a. independientes con media 0 y varianzas (no necesariamente iguales) finitas. Sea,*

$$S_n = X_1 + \ldots + X_n$$

Para cualquier $\epsilon > 0$ se verifica:

$$Prob\left\{\max_{1\le k\le n} |S_k| \ge \epsilon\right\} \le \frac{E[S_n^2]}{\epsilon^2} \tag{A.8}$$

DEMOSTRACION:

Definamos para $1 \le k \le n$ los sucesos

$$A_k = \{\omega : (|S_k(\omega)| \ge \epsilon) \cap (|S_i(\omega)| < \epsilon, 1 \le i < k)\}$$

("la suma parcial formada por k sumandos es la primera que excede en valor absoluto de ϵ"). Sea $A_0 = \{\omega : (|S_k(\omega)| < \epsilon, 1 \le k < n)\}$ ("la suma parcial formada por k sumandos nunca excede de ϵ").

[3]El desarrollo sigue el efectuado por Fourgeaud y Fuchs (1967), pág. 45 y ss. y Billingsley (1986), pág. 296.

Los sucesos $A_0, \ldots, A_n$ son disjuntos, y podemos calcular $E[S_n^2]$ así ($f_{\boldsymbol{X}}(\boldsymbol{x})$ es la función de densidad marginal que proceda):

$$\begin{aligned} E[S_n^2] &= \sum_{k=0}^{n} \int_{A_k} S_n^2 f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \\ &\geq \sum_{k=1}^{n} \int_{A_k} [S_k + (S_n - S_k)]^2 f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \\ &= \sum_{k=1}^{n} \int_{A_k} [S_k^2 + (S_n - S_k)^2 + 2S_k(S_n - S_k)] f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \\ &\geq \sum_{k=1}^{n} \int_{A_k} [S_k^2 + 2S_k(S_n - S_k)] f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \end{aligned}$$

Pero S_k y $(S_n - S_k)$ son v.a. independientes y de media 0, y por tanto:

$$\sum_{k=1}^{n} \int_{A_k} 2S_k(S_n - S_k) f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} = 0$$

En consecuencia:

$$E[S_n^2] \geq \sum_{k=1}^{n} \int_{A_k} S_k^2 f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \geq \sum_{k=1}^{n} \epsilon^2 \text{Prob}\{A_k\}$$

desigualdad equivalente a (A.8).

Teorema A.4 (Kintchine-Kolmogorov) *Si $\{X_n\}$ es una sucesión de v.a. centradas, independientes y con momento de orden dos finito, y se verifica además que $\sum_{i=1}^{\infty} \sigma_i^2 < \infty$, entonces $S_n = \sum_{i=1}^{n} X_i$ converge casi seguramente.*

DEMOSTRACION:

Si S_n converge casi seguramente, quiere decir que casi seguramente verifica la condición de convergencia de Cauchy. Es decir, $|S_{n+k} - S_n| \xrightarrow{c.s.} 0$, para $n, k \to \infty$. Para que no hubiera convergencia de $S_n(\omega)$, debería ocurrir que existiera $\epsilon > 0$ tal que $\forall n \geq 1$ hubiera algún $k \geq 1$ para el que $|S_{n+k} - S_n| \geq \epsilon$. Vamos a comprobar que el conjunto $D = \{\omega\}$ para el que se verifica lo anterior tiene probabilidad cero. Tenemos que:

$$D = \bigcup_{\epsilon > 0} \left[\bigcap_{n \geq 1} \bigcup_{k \geq 1} \{\omega : |S_{n+k} - S_n| > \epsilon\} \right] = \bigcup_{\epsilon > 0} L(\epsilon)$$

en que $L(\epsilon)$ es el suceso entre corchetes. Entonces,

$$\begin{aligned} \text{Prob}\{L(\epsilon)\} &= \text{Prob}\left\{\bigcap_{n\geq 1}\bigcup_{k\geq 1}[\omega : |S_{n+k} - S_n| > \epsilon]\right\} && \text{(A.9)} \\ &\leq \min_n\left[\text{Prob}\left\{\omega : \max_{k\geq 1}|S_{n+k} - S_n| > \epsilon\right\}\right] && \text{(A.10)} \\ &\leq \min_n\left[\frac{1}{\epsilon^2}\sum_{\ell\geq n+1}\sigma_\ell^2\right]. && \text{(A.11)} \end{aligned}$$

En el último paso se ha hecho uso de la primera desigualdad de Kolmogorov. Como $\sum_{i=1}^{\infty}\sigma_i^2 < \infty$, (A.11) es cero, Prob $\{L(\epsilon)\} = 0$ y por consiguiente $D = \bigcup_{\epsilon>0} L(\epsilon)$ tiene también probabilidad cero.

