Capítulo 5: Teoría de Decision Estadística

Teoría Estadística Avanzada

Sigla DES124

Prof. Jaime Lincovil

5.1. Introducción

Una parte importante del análisis estadístico es tomar decisiones bajo incertidumbre. En muchos casos, hay un costo asociado a tomar decisiones incorrectas, por lo que puede ser una buena estrategia incorporar estos costos en el análisis estadístico y buscar la decisión que minimice (en algún sentido) el costo esperado. Este es el enfoque de la *teoría de decisión estadística*.

La Teoría de Decisión es parte del enfoque más general de la teoría de juegos, que se originó con von Neumann y Morgenstern en la década de 1950 en un contexto económico. En el contexto de la teoría de juegos, hay dos (o más) jugadores compitiendo entre sí y la perspectiva típica es que la victoria de un jugador representa una pérdida para el otro. Dado que ninguno de los jugadores conoce en general la estrategia que tomará el otro, el objetivo de cada jugador es elegir una estrategia que le garantice que no perderá demasiado, en cierto sentido. La película *A Beautiful Mind*, inspirada en la vida del matemático John F. Nash, destaca el desarrollo de su *equilibrio de Nash*, un resultado en teoría de juegos que ahora se enseña a estudiantes de economía.

En el contexto de la teoría de decisión estadística, los jugadores son el *Estadístico* y la *Naturaleza*, un personaje hipotético que conoce el valor verdadero del parámetro.

El planteamiento de la teoría de decisión estadística comienza con los ingredientes familiares: hay un espacio muestral $(X, \mathcal{A})$, un espacio de parámetros Θ, y una familia de medidas de probabilidad $\{P_\theta : \theta \in \Theta\}$ definidas en $(X, \mathcal{A})$ indexadas por Θ. En algunos casos, también puede haber una distribución a priori Π en $(\Theta, \mathcal{B})$, donde $\mathcal{B}$ es una σ-álgebra en Θ, aunque esto no siempre es necesario. Los dos "nuevos" ingredientes son los siguientes:

- Un espacio de acciones $\mathcal{A}$. Cuando consideramos los resultados del análisis estadístico como la determinación de una acción a tomar o una decisión a ser realizada por el tomador de decisiones, entonces debe existir un conjunto de todas esas acciones.
- Una función de pérdida (no negativa) $L(\theta, a)$ definida en $\Theta \times \mathcal{A}$. Esta función tiene el propósito de representar los "costos" de tomar decisiones incorrectas. En particular,

$L(\theta, a)$ representa el costo de tomar la acción a cuando el parámetro es θ.

Ejemplo 5.1: Prueba de hipótesis

En un problema de prueba de hipótesis, podemos considerar el espacio de parámetros como $\Theta = \{0, 1\}$, donde "0" significa que H_0 es verdadero y "1" significa que H_1 es verdadero.
El espacio de acciones también es $\mathcal{A} = \{0, 1\}$, donde 0 corresponde a "aceptar H_0" y 1 corresponde a "rechazar H_0". Una función de pérdida típica en este caso es la llamada pérdida 0-1, es decir,

$$L(0,0) = L(1,1) = 0 \quad \text{y} \quad L(1,0) = L(0,1) = 1.$$

Es decir, las decisiones correctas no tienen costo, pero los errores de Tipo I y Tipo II tienen un costo de 1 unidad cada uno. Sin embargo, no siempre es el caso que los errores de Tipo I y Tipo II tengan el mismo costo; es fácil extender la función de pérdida para considerar estos casos.

Ejemplo 5.2: Estimación puntual

Supongamos que el objetivo es estimar $\psi(\theta)$, donde θ es desconocido pero ψ es una función real conocida. Entonces, $\mathcal{A} = \psi(\Theta)$ es la imagen de Θ bajo ψ. La función de pérdida típica es la pérdida de error cuadrático, es decir,

$$L(\theta, a) = (a - \psi(\theta))^2.$$

Sin embargo, también se pueden considerar otras funciones de pérdida como $L(\theta, a) = |\psi(\theta) - a|$.
Ahora supongamos que se observa un dato $X \sim P_\theta$. Nos gustaría usar la información $X = x$ para ayudar a elegir una acción en $\mathcal{A}$ a tomar. Una elección de acción en $\mathcal{A}$ basada en los datos x se denomina *regla de decisión*.

Definición 5.1

Una *regla de decisión no aleatorizada* δ es una función que mapea $\mathcal{X}$ en $\mathcal{A}$, y la pérdida incurrida al usar $\delta(x)$ está dada simplemente por $L(\theta, \delta(x))$.
Una *regla de decisión aleatorizada* δ es una función que mapea $\mathcal{X}$ en el conjunto de medidas de probabilidad definidas sobre $\mathcal{A}$. En este caso, la pérdida incurrida al usar $\delta(x)$ está dada por la esperanza

$$L(\theta, \delta(x)) = \int_{\mathcal{A}} L(\theta, a)\, \delta(x)(da).$$

Una regla de decisión no aleatorizada δ es un caso especial de una regla aleatorizada, donde $\delta(x)$ se considera como una masa puntual en el número $\delta(x) \in \mathcal{A}$.

Estamos familiarizados con el concepto de reglas de decisión aleatorizadas en el contexto de las pruebas de hipótesis. En este escenario, recordemos que en algunos problemas (usualmente discretos), no es posible alcanzar un error de Tipo I específico con una prueba no aleatorizada. La idea, entonces, es lanzar una moneda con probabilidad $\delta(x) \in [0, 1]$ para decidir entre aceptar y rechazar. Por razones obvias, las reglas de decisión no aleatorizadas son preferidas sobre las aleatorizadas. Mostraremos más adelante (Teorema 2) que, para funciones de pérdida .ªdecuadas", podemos ignorar de manera segura las reglas de decisión aleatorizadas.

Dado un conjunto de acciones $\mathcal{A}$, una función de pérdida L y un dato x, el objetivo de la teoría de decisión es elegir una regla de decisión δ que minimice $L(\theta, \delta(x))$. Sin embargo, esto típicamente no se puede hacer sin el conocimiento de θ. Para simplificar la tarea, buscaremos reglas de decisión que posean buenas propiedades en promedio, dado que $X \sim P_\theta$. En esta dirección, se define la *función de riesgo*:

$$R(\theta, \delta) = \int_{\mathcal{X}} L(\theta, \delta(x)) P_\theta(dx), \tag{1}$$

lo cual es simplemente la pérdida esperada incurrida al usar la regla de decisión δ. El objetivo de la teoría de decisión clásica es encontrar una regla de decisión δ que minimice $R(\theta, \delta)$ en algún sentido.

El problema es que, en general, no existe una única δ que minimice $R(\theta, \delta)$ para todos los θ. En tales casos, se buscan otras formas de minimizar el riesgo que sean menos restrictivas que la minimización uniforme. Estas incluyen diversas maneras de eliminar θ del riesgo, haciendo que dependa solo de δ (ver Sección 5.3) o introduciendo restricciones sobre δ (ver Sección 5.4). En estos casos, se puede hablar de una regla de decisión que minimiza el riesgo.

Como primer paso, es útil reducir la búsqueda a reglas de decisión que sean *admisibles*, lo cual discutimos a continuación en la Sección 5.2.

5.2. Admisibilidad

Al buscar una regla de decisión "óptima", puede ser útil descartar algunos procedimientos que se sabe que son subóptimos, reduciendo así el tamaño del espacio de búsqueda.

Definición 5.2

Una regla de decisión δ es *inadmisible* si existe otra regla de decisión δ' tal que $R(\theta, \delta') \leqslant R(\theta, \delta)$ para todo θ con desigualdad estricta para algún θ. Decimos que δ' *domina* a δ. Si no existe tal δ', entonces δ es *admisible*.

En términos generales, solo deben considerarse aquellas reglas de decisión que sean admisibles. Sin embargo, no todas las reglas de decisión admisibles son razonables. Por ejemplo, si el objetivo es estimar θ bajo pérdida cuadrática, la regla de decisión $\delta(x) \equiv \theta_0$ es admisible, ya que es la única regla de decisión con riesgo cero en $\theta = \theta_0$. Sin embargo, la regla $\delta(x) \equiv \theta_0$, que se enfoca únicamente en un único valor de θ, incurrirá en un alto costo en términos de riesgo si $\theta \neq \theta_0$.

Resulta que la admisibilidad de una regla de decisión está estrechamente relacionada con las propiedades de la función de pérdida. En particular, cuando la función de pérdida $L(\theta, a)$ es *convexa* en a, surgen algunas propiedades interesantes. A continuación, se presenta un resultado importante.

