Teoría de la decisión estadística

Julia Teresa López

Trabajo de fin de grado en Matemáticas
Universidad de Zaragoza

Director del trabajo: F. Javier López Lorente
7 de Febrero de 2020

Prólogo

El presente trabajo de fin de grado trata sobre la teoría de la decisión estadística. Esta teoría es necesaria para resolver problemas de decisión reales. Se basa en analizar cómo elegir la acción que, de entre un conjunto de acciones posibles, le conduce al mejor resultado. Qué carrera voy a estudiar, qué coche me compraré o, incluso, dónde me iré de vacaciones, son problemas muy comunes que nos afectan en nuestra vida real y a los que se enfrenta la teoría de la decisión.

Este trabajo recopila diferentes criterios para la resolución de problemas dentro de la teoría de la decisión y contiene una buena cantidad de material de cada criterio. Los métodos explicados a lo largo del trabajo se desarrollan de forma sencilla, de manera que sea fácil de comprender para los lectores. Únicamente se requiere cierto conocimiento de la probabilidad. Los ejemplos que aparecen están desarrollados de manera que puedan ser útiles sin necesidad de leer el capítulo completo. En términos de enseñanza, el trabajo se puede usar como un curso de teoría de decisión estadística.

El Capítulo 2 trata sobre la estimación de Bayes y su relación con la teoría de la decisión estadística, es tan fuerte esta relación que es poco habitual hablar de la teoría de la decisión sin nombrar a la estimación Bayesiana. También, existe una teoría de la decisión, que evita la utilización formal de distribuciones a priori, esta está desarrollada en el Capítulo 3 llamado minimax.

Estoy muy agradecida con varias personas que contribuyeron, de una forma u otra, al trabajo. Especialmente con F. Javier López Lorente quien me proporcionó el material del trabajo y fue mi director.

Abstract

The goal of this project is to introduce basic concepts and results of statistical decision theory.

In the first chapter there is an introduction about decision approach, in which it is understood that any problem consider can be seen as a problem where a decision has to be made. The general objective in a problem in this context is the study of decisions that can be taken in a situation of uncertainty defined by the possible "states of nature" and about which we have certain information. The decisions that are taken will have associated costs (which will depend on the state of nature) and that decision that minimized these costs will be sought.

In the problems of statistical decision, to facilitate decision making, we will have observations of a random quantity that will come from a family of distributions indexed by the states of nature, as we will deal with parametric problems, the states will be identified with the set of possible values of the parameter $\theta \in \Theta$. When the data is observed a decision must be taken that will involve θ; this decision will be given by an action a within a set A of possible actions and, once this action is taken, the error made can be quantified, through a loss function.

A decision rule is a function that specifies, for each observed sample $\mathbf{x}$, the action to be taken, that is, a function of the sample space in the set of actions. To analyze a decision rule $d \in D$ (the set of rules considered), it must be done on all possible values of the sample and, therefore, the risk of the decision rule is defined in a parameter value θ as the expectation of the loss function (with respect to the population distribution for that parameter value).

Following, problems of punctual estimation and hypothesis testing will be seen within the framework of the decision theory. The risk for specific loss function in specific problems will be calculated. Through the examples, comparing different rules according to their risk, it will be verified that the fact that it depends on the parameter will not generally allow finding a rule that is uniformly superior to the others.

Chapter 2 will be dedicated to Bayesian decision theory. In this case, the criterion to minimize the risk is obtained through the allocation of weights π to the different values of the parameter θ, so that the rule that minimizes the weighted risk is sought. For the hypothesis testing, the loss function must be such that it includes the importance of each error and the test that minimizes the risk must be found. The chapter will conclude with examples.

In chapter 3 we study another way to find optimal rules that have their origins in game theory, considering the decision problem as a game against a generic opponent (nature). The objective in this case is to find the rule that minimized the maximum risk. From the perspective of game theory, we try to minimize the risk in the worst possible situation, that is, it is thought that nature is an opponent who will try to put things in the worst possible way for our interests. From the intuitive point of view, it seems that the minimax rules are going to be too conservative, because they consider the worst case.

Finally, chapter 4 is dedicated to two important concepts within the theory of decision, admissibility and completeness. It is important to note that admissibility cannot be taken as a guarantee of optimum, but rather as the absence of a negative property. It seems logical to study whether the search for a rule can be restricted to some classes that basically contain all the admissible rules; for this, the concept of completeness is introduced.

Throughout each chapter, the problems os parametric estimation and hypothesis testing will be tackled. You will find definitions, theorems and exercises.