El siguiente lema no tiene ningún contenido probabilístico, y se limita a establecer una relación entre la convergencia (en el sentido habitual del Análisis Matemático) de dos diferentes series.

Lema A.1 *Si $\{a_i\}$ es una sucesión de números reales y $\sum_{i=1}^{n} a_i/i$ converge a un límite finito ℓ, entonces $n^{-1}\sum_{i=1}^{n} a_i$ converge a cero.*

DEMOSTRACION:

Sea $v_n = \sum_{i=1}^{n} a_i/i$, y $v_0 = 0$. Entonces, $a_i = i(v_i - v_{i-1})$ y:

$$\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} iv_i - \sum_{i=1}^{n} iv_{i-1} = nv_n - \sum_{i=0}^{n-1} v_i$$

Por tanto:

$$\frac{1}{n}\sum_{i=1}^{n} a_i = v_n - \frac{1}{n}\sum_{i=0}^{n-1} v_i = v_n - \frac{n-1}{n}\frac{1}{n-1}\sum_{i=0}^{n-1} v_i$$

y si $v_n \to \ell$, $(n-1)^{-1}\sum_{i=0}^{n-1} v_i \to \ell$ y $n^{-1}\sum_{i=1}^{n} a_i \to 0$.

Podemos ya, con ayuda de los resultados precedentes, establecer la siguiente ley fuerte de grandes números:

Teorema A.5 (ley fuerte de grandes números) *Sea $\{X_n\}$ una sucesión de v.a. independientes centradas, con momento de segundo orden finito, y $\sum_{i=1}^{\infty}\sigma_i^2/i^2 < \infty$. Entonces:*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{c.s.} 0$$

Demostraremos que $\sum_{i=1}^{n} X_i/i \xrightarrow{c.s.} \ell$, pues esto, en virtud del lema precedente, implica $n^{-1}\sum_{i=1}^{n} X_n \xrightarrow{c.s.} 0$. Que la primera serie converge c.s. es inmediato, pues como $\mathrm{Var}(X_i/i) = \sigma_i^2/i^2$ y $\sum_{i=1}^{\infty} \sigma_i^2/i^2 < \infty$, su convergencia es resultado del Teorema A.4

CUESTIONES, COMPLEMENTOS Y COSAS PARA HACER

A.1 Demuéstrese que, en el caso particular en que una sucesión aleatoria converge en distribución a una constante, es decir $X_n \xrightarrow{\mathcal{L}} c$, entonces $X_n \xrightarrow{p} c$.

A.2 Compruébese que $X_n \xrightarrow{m.c.} X \Rightarrow X_n \xrightarrow{p} X$. (Ayuda: Hágase uso de la desigualdad de Tchebichev.)

Apéndice B

Soluciones a problemas seleccionados

3.2 La función de verosimilitud es

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = H(x_{(1)} - 1)H(x_{(n)} + 1)$$

en que $H(.)$ es una función que toma el valor cero si su argumento es negativo y valor 1 si su argumento es no negativo (función "escalón" o de Heaviside). Por tanto, el teorema de factorización (ver (3.8)) se verifica con $g(s,\theta) = H(x_{(1)}-1)H(x_{(n)}+1)$ y $(x_{(1)}, x_{(n)})$ forman un estadístico suficiente.

Sin embargo, este estadístico no es completo: es fácil ver que (por ej.) $(x_{(n)} - x_{(1)})$ tiene una distribución que no depende de θ y es por tanto ancilar.