Teorema 5.1 Rao-Blackwell

Sea $X \sim P_\theta$ y sea T un estadístico suficiente. Sea δ_0 una regla de decisión no aleatorizada, tomando valores en un conjunto convexo $\mathcal{A} \subseteq \mathbb{R}^d$, con $\mathbb{E}_\theta[\|\delta_0(X)\|] < \infty$ para todo θ. Si $\mathcal{A}$ es convexa y $L(\theta, \cdot)$ es una función convexa para cada θ, entonces

$$\delta_1(x) = \delta_1(t) = \mathbb{E}[\delta_0(X) \mid T = t]$$

satisface $R(\theta, \delta_1) \leqslant R(\theta, \delta_0)$ para todo θ.
Demostración. De la desigualdad de Jensen,

$$L(\theta, \delta_1(t)) \leqslant \mathbb{E}[L(\theta, \delta_0(X)) \mid T = t] \quad \forall \theta.$$

Tomando la esperanza en ambos lados (con respecto a la distribución de T bajo $X \sim P_\theta$), se obtiene que $R(\theta, \delta_1) \leqslant R(\theta, \delta_0)$. ■

Este teorema muestra que, para una función de pérdida convexa, solo las reglas de decisión que son funciones de estadísticos suficientes pueden ser admisibles. Además, muestra cómo mejorar una regla de decisión dada: simplemente *"Rao-Blackwellízala"* tomando la esperanza condicional dado un estadístico suficiente.

Ejemplo 5.3

Supongamos que $X_1, \ldots, X_n$ son variables aleatorias independientes $N(\theta, 1)$. El objetivo es estimar $\Phi(c - \theta)$, la probabilidad de que $X_1 \leqslant c$, para algún valor constante c, bajo pérdida cuadrática. Es decir, $L(\theta, a) = (a - \Phi(c - \theta))^2$.
Un estimador directo es $\delta_0(x) = \frac{1}{n}\sum_{i=1}^n I_{(-\infty,c]}(X_i)$. Sin embargo, este no es una función del estadístico suficiente $T = \bar{X}$. Dado que $L(\theta, \cdot)$ es convexa, el teorema de Rao-Blackwell nos dice que podemos mejorar δ_0 tomando su esperanza condicional dado $T = t$. Es decir,

$$\mathbb{E}[\delta_0(X) \mid T = t] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[I_{(-\infty,c]}(X_i) \mid T = t]$$

$$= P(X_1 \leqslant c \mid T = t) = \Phi\left(\frac{c - t}{\sqrt{(n-1)/n}}\right), \tag{2}$$

donde la última igualdad se obtiene del hecho de que $X_1 \mid (T = t) \sim N(t, n^{-1})$; ver Ejercicio 5.
Por lo tanto, δ_0 es inadmisible y el lado derecho de (5.2) es un estimador que la supera. Se deduce (casi) inmediatamente del teorema de Rao-Blackwell que, en el caso de una función de pérdida convexa, no es necesario considerar reglas de decisión aleatorizadas.

Teorema 5.2

Supongamos que la función de pérdida es convexa. Entonces, para cualquier regla de decisión aleatorizada δ_0, existe una regla no aleatorizada δ_1 que no es peor que δ_0 en términos de riesgo.
Demostración. Una regla de decisión aleatorizada δ_0 define una distribución sobre $\mathcal{A}$ en el sentido de que, dado $X = x$, se toma una acción muestreando $A \sim \delta_0(x)$. Podemos expresar esta regla aleatorizada como una no aleatorizada, denotémosla $\delta_0(X, U)$, que es una función de X y de una variable aleatoria independiente U, cuya distribución Q es independiente de θ.
La idea es escribir $A = f_x(U)$, para alguna función f_x dependiente de los datos. Es decir, U cumple el papel del mecanismo de aleatorización. Pero sabemos que $T(X, U) = X$ es suficiente, por lo que el teorema de Rao-Blackwell nos dice cómo mejorar δ_0:

$$\delta_1(x) = \mathbb{E}[\delta_0(X,U) \mid X = x] = \int f_x(u)Q(du) = \int_{\mathcal{A}} a\,\delta_0(x)(da).$$

Es decir, la regla no aleatorizada $\delta_1(x)$ que domina es simplemente la esperanza bajo $\delta_0(x)$. ■

Este resultado explica por qué casi nunca consideramos estimadores aleatorizados—en problemas de estimación, la función de pérdida (por ejemplo, el error cuadrático) es casi siempre convexa, por lo que el Teorema 2 indica que no es necesario considerar estimadores aleatorizados.

La admisibilidad es una propiedad interesante en el sentido de que solo se deben usar reglas admisibles, pero existen muchas reglas admisibles que son deficientes. Por lo tanto, la admisibilidad por sí sola no es suficiente para justificar el uso de una regla de decisión—compárese esto con el control de la probabilidad de error de Tipo I en un problema de prueba de hipótesis, donde también se deben considerar las probabilidades de error de Tipo II.

Existen otras propiedades generales de admisibilidad (por ejemplo, todas las reglas de Bayes propias son admisibles), que discutiremos más adelante. Curiosamente, hay algunos resultados sorprendentes que indican que algunos procedimientos .estándar"son, en ciertos casos, inadmisibles. El ejemplo más famoso de esto es la *paradoja de Stein*: si $X \sim N_d(\theta, I)$ y $d \geqslant 3$, entonces el estimador de verosimilitud máxima / mínimos cuadrados $\hat{\theta} = X$ es, de hecho, *inadmisible*.

5.3. Minimización de una medida "global" de riesgo

Encontrar δ que minimice $R(\theta, \delta)$ uniformemente sobre θ es una tarea imposible. El problema es que existen reglas de decisión absurdas que funcionan muy bien para ciertos valores de θ pero muy mal para otros.

Si introducimos una medida global de riesgo—una que no dependa de un solo valor de θ—entonces es un poco más fácil encontrar reglas de decisión óptimas. Las dos formas más comunes de eliminar θ del riesgo son integrarlo (riesgo promedio) o maximizarlo sobre θ (riesgo máximo), y cada una de estas tiene su propio desarrollo teórico.

5.3.1. Minimización del riesgo promedio

La primera forma en que se podría considerar eliminar θ de la función de riesgo $R(\theta, \delta)$ es promediándolo/integrándolo con respecto a alguna distribución de probabilidad Π sobre Θ. Esto lleva a la noción de *reglas de Bayes*.

Definición 5.3

Para un problema de decisión como el anterior, supongamos que también existe una medida de probabilidad Π en Θ. Entonces, para una regla de decisión δ, el *riesgo de Bayes* es

$$r(\Pi, \delta) = \int R(\theta, \delta)\, \Pi(d\theta), \tag{3}$$

que representa el riesgo promedio de δ con respecto a Π. Si existe una $\delta = \delta_\Pi$ que minimiza $r(\Pi, \delta)$, entonces δ_Π se llama la *regla de Bayes* (con respecto a Π).

En este contexto, la medida de probabilidad Π no necesariamente tiene el propósito de describir el conocimiento previo sobre θ. En su lugar, una Π adecuadamente elegida puede ayudar en la construcción de procedimientos (no bayesianos) con buenas propiedades. Esta es la gran idea detrás del trabajo moderno sobre estimación por contracción a través de penalizaciones basadas en distribuciones a priori.

Para encontrar reglas de Bayes, es importante notar que

$$r(\Pi, \delta) = \int_{\mathcal{X}} \int_{\Theta} L(\theta, \delta(x)) \Pi_x(d\theta) P_\Pi(dx),$$

donde la integral interna, $\int_\Theta L(\theta, \delta(x))\Pi_x(d\theta)$, se conoce como el *riesgo posterior*, y P_Π es la distribución marginal de X. Esto es una consecuencia del teorema de Fubini. Se puede demostrar que (en la mayoría de los casos) un minimizador del riesgo posterior también será la *regla de Bayes*. Esto es importante porque minimizar el riesgo posterior suele ser más sencillo.

Ejemplo 5.4

Consideremos la estimación de un parámetro real θ bajo pérdida cuadrática. Entonces, el riesgo posterior es

$$\int_\Theta (\theta - \delta(x))^2 \Pi_x(d\theta) = \mathbb{E}(\theta^2 \mid x) - 2\delta(x)\mathbb{E}(\theta \mid x) + \delta(x)^2.$$

De esto se deduce que (si todas las esperanzas existen) el riesgo posterior se minimiza tomando $\delta(x) = \mathbb{E}(\theta \mid x)$. Argumentos similares pueden usarse para demostrar que, si la función de pérdida es el error absoluto, entonces la regla de Bayes es la mediana posterior.

Ejemplo 5.5

Supongamos que $X_1, \ldots, X_n$ son observaciones independientes Ber(θ). El objetivo es probar $H_0 : \theta \leqslant 0{,}5$ frente a $H_1 : \theta > 0{,}5$ bajo pérdida 0-1.
Por simplicidad, asumamos que n es par y que $T = \sum_{i=1}^n X_i$. Si la distribución a priori es Unif$(0, 1)$, entonces la distribución a posteriori es Beta$(t + 1, n - t + 1)$.
El riesgo posterior, denotado $r_t(H_0)$, para elegir H_0 es la probabilidad a posteriori de H_1, es decir, $r_t(H_0) = \Pi_t(H_1)$. De manera similar, el riesgo posterior de elegir H_1 es la probabilidad a posteriori de H_0, es decir, $r_t(H_1) = \Pi_t(H_0) = 1 - r_t(H_0)$.
La regla de Bayes elige H_0 o H_1 dependiendo de cuál de $r_t(H_0)$ y $r_t(H_1)$ sea menor. Es decir, la regla de Bayes rechaza H_0 si y solo si $r_t(H_0) < 0{,}5$. Sin embargo, si $t = n/2$, entonces $r_t(H_0) = r_t(H_1) = 0{,}5$ y, por lo tanto, no está claro cuál elegir. Por lo tanto, la regla de Bayes es una regla aleatorizada dada por:

$$\delta(x) = \begin{cases} \text{Elegir } H_0 & \text{si } t(x) < n/2, \\ \text{Elegir } H_1 & \text{si } t(x) > n/2, \\ \text{Lanzar una moneda justa para decidir} & \text{si } t(x) = n/2. \end{cases}$$

Los principales resultados que consideraremos aquí indican que, bajo ciertas condiciones, las reglas de Bayes son admisibles. Por lo tanto, no se puede descartar ser bayesiano basándose únicamente en restricciones de admisibilidad.
Aún más interesante es que existen teoremas que establecen que, en esencia, todas las reglas de decisión admisibles son de Bayes; ver Sección 5.5.