3.5 En efecto,

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) \quad = \quad \prod_{i=1}^{n} \exp\{\theta x_j\} \exp\left\{-e^{\theta x_j} y_j\right\} = \exp\left\{-\sum_{i=1}^{n} \exp\{\theta x_j\} y_j + \theta \sum_{i=1}^{n} x_j\right\},$$

que no es de rango completo.

4.5 Es fácil encontrar un estadístico suficiente empleando el teorema de factorización:

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^{n} \theta x^{\theta-1} = \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1};$$

vemos que $\prod_{i=1}^{n} x_i$ (o alternativamente $\sum_{i=1}^{n} \log x_i$) es un estadístico suficiente.

Para comprobar que $Z = -\log X_1$ es insesgado para θ^{-1}, veamos cual es su distribución. La de X es $F_{X|\theta}(x\,|\theta) = x^{\theta}$. Entonces,

$$\begin{aligned}
\text{Prob}\,\{Z \le z\} &= \text{Prob}\,\{-\log(X) \le z\} \\
&= \text{Prob}\,\{\log(X) > -z\} \\
&= \text{Prob}\,\left\{X > e^{-z}\right\} \\
&= 1 - \text{Prob}\,\left\{X \le e^{-z}\right\} \\
&= 1 - e^{-z\theta};
\end{aligned}$$

derivando, $f_{X|\theta}(x\,|\theta) = \theta e^{-z\theta}$, en la que reconocemos una exponencial de media θ^{-1}. Por tanto, $Z = -\log X_1$ es efectivamente insesgado.

Vemos además que $T = -n^{-1}\sum_{i=1}^{n}\log X_i$ será también insesgado, y es función de un estadístico suficiente. Es claro entonces que T será insesgado de varianza mínima.

5.5 Calculemos en primer lugar la cota de Cramér-Rao para el estimador proporcionado. En los cálculos que siguen, $\theta = (\mu, \sigma^2)$ y tratamos a σ^2 como un parámetro respecto del cual derivamos.

$$\begin{aligned}
f_{X|\theta}(x\,|\theta) &= \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2} \\
\log f_{X|\theta}(x\,|\theta) &= -\frac{1}{2}\log\sigma^2 - \log\sqrt{2\pi} - (x-\mu)^2/2\sigma^2 \\
\frac{\partial}{\partial\sigma^2}\log f_{X|\theta}(x\,|\theta) &= -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4} \\
E\left[\frac{\partial}{\partial\sigma^2}\log f_{X|\theta}(x\,|\theta)\right]^2 &= E\left[\left(\frac{(x-\mu)^2}{2\sigma^4}\right)^2 + \left(\frac{1}{2\sigma^2}\right)^2 - 2\frac{1}{2}\frac{1}{2\sigma^4}(x-\mu)^2\right] \\
&= \frac{\mu_4}{4\sigma^8} + \frac{1}{4\sigma^4} - \frac{1}{2\sigma^4} \qquad \text{(B.1)}
\end{aligned}$$

Teniendo en cuenta que μ_{2k}, el momento centrado de orden $2k$, en una distribución normal toma el valor $\sigma^{2k}(2k)!2^{-k}(k!)^{-1}$, tenemos sustituyendo μ_4 en (B.1) que:

$$E\left[\frac{\partial}{\partial\sigma^2}\log f_{X|\theta}(x\,|\theta)\right]^2 = \frac{\sigma^4 4!}{4\cdot 4\cdot 2!\cdot\sigma^8} + \frac{1}{4\sigma^4} - \frac{1}{2\sigma^4} = \frac{1}{2\sigma^4}. \quad \text{(B.2)}$$

La cota de Cramér-Rao es por tanto $1/nI_X(\theta) = 2\sigma^4/n$.

Calculemos ahora la varianza del estimador. Para ello requerimos los momentos $E[S^2]$ y $E[(S^2)^2]$. Sabemos que $E[S^2] = \sigma^2$ —el S^2 proporcionado es el habitual estimador insesgado de la varianza—. Por otra parte, viendo $\sum_{i=1}^{n}(X_i - \overline{X})^2$ como la suma de cuadrados de los residuos cuando regresamos X sobre la columna de "unos", por teoría básica de regresión lineal sabemos que se distribuye como $\sigma^2\chi^2_{n-1}$. Entonces,

$$\begin{aligned}
E[S^2] &= \frac{\sigma^4}{(n-1)^2}E[\chi^2_{n-1}]^2 \\
&= \frac{\sigma^4}{(n-1)^2}E[Z_1^2 + \ldots Z_{n-1}^2]^2 \\
&= \frac{\sigma^4}{(n-1)^2}E\left[Z_1^4 + \ldots Z_{n-1}^4 + \sum_i\sum_{j\neq i} Z_i^2 Z_j^2\right], \qquad \text{(B.3)}
\end{aligned}$$

en que $Z_1, \ldots, Z_{n-1}$ son variables aleatorias $N(0,1)$. Sabiendo que el momento de orden cuatro de tal distribución tiene la expresión indicada antes y sustituyendo en

(B.3) obtenemos:

$$\begin{aligned} E[S^2] &= \frac{\sigma^4}{(n-1)^2}\left[(n-1)\cdot 3 + (n-1)(n-2)\right] \\ &= \frac{\sigma^4(n+1)(n-1)}{(n-1)^2} \\ &= \frac{\sigma^4(n+1)}{(n-1)}. \end{aligned}$$

Por consiguiente, la varianza buscada es:

$$\text{Var}(S^2) = E[(S^2)^2] - [E(S^2)]^2 = \frac{\sigma^4(n+1)}{(n-1)} - \sigma^4 = \frac{2\sigma^4}{n-1}. \qquad \text{(B.4)}$$

Comparando ahora las expresiones (B.4) y (B.2) llegamos a la conclusión de que la varianza del estimador no alcanza la cota de Cramér-Rao, pero la diferencia tiende a cero al crecer n.

Bibliografía

Abramson, N. (1966). *Teoría de la Información y Codificación*. Paraninfo, Madrid, 1973ª edón.

Akaike, H. (1969). Fitting Autoregressive Models for Prediction. *Annals of the Institute of Statistical Mathematics*, vol. 21, págs. 243–247.

Akaike, H. (1970). Statistical Predictor Identification. *Annals of the Institute of Statistical Mathematics*, vol. 22, págs. 203–217.

Akaike, H. (1972). Use of an Information Theoretic Quantity for Statistical Model Identification. En *Proc. 5th. Hawai Int. Conf. on System Sciences*, págs. 249–250.

Akaike, H. (1974). Information Theroy and an Extension of the Maximum Likelihood Principle. En *Second International Symposium on Information Theory* (eds. B. Petrov y F. Csaki), págs. 267–281. Akademia Kiado, Budapest. Reimpreso en Johnson-Kotz(1991), vol. 1, p. 610 y ss.

Akaike, H. (1991). Information Theory and an Extension of the Maximum Likelihood Principle. En *Breakthroughs in Statistics* (eds. Johnson y Kotz), vol. 1, pág. 610 y ss. Springer Verlag.

Berkson, J. (1980). Minimum chi.square, not maximum likelihood! *Annals of Statistics*, vol. 8, págs. 457–487.

Billingsley, P. (1986). *Probability and Measure*. John Wiley and Sons, New York, 2ª edón.

Chaitin, G. (1987). *Algorithmic Information Theory*. Cambridge University Press, Cambridge, 1992ª edón.

Cover, T., P. Gacs, y R. Gray (1989). Kolmogorov's contributions to information theory and algorithmic complexity. *Annals of Probability*, vol. 17(3), págs. 840–865.

Cox, D. R. y D. V. Hinkley (1974). *Theoretical Statistics*. Chapman and Hall, London, 1979ª edón.

Cramér, H. (1960). *Métodos Matemáticos de Estadística*. Ed. Aguilar, Madrid, 1970ª edón.

Cullman, G., M. Denis-Papin, y A. Kaufmann (1967). *Elementos de Cálculo Informacional*. Ed. Urmo, Bilbao, 1967ª edón.

D'Agostino, R. (1971). An Omnibus Test of Normality for Moderate and Large Sample Sizes. *Biometrika*, vol. 58, págs. 341–348.

de Leeuw, J. (2000). Information Theroy and an Extension of the Maximum Likelihood Principle by Hirotugu Akaike. Disponible en `http://www.stat.ucla.edu/~deleeuw/work/research.phtml`.