Teorema 5.3

Supongamos que Θ es un subconjunto de $\mathbb{R}^d$ tal que cualquier vecindad de cada punto en Θ interseca el interior de Θ. Sea Π una medida en Θ tal que $\lambda \ll \Pi$, donde λ es la medida de Lebesgue. Supongamos que $R(\theta, \delta)$ es continua en θ para cada δ cuyo riesgo sea finito. Si la regla de Bayes δ_Π tiene riesgo finito, entonces es admisible.
Demostración. Supongamos que δ_Π no es admisible. Entonces, existe una δ_1 tal que $R(\theta, \delta_1) \leqslant R(\theta, \delta_\Pi)$ para todo θ con desigualdad estricta para algún θ_0.
Por continuidad de la función de riesgo, existe un vecindario abierto N de θ_0, que interseca el interior de Θ, tal que $R(\theta, \delta_1) < R(\theta, \delta_\Pi)$ para todo $\theta \in N$.
Dado que la medida de Lebesgue está dominada por Π y N es un conjunto abierto, debemos tener $\Pi(N) > 0$. Esto implica que $r(\Pi, \delta_1) < r(\Pi, \delta_\Pi)$, lo cual es una contradicción.
Por lo tanto, δ_Π es admisible. ■

Ejemplo 5.6

Consideremos una distribución de familia exponencial con espacio de parámetros naturales Θ que contiene un conjunto abierto. Para estimar $g(\theta)$ para alguna función continua g, consideremos la pérdida cuadrática $L(\theta, a) = (a - g(\theta))^2$.
Dado que Θ es convexa (Capítulo 2), la condición de vecindad se satisface. Además, la función de riesgo para cualquier δ con varianza finita será continua en θ.
Finalmente, si Π tiene una densidad positiva π sobre Θ con respecto a la medida de Lebesgue, entonces Lebesgue $\ll \Pi$ también se cumple. Por lo tanto, la regla de Bayes δ_Π es admisible.

Teorema 5.4

Supongamos que $\mathcal{A}$ es convexa y que todas las P_θ son absolutamente continuas entre sí. Si $L(\theta, \cdot)$ es estrictamente convexa para cada θ, entonces, para cualquier medida de probabilidad Π sobre Θ, la regla de Bayes δ_Π es admisible.
Demostración. Supongamos que δ_Π no es admisible. Entonces, existe δ_0 tal que $R(\theta, \delta_0) \leqslant R(\theta, \delta_\Pi)$ con desigualdad estricta para algún θ. Definamos una nueva regla de decisión $\delta_1(x) = \frac{1}{2}(\delta_\Pi(x) + \delta_0(x))$, lo cual es válido ya que $\mathcal{A}$ es convexa. Entonces, para todo θ tenemos:

$$R(\theta, \delta_1) = \int_{\mathcal{X}} L(\theta, \tfrac{1}{2}(\delta_\Pi(x) + \delta_0(x)))P_\theta(dx)$$

$$\leqslant \int_{\mathcal{X}} \tfrac{1}{2}\{L(\theta, \delta_\Pi(x)) + L(\theta, \delta_0(x))\}P_\theta(dx)$$

$$= \tfrac{1}{2}\{R(\theta, \delta_\Pi) + R(\theta, \delta_0)\}$$

$$\leqslant R(\theta, \delta_\Pi).$$

La primera desigualdad será estricta a menos que $\delta_\Pi(X) = \delta_0(X)$ con probabilidad P_θ igual a 1. Sin embargo, dado que las P_θ son absolutamente continuas entre sí, se sigue que la primera desigualdad es estricta a menos que $\delta_\Pi(X) = \delta_0(X)$ con probabilidad P_θ igual a 1 para todo θ.
Por lo tanto, la primera desigualdad anterior es estricta a menos que $\delta_\Pi(X)$ y $\delta_0(X)$ tengan exactamente la misma distribución. Dado que esto violaría la suposición de que δ_0 domina a δ_Π, se debe concluir que la primera desigualdad es estricta para todo θ. Es decir, $R(\theta, \delta_1) < R(\theta, \delta_\Pi)$ para todo θ.
Promediando ambos lados sobre Π, se obtiene la conclusión de que $r(\Pi, \delta_1) < r(\Pi, \delta_\Pi)$, lo cual contradice la suposición de que δ_Π es la regla de Bayes. Por lo tanto, δ_Π debe ser admisible. ■

Se puede extender la definición de reglas de Bayes a casos en los que Π es una medida, no necesariamente una medida de probabilidad. En mi opinión, el uso de distribuciones a priori impropias es más razonable en este caso (en comparación con el caso puramente bayesiano), ya que el objetivo aquí es simplemente construir reglas de decisión con buenas propiedades de riesgo.

Definición 5.4
Sea $\frac{dP_\theta}{d\mu}(x) = p_\theta(x)$. Sea Π una medida en Θ y supongamos que, para cada x, existe $\delta(x)$ tal que

$$\int_\Theta L(\theta, \delta(x)) p_\theta(x) \Pi(d\theta) = \min_{a \in \mathcal{A}} \int_\Theta L(\theta, a) p_\theta(x) \Pi(d\theta).$$

Entonces, $\delta = \delta_\Pi$ se denomina una *regla de Bayes generalizada* con respecto a Π.
La diferencia entre la Definición 5.4 y la Definición 5.3 es que la primera no requiere que Π sea una medida de probabilidad o finita. La situación cambia ligeramente en este caso, porque si la distribución a priori es impropia, la distribución a posteriori también podría serlo, lo que puede hacer que definir la regla de Bayes como un minimizador del riesgo posterior sea problemático.
Por ejemplo, si $X_1, \ldots, X_n$ son i.i.d. $N(\theta, 1)$ y Π es la medida de Lebesgue sobre $(-\infty, \infty)$, entonces $\delta_\Pi(x) = \bar{x}$ es la correspondiente regla de Bayes generalizada.
Sin embargo, el resultado de admisibilidad del Teorema 5.3 no se cumple, en general, para las reglas de Bayes generalizadas. La condición adicional es que la función de riesgo sea Π-integrable.

Teorema 5.5

Supongamos que Θ es como en el Teorema 3. Supongamos que $R(\theta, \delta)$ es continua en θ para toda δ. Sea Π una medida que domina la medida de Lebesgue, y sea δ_Π la correspondiente regla de Bayes generalizada. Si la función $(x, \theta) \mapsto L(\theta, \delta_\Pi(x)) p_\theta(x)$ es $\mu \times \Pi$-integrable, entonces δ_Π es admisible.
Demostración. Usar el teorema de Fubini y la Definición 5.4. ■

Volviendo al ejemplo normal, consideremos nuevamente la regla de Bayes generalizada $\delta_\Pi(x) = \bar{x}$ (con respecto a la prior Π igual a la medida de Lebesgue) para estimar θ bajo pérdida

cuadrática. La función de riesgo $R(\theta, \delta_\Pi)$ para $\delta_\Pi(x) = \bar{x}$ es constante (igual a $1/n$) y, por lo tanto, no es integrable con respecto a $\Pi =$ Lebesgue. Por lo tanto, el Teorema 5.5 no es suficiente para demostrar la admisibilidad del estimador de máxima verosimilitud.

Un enfoque alternativo es considerar una secuencia $\{\Pi_s : s \geqslant 1\}$ de priors propias o impropias, y la correspondiente secuencia de reglas de Bayes $\{\delta_{\Pi_s} : s \geqslant 1\}$. El siguiente teorema es una herramienta poderosa y general para demostrar admisibilidad. Un resultado similar se encuentra en Schervish (1995, p. 158–159).

Teorema 5.6

$$\lim_{s\to\infty} \{r(\Pi_s, \delta) - r(\Pi_s, \delta_s)\} = 0,$$

entonces δ es admisible.
Demostración. Ver Keener (2010, p. 215). ■

Para ilustrar el uso de este teorema, probaremos que, para $X \sim N(\theta, 1)$, el estimador de máxima verosimilitud $\hat{\theta} = x$ es admisible bajo pérdida cuadrática. El caso más general $n > 1$ se deduce de este haciendo un cambio de escala.