Dempster, A., N. Laird, y D. Rubin (1976). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, vol. 39, págs. 1–38.

Dowe, D., K. Korb, y J. Oliver (eds.) (1996). *Information, Statistics and Induction in Science – ISIS'96*, Melbourne, Australia. World Scientific, Singapore.

Fourgeaud, C. y A. Fuchs (1967). *Statistique*. Dunod, Paris.

Garín, A. y F. Tusell (1991). *Problemas de Probabilidad e Inferencia Estadística*. Ed. Tébar-Flores, Madrid.

Garthwaite, P., I. Jolliffe, y B. Jones (1995). *Statistical Inference*. Prentice Hall, London.

Gell-Mann, M. (1994). *El quark y el jaguar*. Tusquets, Barcelona, 1995ª edón.

G.J.McLachlan y T. Krishnan (1997). *The EM Algorithm and Extensions*. Wiley.

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

Kiefer, J. C. (1983). *Introduction to Statistical Inference*. Springer-Verlag, New York, 1987ª edón. (ed. Gary Lorden).

Laird, N. (1993). The EM algorithm. En *Handbook of Statistics*, vol. IX, págs. 509–520.

Lange, K. (1998). *Numerical Analysis for Statisticians*. Springer. Signatura: 519.6 LAN.

Lehmann, E. L. (1959). *Testing Statistical Hypothesis*. Wiley, New York.

Lehmann, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

Levy, M. (1985). A note on nonunique MLEs and sufficient statistics. *Annals of Mathematical Statistics*, vol. 39, págs. 66.

Li, M. y P. Vitányi (1993). *An introduction to Kolmogorov complexity and its applications*. Springer-Verlag, New York.

Mallows, C. (1973). Some comments on C_p. *Technometrics*, vol. 15, págs. 661–675.

Mann, H. y A. Wald (1943). On stochastic limit and order relationships. *Annals of Mathematica Statistics*, vol. 14, págs. 217–226.

Meeden, G. y S. Varderman (1985). Bayes and admissible set estimation. *Journal of the American Statistical Association*, vol. 80, págs. 465–471.

Navidi, W. (1997). A Graphical Illustration of the EM Algorithm. *Annals of Mathematical Statistics*, vol. 51(1), págs. 29–31.

Quenouille, M. (1956). Notes on bias estimation. *Biometrika*, vol. 43, págs. 353–360.

Rao, C. R. (1962). Efficient Estimates and Optimum Inference Procedures in Large Samples. *Journal of the Royal Statistical Society, Ser. B*, vol. 24, págs. 46–72.

Rao, C. R. (1965). *Linear Statistical Inference and its Applications*. Wiley, New York.

Rissanen, J. (1983). A Universal Prior for Integers and Estimation by Minimum Description Length. *Annals of Statistics*, vol. 11(2), págs. 416–431.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore.

Romano, J. P. y A. F. Siegel (1986). *Counterexamples in Probability and Statistics*. Wadsworth and Brooks/Cole, Monterrey, California.

Ruelle, D. (1991). *Chance and Chaos*. Penguin, London.

Russell, B. (1912). *The problems of philosophy*. Oxford University Press, 1989ª ed$^{\underline{\text{ón}}}$.

Shannon, C. (1948). The mathematical theory of communication. *Bell System Tech. Journal*, vol. 27, págs. 379–423, 623–656.

Shannon, C. y W. Weaver (1949). *The mathematical theory of communication*. University of Illinois Press, Urbana. Eight reprint, 1980.

Shapiro, S. y R. Francia (1972). An Approximate Analysis of Variance Test for Normality. *Journal of the American Statistical Association*, vol. 67, págs. 215–216.

Trocóniz, A. F. (1987). *Probabilidades. Estadística. Muestreo.* Tebar-Flores, Madrid.

Wang, C. (1993). *Sense and Nonsense of Statistical Inference*. Marcel Dekker, New York.

Young, G. y R. Smith (2005). *Essentials of Statistical Inference*. Cambridge Univ. Press. Signatura: 519.22 YOU.

Índice alfabético