Ejemplo 5.7

Consideremos una secuencia de medidas $\Pi_s = \sqrt{s}N(0, s)$, $s \geqslant 1$; estas son finitas pero no son medidas de probabilidad. Sea $\delta(x) = x$ el estimador de máxima verosimilitud. Las reglas de Bayes generalizadas están dadas por $\delta_s(x) = sx/(s + 1)$ y los riesgos de Bayes son

$$r(\Pi_s, \delta) = \sqrt{s}, \quad r(\Pi_s, \delta_s) = s^{3/2}/(s + 1).$$

La diferencia $r(\Pi_s, \delta) - r(\Pi_s, \delta_s) = s^{1/2}/(s + 1)$ tiende a cero cuando $s \to \infty$. Lo único que queda por verificar es que $\Pi_s(B)$ está acotada lejos de cero para todos los intervalos abiertos $x \pm m$. Para esto, tenemos

$$\Pi_s(x \pm m) = \frac{1}{\sqrt{2\pi}} \int_{x-m}^{x+m} e^{-u^2/2s}\, du,$$

y dado que el integrando está acotado por 1 para todo s y converge a 1 cuando $s \to \infty$, el teorema de convergencia dominada implica que $\Pi_s(x \pm m) \to 2m(2\pi)^{-1/2} > 0$. Se sigue del Teorema 6 que $\delta(x) = x$ es admisible.

Ejemplo 5.8

Sea $X \sim \text{Bin}(n, \theta)$, de modo que el estimador de máxima verosimilitud de θ es la media muestral $\delta(X) = X/n$.
El objetivo es demostrar, usando el Teorema 6, que δ es admisible bajo pérdida cuadrática. Para esto, necesitamos una secuencia de distribuciones a priori propias $\{\Pi_s : s \geqslant 1\}$ para θ.
Dado que las distribuciones beta son conjugadas para el modelo binomial, un punto de partida razonable es considerar $\theta \sim \text{Beta}(s^{-1}, s^{-1})$. Para dicho modelo, la regla de Bayes es

$$\mathbb{E}(\theta \mid X) = \frac{X + s^{-1}}{n + 2s^{-1}}.$$

Es claro que, cuando $s \to \infty$, la regla de Bayes $\delta_{\Pi_s}(X)$ converge a la media muestral $\delta(X) = X/n$. Sin embargo, el límite de las distribuciones beta no es una distribución a priori propia para θ; de hecho, el límite de las distribuciones a priori tiene una densidad proporcional a $\{\theta(1-\theta)\}^{-1}$, lo cual es impropio.
Las distribuciones beta a priori en sí mismas no satisfacen las condiciones del Teorema 5.6 (ver Ejercicio 7). Afortunadamente, existe una modificación sencilla de la distribución beta a priori que sí funciona. Tomemos Π_s con densidad π_s, que es simplemente la Beta(s^{-1}, s^{-1}) sin la constante de normalización:

$$\pi_s(\theta) = \{\theta(1-\theta)\}^{s^{-1}-1}$$

o, en otras palabras,

$$\pi_s(\theta) = \lambda(s)\text{Beta}(\theta \mid s^{-1}, s^{-1}),$$

donde

$$\lambda(s) = \frac{\Gamma(s^{-1})^2}{\Gamma(2s^{-1})}.$$

Dado que Π_s es simplemente una reescalación de la distribución beta a priori anterior, es claro que la regla de Bayes $\delta_{\Pi_s}(X)$ para la nueva distribución a priori es la misma que para la distribución beta anterior. Luego, los cálculos del riesgo de Bayes son relativamente simples (ver Ejercicio 8).
Con la secuencia de distribuciones a priori Π_s, que son propias pero no medidas de probabilidad, se sigue del Teorema 5.6 que la media muestral $\delta(X) = X/n$ es un estimador admisible de θ.

También existen algunos teoremas generales sobre la admisibilidad de los .[es]timadores estándar.[en] familias exponenciales que abarcan el resultado demostrado en el Ejemplo 5.7. Las condiciones de tales teoremas son bastante técnicas, por lo que no las abordaremos aquí. Para una declaración detallada y una demostración de uno de estos teoremas, véase Schervish (1995), páginas 160–161.

Otro uso importante de las reglas de Bayes se verá en la siguiente sección, donde las reglas de Bayes con respecto a ciertos priors de "peor caso"producirán reglas minimax.

5.3.2. Minimización del riesgo máximo (MINIMAX)

En la sección anterior medimos el desempeño global de una regla de decisión promediando su función de riesgo con respecto a una medida de probabilidad Π en Θ. Sin embargo, esta no es la única forma de resumir el desempeño de una regla de decisión. Otro criterio es considerar el máximo de la función de riesgo $R(\theta, \delta)$ cuando θ varía en Θ. Este riesgo máximo representa el peor desempeño que puede tener la regla de decisión δ.

La idea, entonces, es elegir δ de modo que este desempeño en el peor caso sea lo más pequeño posible.

Definición 5.5

Para un problema de decisión con función de riesgo $R(\theta, \delta)$, una regla de decisión minimax δ_0 satisface:

$$\sup_{\theta} R(\theta, \delta_0) \leqslant \sup_{\theta} R(\theta, \delta)$$

para todas las reglas de decisión δ. La regla minimax protege contra el peor escenario en el sentido de que minimiza el riesgo máximo.

El origen de este enfoque se encuentra en el escenario de la teoría de juegos, donde un jugador compite contra un oponente. El objetivo del oponente es maximizar tu propia pérdida, por lo que una estrategia en este contexto sería elegir una estrategia que minimice la pérdida máxima, es decir, la estrategia minimax.

Sin embargo, en un problema de decisión estadística, esta estrategia podría considerarse demasiado conservadora o pesimista, ya que no tiene en cuenta la probabilidad del valor de θ en el que ocurre el máximo. No obstante, las reglas de decisión minimax tienen una larga historia.

Teorema 5.7

Sea Π una medida de probabilidad y δ_Π la correspondiente regla de Bayes. Si

$$r(\Pi, \delta_\Pi) = \sup_{\theta} R(\theta, \delta_\Pi),$$

entonces δ_Π es minimax.

Demostración. Sea δ otro procedimiento. Entonces,

$$\sup_{\theta} R(\theta, \delta) \geqslant r(\Pi, \delta) \geqslant r(\Pi, \delta_\Pi).$$

Dado que $r(\Pi, \delta_\Pi) = \sup_\theta R(\theta, \delta_\Pi)$ por suposición, se sigue que δ_Π es minimax. ■

Corolario 5.1

Una regla de Bayes δ_Π con riesgo constante es minimax.

Demostración. Si el riesgo es constante, entonces las condiciones del Teorema 5.7 se cumplen trivialmente. ■

Como se podría suponer, una distribución a priori Π para la cual el riesgo promedio es igual al riesgo máximo es un tipo particular de distribución extraña. Es decir, debe concentrar toda su masa en valores de θ donde el riesgo de la regla de Bayes δ_Π es grande. Estas son las distribuciones a priori de "peor caso"mencionadas al final de la sección anterior. Tal distribución a priori se denomina *prior menos favorable* y satisface:

$$r(\Pi, \delta_\Pi) \geqslant r(\Pi', \delta_{\Pi'})$$

para todas las distribuciones a priori Π'.

Para ver esto, sea Π una distribución a priori que satisface $r(\Pi, \delta_\Pi) = \sup_\theta R(\theta, \delta_\Pi)$. Para otra distribución a priori Π' tenemos:

$$r(\Pi', \delta_{\Pi'}) \leqslant r(\Pi', \delta_{\Pi}) \leqslant \sup_{\theta} R(\theta, \delta_{\Pi}) = r(\Pi, \delta_{\Pi}),$$

por lo que Π es la menos favorable.

En la práctica, las distribuciones a priori menos favorables no son particularmente útiles. Sin embargo, esta conexión entre distribuciones a priori menos favorables y estimadores minimax proporciona una técnica poderosa para encontrar estimadores minimax.

Ejemplo 5.9

Sea $X \sim \text{Bin}(n, \theta)$. El objetivo es encontrar un estimador minimax de θ bajo pérdida cuadrática. Consideremos una distribución a priori conjugada Beta(α, β). Entonces, la media a posteriori es

$$\delta(X) = \mathbb{E}(\theta \mid X) = aX + b = \frac{1}{\alpha + \beta + n} X + \frac{\alpha}{\alpha + \beta + n}.$$

La función de riesgo para δ es

$$R(\theta, \delta) = V_\theta\{aX + b - \theta\} + \mathbb{E}^2_\theta\{aX + b - \theta\} = A\theta^2 + B\theta + C,$$

donde A, B y C dependen de (α, β, n), y se invita al lector a encontrar las expresiones exactas en el Ejercicio 9.
La función de riesgo es constante si y solo si $A = B = 0$, lo que ocurre si $\alpha = \beta = \frac{1}{2}\sqrt{n}$. Por lo tanto, la regla de Bayes con riesgo constante es

$$\delta(x) = \frac{x + \frac{1}{2}\sqrt{n}}{n + \sqrt{n}},$$

y así, este estimador es minimax según el Corolario 1.

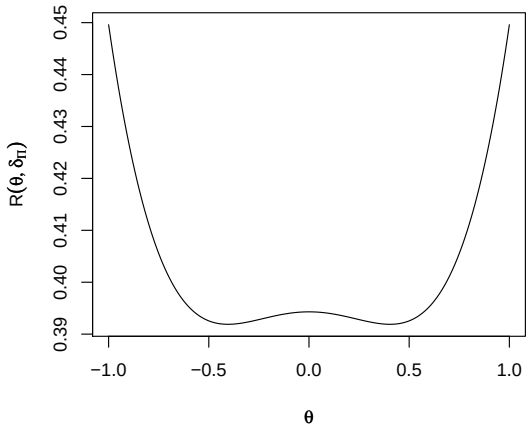

Figura 1: Grafica de la Funcion Riesgo del ejemplo 10.

Ejemplo 5.10

Sea $X \sim N(\theta, 1)$ donde se sabe que $|\theta| \leqslant 1$. Se puede demostrar (Ejercicio 12) que el estimador de máxima verosimilitud $\hat{\theta} = X$ es inadmisible en este caso. Aquí encontraremos un estimador minimax δ de θ bajo pérdida cuadrática.
Consideremos una medida de probabilidad Π que asigna probabilidad 0.5 a los extremos del intervalo $[-1, 1]$. Es decir, $\Pi(\{-1\}) = \Pi(\{1\}) = 0{,}5$. En este caso, la distribución a posteriori está determinada por

$$\Pi_x(\{1\}) = \frac{\varphi(x-1)/2}{\varphi(x-1)/2 + \varphi(x+1)/2} = \frac{\varphi(x-1)}{\varphi(x-1) + \varphi(x+1)},$$

donde φ es la función de densidad normal estándar. Entonces, la media a posteriori es

$$\delta_\Pi(x) = \frac{\varphi(x-1) - \varphi(x+1)}{\varphi(x-1) + \varphi(x+1)} = \frac{e^{-x} - e^{x}}{e^{-x} + e^{x}} = \tanh(x).$$

Se puede demostrar que la función de riesgo $R(\theta, \delta_\Pi)$ es simétrica y se maximiza en $\theta = \pm 1$ (ver Figura 5.1). En este caso, el riesgo máximo es igual al promedio de $R(\pm 1, \delta_\Pi)$, por lo que, según el Teorema 7, δ_Π es minimax.

En el Ejercicio 10(b) se invita al lector a demostrar, usando argumentos directos, que en un problema de media normal, la media muestral es un estimador minimax bajo pérdida cuadrática. El caso multivariante con una función de pérdida más general se considera en la Sección 5.6. Este argumento se basa en un resultado interesante del análisis convexo llamado el lema de Anderson.

Los procedimientos minimax son pesimistas por naturaleza y, en los problemas relativamente simples considerados hasta ahora, no es demasiado difícil encontrar mejores procedimientos, por ejemplo, los estimadores de máxima verosimilitud. Sin embargo, si nos alejamos de estos problemas relativamente simples", la máxima verosimilitud podría no ser adecuada y necesitaríamos criterios diferentes para construir estimadores, etc.

5.4. Minimización del riesgo bajo restricciones

Anteriormente vimos que encontrar un δ que minimice el riesgo $R(\theta, \delta)$ uniformemente sobre θ es imposible. En la Sección 5.3 vimos dos estrategias comunes para introducir un riesgo global y encontrar reglas de decisión óptimas.

Un enfoque alternativo es introducir una restricción razonable.[en] el conjunto de reglas de decisión que se está dispuesto a considerar. En este caso, puede ser posible encontrar un (restringido) δ para el cual $R(\theta, \delta)$ se minimice uniformemente sobre θ.

5.4.1. Restricciones de insesgadez

Estamos familiarizados con la insesgadez en el contexto de estimación. Sin embargo, la insesgadez es una condición general para las reglas de decisión. Es decir, para una función de pérdida $L(\theta, a)$, una regla de decisión δ es *insesgada* si

$$E_{\theta'}\{L(\theta', \delta(X))\} \geqslant E_{\theta}\{L(\theta, \delta(X))\}, \quad \forall \theta'. \tag{4}$$

En el Ejercicio 13, se te invita a demostrar que, si el objetivo es estimar $g(\theta)$ bajo la pérdida de error cuadrático, entonces la condición de insesgadez (5.4) es equivalente a la definición familiar, es decir,

$$E_{\theta}\{\delta(X)\} = g(\theta) \quad \text{para todo } \theta.$$

Aunque la insesgadez es un concepto más general (ver Sección 5.4.3), nos centraremos aquí en el problema de estimación.

Un resultado interesante es que no necesitamos considerar los estimadores de Bayes en este contexto, ya que (excepto en casos extraños) no pueden ser insesgados.

Teorema 5.8

Ningún estimador insesgado $\delta(X)$ puede ser un estimador de Bayes a menos que la distribución previa Π satisfaga $\Pi\{\theta : R(\theta, \delta) = 0\} = 1$.

Demostración. Supongamos que δ es una regla de Bayes (bajo la pérdida de error cuadrático con respecto a Π) y es insesgada. Entonces sabemos que

$$\delta(X) = E\{g(U) \mid X\} \quad \text{y} \quad g(U) = E\{\delta(X) \mid U\}.$$

Luego, dependiendo del orden en el que condicionemos, obtenemos

$$E[g(U)\delta(X)] = \begin{cases} E[g(U)E\{\delta(X) \mid U\}] = E[g(U)^2] & \text{condicionando en } U, \\ E[\delta(X)E\{g(U) \mid X\}] = E[\delta(X)^2] & \text{condicionando en } X. \end{cases}$$

Por lo tanto, $E[g(U)^2] = E[\delta(X)^2]$ y, en consecuencia,

$$r(\Pi, \delta) = E[\delta(X) - g(U)]^2 = E[\delta(X)^2] - 2E[g(U)\delta(X)] + E[g(U)^2] = 0.$$

Pero el riesgo de Bayes también satisface

$$r(\Pi, \delta) = \int R(\theta, \delta)\, d\Pi(\theta).$$

Dado que $R(\theta, \delta) \geqslant 0$ para todo θ, la única manera en que la integral con respecto a Π sea cero es si Π asigna probabilidad 1 al conjunto de valores de θ donde $R(\theta, \delta)$ se anula. Esto prueba la afirmación. ■

Por lo tanto, restringirse a estimadores insesgados necesariamente excluye a los estimadores de Bayes (razonables). Sin embargo, esto no ayuda a encontrar el mejor estimador, ni siquiera sugiere que exista un "mejor.^{es}timador.

Afortunadamente, existe un resultado muy poderoso: el **teorema de Lehmann-Scheffé**, que establece que, en efecto, hay una regla que minimiza uniformemente el riesgo y, además, proporciona condiciones suficientes fácilmente verificables para identificar este mejor estimador.

Teorema 5.9 Lehmann-Scheffé

Sea $X \sim P_\theta$ y supongamos que T es un estadístico suficiente completo. Supongamos que el objetivo es estimar $g(\theta)$ bajo una pérdida convexa, y que existe un estimador insesgado. Entonces, existe un estimador insesgado esencialmente único que es una función de T y minimiza uniformemente el riesgo.

Hay algunas cosas importantes que vale la pena mencionar sobre este teorema. Primero, nótese que la pérdida cuadrática no es fundamental; lo que realmente importa es que la función de pérdida sea convexa.

En segundo lugar, este teorema no garantiza que exista un estimador insesgado; hay ejemplos donde no existe un estimador insesgado (por ejemplo, al estimar $1/\theta$ en una $\text{Bin}(n, \theta)$). Si existe un estimador insesgado, entonces el **teorema de Rao-Blackwell** muestra cómo mejorarlo condicionando sobre T.

El hecho de que T sea también completo es lo que conduce a la unicidad. En efecto, si existen dos estimadores insesgados que son funciones de T, entonces su diferencia

$$f(T) = \delta_1(T) - \delta_2(T)$$

satisface $E_\theta\{f(T)\} = 0$ para todo θ. La completitud de T implica que $f = 0$ casi en todas partes, lo que a su vez implica que δ_1 y δ_2 son (c.a.) el mismo estimador.

5.4.2. Restricciones de Equivarianza

Para un espacio muestral $\mathcal{X}$, sea $\{P_\theta : \theta \in \Theta\}$ un modelo de transformación de grupo con respecto a un grupo $\mathcal{G}$ de transformaciones $g : \mathcal{X} \to \mathcal{X}$. Es decir, si $X \sim P_\theta$, entonces $gX \sim P_{\theta'}$ para algún θ' en Θ. Este θ' particular está determinado por θ y la transformación g. En otras palabras, las transformaciones g también actúan sobre Θ, pero posiblemente de una manera diferente a como actúan sobre $\mathcal{X}$. Nos referiremos a esta transformación en Θ determinada por g como g_Θ, y a la colección de todas estas transformaciones como $\mathcal{G}_\Theta$. Se puede demostrar que $\mathcal{G}_\Theta$ también es un grupo.

En resumen, tenemos los grupos $\mathcal{G}$ y $\mathcal{G}_\Theta$ actuando sobre $\mathcal{X}$ y Θ, respectivamente, los cuales están relacionados con la distribución P_θ de la siguiente manera:

$$X \sim P_\theta \iff gX \sim P_{g_\Theta \theta},$$

donde $g_\Theta \in \mathcal{G}_\Theta$ está determinado por $g \in \mathcal{G}$.

Esta es una estructura especial impuesta sobre la distribución P_θ. Estamos familiarizados con las estructuras de localización y escala, donde $\mathcal{G}$ y $\mathcal{G}_\Theta$ son el mismo grupo, pero existen otros casos; por ejemplo, ya has visto la distribución Weibull como un modelo de transformación de grupo, y se ha escrito explícitamente la función g_Θ para un g dado.

En esta sección investigaremos el efecto de tal estructura en el problema de decisión estadística. Primero, necesitamos imponer esta estructura en algunos de los otros elementos, como

el espacio de acción, la función de pérdida y las reglas de decisión. Lo haremos rápidamente aquí.

- La función de pérdida se llama *invariante* (con respecto a $\mathcal{G}$ o $\mathcal{G}_\Theta$) si, para cada $g \in \mathcal{G}$ (o cada $g \in \mathcal{G}_\Theta$) y cada $a \in \mathcal{A}$, existe un único $a' \in \mathcal{A}$ tal que $L(g\theta, a') = L(\theta, a)$ para todo θ. En este caso, el grupo también actúa (de manera directa o indirecta) sobre el espacio de acciones $\mathcal{A}$, es decir, a' está determinado por a y g. Escribimos $a' = ga$. Se puede demostrar que la colección $\mathcal{G}$ de transformaciones $g : \mathcal{A} \to \mathcal{A}$ también forma un grupo.
- Una función h definida en $\mathcal{X}$ (o en algún otro espacio como Θ o $\mathcal{A}$, equipado con un grupo de transformaciones) se llama *invariante* si $h(gx) = h(x)$ para todo $x \in \mathcal{X}$ y todo $g \in \mathcal{G}$. Alternativamente, una función $f : \mathcal{X} \to \mathcal{A}$ es *equivariante* si $f(gx) = gf(x)$.
- Nos enfocaremos en reglas de decisión δ que sean *equivariantes*. La intuición es que nuestras reglas de decisión deben ser consistentes con la estructura asumida. Por ejemplo, en un problema de parámetro de localización, un desplazamiento de los datos por una constante debería causar un desplazamiento de nuestro estimador de la localización en la misma cantidad.

Un problema de decisión cuyos elementos satisfacen todas estas propiedades se denomina, en general, un *problema de decisión invariante*.

Consideraremos la exigencia de que la regla de decisión sea equivariante como una restricción sobre las posibles reglas de decisión, al igual que la insesgadez es una restricción. Entonces, la pregunta es si existe una regla *equivariante* que minimice uniformemente el riesgo. El primer resultado es un paso en esta dirección.

Teorema 5.10

En un problema de decisión invariante, la función de riesgo $R(\theta, \delta)$ de una regla de decisión equivariante δ es una función invariante en Θ, es decir, es constante en las órbitas de $\mathcal{G}$.

Las órbitas mencionadas en el Teorema 5.10 son los conjuntos

$$O_\theta = \{\theta' \in \Theta : \theta' = g\theta,\ g \in \mathcal{G}\}.$$

Este conjunto O_θ consiste en todas las posibles imágenes de θ bajo transformaciones $g \in \mathcal{G}$. Entonces, una definición equivalente de una función invariante es aquella que es constante en las órbitas. Una función invariante se llama *maximal* si los valores constantes son diferentes en distintas órbitas. Los invariantes maximales son importantes, pero no los discutiremos más aquí.

Un caso especial interesante es cuando el grupo $\mathcal{G}$ tiene solo una órbita, en cuyo caso, la función de riesgo en el Teorema 10 es constante en todas partes. Los grupos que tienen una única órbita se denominan *transitivos*. Las transformaciones de localización unidimensional corresponden a grupos transitivos; lo mismo ocurre con las transformaciones de escala. En este caso, es fácil comparar las funciones de riesgo de las reglas de decisión equivariante.

La pregunta es si existe una regla equivariante que minimice el riesgo. Existe un resultado general en esta dirección, pero no daremos una formulación precisa aquí.

Teorema 5.11

Consideremos un problema de decisión invariante. Bajo algunas suposiciones, si la regla de Bayes formal con respecto a la medida de Haar invariante a derecha en $\mathcal{G}$ existe, entonces es la regla equivariante de mínimo riesgo.

El principal desafío para comprender este teorema es la definición de la medida de Haar. Esto puede estar más allá del alcance de nuestro análisis, pero podemos considerar un ejemplo simple pero importante: la estimación equivariante de un parámetro de localización.

Ejemplo 5.11

Consideremos un problema de parámetro de localización, donde la densidad de $X_1, \ldots, X_n$ bajo P_θ tiene la forma

$$p_\theta(x_1, \ldots, x_n) = p_0(x_1 - \theta, \ldots, x_n - \theta), \quad \theta \in \mathbb{R}.$$

Es decir, $X_i = \theta + Z_i$, donde $Z_1, \ldots, Z_n$ tienen distribución P_0. En este caso, todos los grupos $\mathcal{G}, \mathcal{G}_\Theta$ y $\mathcal{G}_\mathcal{A}$ son (isomorfos a) el grupo de los números reales bajo la adición. Para los números reales bajo la adición, la medida invariante (izquierda y derecha) es la medida de Lebesgue λ (¿por qué?).
Una función de pérdida invariante es de la forma $L(\theta, a) = L(a-\theta)$. Entonces, el teorema establece que el estimador equivariante de riesgo mínimo δ_λ es la regla de Bayes formal basada en una medida de Lebesgue formal como prior. Es decir, $\delta_\lambda(x)$ es el $\delta(x)$ que minimiza

$$\frac{\int_\Theta L(\delta(x) - \theta) p_0(x - \theta)\, d\theta}{\int_\Theta p_0(x - \theta)\, d\theta}.$$

En el caso en que $L(a - \theta) = (a - \theta)^2$, es decir, pérdida cuadrática, sabemos que δ_λ es simplemente la media a posteriori bajo la medida de Lebesgue formal λ, es decir,

$$\delta_\lambda(x) = \frac{\int_\Theta \theta p_0(x - \theta)\, d\theta}{\int_\Theta p_0(x - \theta)\, d\theta}.$$

Este estimador—conocido como el *estimador de Pitman*, $\hat{\theta}_{\text{pit}}$—es el estimador equivariante de riesgo mínimo.
Nótese que en el caso en que $P_0 = N(0, 1)$, el estimador de Pitman es simplemente $\hat{\theta}_{\text{pit}}(x) = \bar{x}$.

5.4.3. Restricciones sobre el error de Tipo I

En un problema de prueba de hipótesis con pérdida 0-1, se puede demostrar (Ejercicio 3) que la función de riesgo es la suma de las probabilidades de error de Tipo I y Tipo II.

A partir de nuestro conocimiento previo sobre pruebas de hipótesis, sabemos que si hacemos que la probabilidad de error de Tipo I sea pequeña, entonces la probabilidad de error de Tipo II aumenta, y viceversa. Por lo tanto, no es evidente cómo minimizar estrictamente este riesgo.

La estrategia habitual es fijar la probabilidad de error de Tipo I en algún $\alpha \in (0,1)$ y tratar de encontrar una prueba que satisfaga esta restricción y que minimice la probabilidad de error de Tipo II (o maximice el poder). Esta es la idea detrás de las *pruebas más potentes*.

Aquí nos enfocaremos en la situación más simple. Supongamos que X es una realización de uno de dos modelos P_0 y P_1, ambos con densidades p_0 y p_1 en $\mathcal{X}$ con respecto a una medida μ. Entonces, el objetivo es probar la hipótesis

$$H_0 : X \sim P_0 \quad \text{versus} \quad H_1 : X \sim P_1.$$

Este es el llamado *problema de prueba simple contra simple*. En este caso, una regla de decisión es una función δ que asigna $\mathcal{X}$ a $[0,1]$. En un problema no aleatorizado, δ asigna $\mathcal{X}$ a $\{0,1\}$.

El siguiente teorema es un resultado importante en esta línea de estudio.

Teorema 5.12 Neyman-Pearson

Para un $\alpha \in (0,1)$ fijo, la prueba más potente de nivel α está dada por

$$\delta(x) = \begin{cases} 0, & \text{si } p_1(x) < k_\alpha p_0(x), \\ \gamma, & \text{si } p_1(x) = k_\alpha p_0(x), \\ 1, & \text{si } p_1(x) > k_\alpha p_0(x), \end{cases} \tag{5}$$

donde γ y $k(\alpha)$ están determinados de manera única por la restricción

$$\alpha = P_0\left\{\frac{p_1(X)}{p_0(X)} > k_\alpha\right\} + \gamma P_1\left\{\frac{p_1(X)}{p_0(X)} = k_\alpha\right\}.$$

Nótese que la parte γ del teorema permite reglas de decisión aleatorizadas. Es decir, si la observación particular $X = x$ satisface $p_1(x) = k_\alpha p_0(x)$, entonces la regla establece que se debe lanzar una moneda con probabilidad de éxito γ y rechazar H_0 si la moneda cae en cara. Este mecanismo de aleatorización típicamente no es necesario en problemas con datos continuos, ya que el evento de que la razón de verosimilitud sea exactamente igual a k_α tiene probabilidad 0. Sin embargo, no podemos descartar pruebas aleatorizadas desde el inicio, porque la pérdida 0-1 no es convexa.

Aquí hay una interpretación alternativa del conocido *lema de Neyman-Pearson*. Podemos indexar los tests δ en (5) por el valor particular de α; escribimos estos tests como δ_α. Ahora, consideremos cualquier otro test δ' para este problema en particular. Supongamos que tiene alguna probabilidad de error de Tipo I α'. Entonces, el teorema muestra que δ_α domina a δ' en términos de riesgo. Por lo tanto, δ' es inadmisible.

En la discusión anterior, nos centramos en el caso de pruebas de hipótesis simples contra simples. A continuación, algunas observaciones sobre problemas más generales:

- Si la alternativa es unilateral (por ejemplo, $H_1 : \theta > \theta_0$), a menudo el test simple contra simple derivado del lema de Neyman-Pearson sigue siendo el mejor. El punto clave es que la prueba de Neyman-Pearson en realidad no depende del valor de θ_1 en la alternativa simple.
- Cuando la alternativa es bilateral, es un hecho bien conocido que generalmente no existe una prueba uniformemente más potente. Para abordar esto, se puede considerar la

restricción a pruebas insesgadas que satisfacen (4). En particular, en muchos casos, existe una prueba uniformemente más potente dentro de la clase de pruebas insesgadas; véase Lehmann y Romano (2005) para un tratamiento detallado de las pruebas uniformemente más potentes y la condición de insesgadez.

5.5. Teoremas de clases completas

Una clase de reglas de decisión se llama *clase completa*, denotada por $\mathcal{C}$, si para cualquier $\delta_1 \notin \mathcal{C}$, existe una regla $\delta_0 \in \mathcal{C}$ tal que

$$R(\theta, \delta_0) \leqslant R(\theta, \delta_1) \quad \text{para todo } \theta,$$

con desigualdad estricta para algún θ. En otras palabras, ninguna δ fuera de $\mathcal{C}$ es admisible.

Aquí hay algunos hechos interesantes:

- Si la función de pérdida es convexa, entonces el conjunto de todas las reglas de decisión que son funciones de un estadístico suficiente forma una clase completa.
- Si la función de pérdida es convexa, entonces el conjunto de todas las reglas de decisión no aleatorizadas forma una clase completa.
- El conjunto de pruebas de la forma (5.5) (indexado por α) forma una clase completa.

Aunque una clase completa $\mathcal{C}$ contiene todas las reglas de decisión admisibles, puede haber muchas reglas en $\mathcal{C}$ que sean inadmisibles. Por lo tanto, sería interesante identificar la clase completa más pequeña. Una clase completa $\mathcal{C}$ se llama *mínima* si no existe un subconjunto propio de $\mathcal{C}$ que sea completa. Se puede demostrar (ver Ejercicio 20) que una clase completa mínima es exactamente el conjunto de reglas de decisión admisibles.

El resultado en el que nos enfocaremos aquí es aquel que establece que los (límites de) reglas de Bayes forman una clase completa o, en otras palabras, para cualquier regla de decisión δ, existe una regla "aproximadamente de Bayes" δ^* tal que el riesgo de δ^* no es mayor en todas partes que el riesgo de δ. Se puede dar a este resultado una interpretación topológica: en términos generales, las reglas de Bayes con priors adecuados forman un subconjunto denso de todas las reglas admisibles.

Teorema 5.13

Los estimadores que satisfacen las condiciones del Teorema 6 forman una clase completa.

Como caso especial de este teorema, si el modelo es parte de una familia exponencial y si δ es un límite de reglas de Bayes, entonces existe una subsecuencia $\{\Pi_j\}$ tal que $\Pi_j \to \Pi$ y δ es la regla de Bayes δ_Π correspondiente a este límite.

Es decir, la clase de todas las reglas de Bayes generalizadas forma una clase completa en el caso de la familia exponencial.

5.6. Sobre la estimación minimax de una media normal

Aquí nos interesa la estimación minimax de un vector de media normal θ, bajo una función de pérdida más general que el error cuadrático, basada en una muestra normal $X \sim N_d(\theta, \Sigma)$, donde la matriz de covarianza Σ es conocida. El tipo de función de pérdida que consideraremos es de la forma $L(\theta, a) = W(a - \theta)$, donde W es una función con forma de "cuenco" (*bowl-shaped*).

Definición 5.6
Una función $W : \mathbb{R}^d \to [0, \infty]$ es *bowl-shaped* si el conjunto $\{x : W(x) \leqslant \alpha\}$ es convexo y simétrico con respecto al origen para todo $\alpha \geqslant 0$.

En el caso $d = 1$, la función $W(x) = x^2$ tiene forma de cuenco; por lo tanto, los resultados que se desarrollarán a continuación se especializarán en el caso de estimación de una media normal escalar bajo la pérdida cuadrática estándar.

El análogo en d-dimensiones de la pérdida cuadrática es

$$L(\theta, a) = \|a - \theta\|^2,$$

donde $\|\cdot\|$ es la norma euclidiana usual en $\mathbb{R}^d$. En el Ejercicio 21 se invita a demostrar que la función correspondiente $W(x) = \|x\|^2$ tiene forma de cuenco.

Un resultado importante relacionado con funciones de este tipo es el siguiente, conocido como *lema de Anderson*.

Lema 5.1
Sea f una densidad de Lebesgue en $\mathbb{R}^d$, con $\{x : f(x) \geqslant \alpha\}$ convexo y simétrico respecto al origen para todo $\alpha \geqslant 0$. Si W es una función con forma de cuenco, entonces

$$\int W(x - c) f(x)\, dx \geqslant \int W(x) f(x)\, dx \quad \forall c \in \mathbb{R}^d.$$

El punto clave es que la función $\int W(x - c) f(x)\, dx$ se minimiza en $c = 0$. Este hecho será útil en nuestra derivación de un estimador minimax para θ más adelante. Antes de esto, me gustaría mencionar una aplicación del lema de Anderson.

Ejemplo 5.12
Sea $X \sim N_d(0, \Sigma)$ y sea A un conjunto convexo simétrico respecto al origen. Entonces, la densidad f de X y la función $W(x) = 1 - I_A(x)$ satisfacen las condiciones del Lema 5.1 (ver Ejercicio 22). Un ejemplo de un conjunto A es una bola centrada en el origen. Se sigue entonces que

$$P(X \in A) \geqslant P(X + c \in A) \quad \forall c \in \mathbb{R}^d. \tag{6}$$

En otras palabras, la distribución normal con media cero asigna la mayor probabilidad al conjunto convexo y simétrico A. Esto puede parecer intuitivamente obvio, pero la demostración no es sencilla. Resultados como este han sido utilizados recientemente en aplicaciones de

métodos Bayesianos en problemas de medias normales de alta dimensión.

A continuación, se presenta el resultado principal de esta sección, es decir, que $\delta(X) = X$ es minimax para la estimación de θ bajo cualquier función de pérdida $L(\theta, a) = W(a - \theta)$ con W de forma de cuenco.

Teorema 5.14

Sea $X \sim N_d(\theta, \Sigma)$ donde Σ es conocida. Entonces, X es un estimador minimax de θ bajo la función de pérdida $L(\theta, a) = W(a - \theta)$ para W con forma de cuenco.

Demostración. Consideremos un enfoque Bayesiano y tomemos un prior $\Theta \sim \Pi_\psi \equiv N_d(0, \psi\Sigma)$ para una escala genérica $\psi > 0$. Entonces, la distribución posterior de Θ, dado X, es

$$\Theta \mid X \sim N_d\left(\frac{\psi}{\psi+1}X, \frac{\psi}{\psi+1}\Sigma\right).$$

Denotemos por $f(z)$ la densidad $N_d(0, \{\psi/(\psi+1)\}\Sigma)$. Para cualquier estimador $\delta(X)$, el riesgo posterior es

$$E\{W(\Theta - \delta(X)) \mid X = x\} = \int W\left(z + \frac{\psi}{\psi+1}x - \delta(x)\right) f(z)\, dz.$$

Dado que W tiene forma de cuenco y f satisface los requisitos de convexidad, el lema de Anderson establece que el riesgo posterior se minimiza en

$$\delta_\psi(x) = \frac{\psi}{\psi+1}x;$$

por lo tanto, esta $\delta(x)$ es la regla de Bayes.
Bajo el modelo Bayesiano, la distribución de X es la misma que la de $\Theta + Z$, donde $Z \sim N_d(0, \Sigma)$ y Z es independiente de Θ. Entonces,

$$\Theta - \delta_\psi(X) = \Theta - \delta_\psi(\Theta + Z) = \frac{\Theta - \psi Z}{\psi+1} \quad \text{(en distribución)},$$

y la distribución del lado derecho es la misma que la de $\left\{\frac{\psi}{\psi+1}\right\}^{1/2} Z$.
Entonces, el riesgo de Bayes correspondiente es

$$r(\Pi_\psi, \delta_\psi) = EW(\Theta - \delta_\psi(X)) = EW\left(\left\{\frac{\psi}{\psi+1}\right\}^{1/2} Z\right).$$

Para cualquier estimador δ, tenemos que

$$\sup_\theta R(\theta, \delta) \geqslant r(\Pi_\psi, \delta) \geqslant r(\Pi_\psi, \delta_\psi).$$

Esto es válido para todo ψ, por lo que también se mantiene en el límite cuando $\psi \to \infty$, lo que implica

$$\sup_\theta R(\theta, \delta) \geqslant \lim_{\psi\to\infty} EW\left(\left\{\frac{\psi}{\psi+1}\right\}^{1/2} Z\right).$$

Dado que $\psi/(\psi+1) \to 1$ cuando $\psi \to \infty$, se sigue del *teorema de convergencia monótona* que la cota inferior anterior es $EW(Z)$, que es exactamente $\sup_\theta R(\theta, \hat{\theta})$, donde $\hat{\theta} = X$.

Dado que

$$\sup_{\theta} R(\theta, \delta) \geqslant \sup_{\theta} R(\theta, \hat{\theta})$$

para todo δ, se concluye que $\hat{\theta} = X$ es minimax. ■

5.7. Ejercicios

1. Suponga que $X_1, \ldots, X_n$ son variables aleatorias independientes Ber(θ). El objetivo es estimar θ bajo la pérdida de error cuadrático.

   (a) Calcule el riesgo para el estimador de verosimilitud máxima $\hat{\theta}_{\text{mle}} = \bar{X}$.

   (b) Encuentre la media a posteriori $\hat{\theta}_{\text{Bayes}} = E(\theta \mid X)$ bajo una prior Unif$(0, 1)$ para θ y calcule su función de riesgo. [*Sugerencia*: Ya ha encontrado la fórmula para la media a posteriori en la Tarea 04—solo use el hecho de que Unif$(0, 1)$ es un caso especial de Beta(a, b).]

   (c) Compare las dos funciones de riesgo.

2. Suponga que $X_1, \ldots, X_n$ son variables aleatorias independientes $N(\theta, 1)$.

   (a) Encuentre la función de riesgo del MLE $\bar{X}$ (bajo pérdida de error cuadrático).

   (b) Encuentre la función de riesgo para la media a posteriori Bayesiana bajo una prior $N(0, 1)$.

   (c) Compare las dos funciones de riesgo, por ejemplo, ¿dónde se intersectan?

3. Sea $X \sim P_\theta$ y considere la prueba de hipótesis $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$. Encuentre la función de riesgo para una prueba no aleatorizada δ basada en la pérdida 0-1. [*Sugerencia*: Esto involucrará las probabilidades de error Tipo I y Tipo II.]

4. Sea X una variable aleatoria con media θ y varianza σ^2. Para estimar θ bajo la pérdida de error cuadrático, considere la clase $\delta_{a,b}(x) = ax + b$. Muestre que si

   $$a > 1 \quad \text{o} \quad a < 0 \quad \text{o} \quad a = 1 \text{ y } b \neq 0,$$

   entonces $\delta_{a,b}$ es inadmisible.

5. Suponga que $X_1, \ldots, X_n$ son iid $N(\theta, 1)$. Si T es la media muestral, muestre que la distribución condicional de X_1 dado $T = t$ es $N\left(t, \frac{n-1}{n}\right)$. [*Sugerencia*: Puede hacer esto usando el teorema de Bayes o utilizando propiedades de la distribución normal multivariada.]

6. Reconsidere el problema del Ejercicio 2 y asuma una prior $N(0, 1)$, denotada por Π.

   (a) Encuentre el riesgo Bayesiano del MLE $\bar{X}$.

   (b) Encuentre el riesgo Bayesiano de la regla de Bayes.

   (c) ¿Cuál estimador tiene menor riesgo Bayesiano?

7. Considere que $\theta \sim \Pi_a = \text{Beta}(a, a)$, y sea $D \subset [0, 1]$ un intervalo abierto que no contiene 0 ni 1. Muestre que $\Pi_a(D) \to 0$ cuando $a \to 0$. [*Sugerencia*: Use el hecho de que $\Gamma(x) = \Gamma(x + 1)/x$.]

8. Considere la admisibilidad de la media muestral como se discute en el Ejemplo 8.

   (a) Muestre que el riesgo Bayesiano de la media muestral $\delta(X) = X/n$ con respecto a Π_s es:
   $$r(\Pi_s, \delta) = \frac{1}{4n}\left(3 - \frac{1}{2s+1}\right).$$

   (b) Muestre que el riesgo Bayesiano de la media a posteriori $\delta_{\Pi_s}(X) = E(\theta \mid X)$ con respecto a la prior Π_s es:
   $$r(\Pi_s, \delta_{\Pi_s}) = \left(\frac{1}{2s+1}\right)^2 \frac{3n}{4} - \frac{n}{4}\left[\frac{s^{-2}}{2s+1} + 1\right].$$

   (c) Muestre que $\{r(\Pi_s, \delta) - r(\Pi_s, \delta_{\Pi_s})\} \to 0$ cuando $s \to \infty$. [*Sugerencia*: Probablemente necesitará la propiedad de la función gamma del Ejercicio 7.]

9. Considere el problema binomial en el Ejemplo 5.9.

   (a) Encuentre expresiones para A, B y C en términos de α, β y n.

   (b) Muestre que $A = B = 0$ si y solo si $\alpha = \beta = \frac{1}{2}\sqrt{n}$.

   (c) Grafique la función de riesgo de la regla minimax y la del estimador de máxima verosimilitud $\delta(x) = x/n$ para $n \in \{10, 25, 50, 100\}$. Compare el desempeño de los dos estimadores en cada caso.

10. (a) Muestre que si una regla de decisión es admisible y tiene riesgo constante, entonces es minimax.

    (b) Use la parte (a) y el Ejemplo 5.7 para argumentar que, si $X_1, \ldots, X_n$ son iid $N(\theta, 1)$, entonces la media muestral $\bar{X}$ es un estimador minimax de θ bajo pérdida de error cuadrático.

    (c) Suponga que $X \sim \text{Bin}(n, \theta)$. Muestre que $\delta(x) = x/n$ es minimax para estimar θ bajo la función de pérdida
    $$L(\theta, a) = \frac{(a - \theta)^2}{\theta(1 - \theta)}.$$
    [*Sugerencia*: Encuentre una prior adecuada Π para que $\delta(x)$ sea una regla de Bayes, y por lo tanto admisible. Para demostrar que es minimax, use la parte (a).]

11. Los estimadores minimax no son únicos. De hecho, muestre que si $X \sim \text{Pois}(\theta)$, entonces todo estimador de θ es minimax bajo la pérdida de error cuadrático. [*Sugerencia*: Para mostrar que todo estimador δ tiene una función de riesgo no acotada $R(\theta, \delta)$, demuestre que existen priors Π y reglas de Bayes correspondientes δ_Π con riesgo de Bayes $r(\Pi, \delta_\Pi)$ arbitrariamente pequeño.]

12. En el Ejemplo 10, muestre que el estimador de máxima verosimilitud $\delta(x) = x$ es inadmisible. [*Sugerencia*: Encuentre otro estimador $\delta'(x)$ con riesgo en todas partes no mayor que el de $\delta(x) = x$; la clave es incorporar la restricción—piense en truncar $\delta(x)$.]

13. Considere el problema de estimación con la función de pérdida $L(\theta, a) = (a - \theta)^2$, es decir, la pérdida de error cuadrático. Muestre que, en este caso, la condición de insesgadez (5.4) en un estimador $\delta(X)$ de θ se reduce a la definición familiar, es decir,

$$E_\theta[\delta(X)] = \theta \quad \text{para todo } \theta.$$

14. Problema 4.6 en Keener (2010, p. 78). [*Sugerencia*: $\delta + cU$ es un estimador insesgado para todo c.]

15. Sean $X_1, \ldots, X_n$ iid Pois(θ). Encuentre el estimador UMVU de $P_\theta(X_1 \text{ es par})$.

16. Demuestre que el estimador de Pitman $\hat{\theta}_{\text{pit}}$ es equivariarte bajo traslación.

17. Para cada problema de localización a continuación, encuentre el estimador de Pitman de θ.

    (a) $X_1, \ldots, X_n$ iid Unif($\theta - 1, \theta + 1$).

    (b) $X_1, \ldots, X_n$ iid con densidad $\frac{1}{2}e^{-|x-\theta|}$ para $x \in \mathbb{R}$. No hay una expresión cerrada para $\hat{\theta}_{\text{pit}}$, pero se puede encontrar numéricamente. Escriba un programa de computadora para hacerlo y aplíquelo a los datos (6.59, 4.56, 4.88, 6.73, 5.67, 4.26, 5.80).

18. Problemas 10.2(a) y 10.3 en Keener (2010, p. 201).

19. Problema 10.8 en Keener (2010, p. 202).

20. Muestre que si $\mathcal{C}$ es una clase completa mínima, entonces es exactamente la clase de todas las reglas de decisión admisibles.

21. Defina $W(x) = \|x\|^2$ para $x \in \mathbb{R}^n$. Muestre que W tiene forma de cuenco.

22. Verifique las afirmaciones en el Ejemplo 5.12 que conducen a (6).