In conclusion, we will see the relations between the Bayes rules, minimax and admissibility through a scheme that relates the different chapters of the project.

Índice general

Prólogo III

Abstract V

1. Introducción 1

2. Teoría de la decisión Bayesiana 7
2.1. Introducción . . . 7
2.2. Enfoque Bayesiano de la estadística . . . 7
2.3. Regla Bayes . . . 9
2.4. Estimación puntual . . . 10
2.5. Test de Hipótesis . . . 11

3. Minimax 15
3.1. Estimación puntual. Relación con las reglas de Bayes . . . 15
3.2. Test de Hipótesis . . . 18

4. Admisibilidad y Completitud 21
4.1. Estimación puntual. Relación con las reglas de Bayes y Minimax . . . 22
4.2. Test de Hipótesis . . . 25

Bibliografía 27

Capítulo 1

Introducción

En este capítulo se introducen conceptos y resultados básicos de la teoría de la decisión estadística. Desde el enfoque decisional que tiene esta teoría, se entiende que cualquier problema de inferencia estadística paramétrica puede verse como un problema donde hay que tomar una decisión. El objetivo general de un problema en este contexto es el estudio de las decisiones que se pueden tomar en una situación de incertidumbre, definida por los posibles estados de la naturaleza, θ, y sobre la que poseemos cierta información. Las decisiones que se tomen llevarán asociados unos costes (que dependerán del estado de la naturaleza) y se buscará aquella decisión que minimice estos costes en algún sentido. Para ayudarnos en la toma de decisiones dispondremos de observaciones $\mathbf{x}$ de la cantidad aleatoria bajo estudio. Tales observaciones provendrán de una familia de distribuciones indexadas por los estados de la naturaleza $\{F_\theta,\ \theta \in \Theta\}$. Cuando se observen los datos se deberá tomar una decisión que involucrará a θ. Esta decisión vendrá dada por una acción $a \in A$, siendo A el conjunto de posibles acciones. Una vez seleccionada esta acción, se dispondrá de una forma de cuantificar el error cometido, a través de una función de pérdida. Comenzamos este capítulo introduciendo notación que vamos a utilizar a lo largo del trabajo:

- Denotaremos por Θ al espacio paramétrico, siendo Θ un subconjunto abierto de $\mathbb{R}^k$ que representa los posibles estados de la naturaleza.

- El conjunto de resultados posibles es el espacio muestral, y se denotará como $\mathscr{X} \in \mathbb{R}^n$. Trabajaremos con muestras aleatorias simples $(X_1,\ X_2,\ ...,\ X_n)$, es decir, $X_1,\ X_2,\ ...,\ X_n$ serán variables aleatoias independientes identicamente distribuidas con distribución F_θ, $\theta \in \Theta$. Denotaremos por $\mathbf{x} = (x_1,\ x_2,\ ...,\ x_n)$ el valor de la muestra, siendo $x_1,\ x_2,\ ...,\ x_n$ las observaciones.

- Sea X una variable aleatoria discreta o absolutamente continua; denotaremos a la distribución de masa de probabilidad o de densidad de X como $f(x|\theta)$. Denotaremos por $f(\mathbf{x}|\theta)$ a la verosimilitud de una muestra $\mathbf{x}$.

- Utilizaremos el subíndice θ para denotar que las probabilidades se calculan con la distribución F_θ. Así, por ejemplo, E_θ denota la esperanza cuando la distribución de la variable se toma con el valor del parámetro igual a θ.

Definición 1.1. Se denomina *función de pérdida* a la función no negativa $L : \Theta \times A \to \mathbb{R}$, donde A es un espacio de acciones arbitrario. $L(\theta,\ a)$ será la pérdida obtenida si se toma la acción a cuando θ es el estado de la naturaleza.

Funciones de pérdida habituales cuando $\Theta,\ A \subseteq \mathbb{R}$:

- Función de pérdida absoluta: $L(\theta,\ a) = |a - \theta|$.

- Función de pérdida cuadrática: $L(\theta,\ a) = (a - \theta)^2$.

Ambas funciones de pérdida aumentan a medida que aumenta la distancia entre θ y a, con un valor mínimo de $L(\theta,\ \theta) = 0$. Es decir, la pérdida es mínima si la acción es correcta. La función de pérdida cuadrática proporciona una penalización mayor que la función de pérdida absoluta para grandes discrepancias entre θ y a, sin embargo, la función de pérdida absoluta genera una penalización mayor que la función de pérdida cuadrática para pequeñas discrepancias entre θ y a, en concreto cuando $|a-\theta| < 1$.

Una variación del error de pérdida cuadrática que penaliza más la sobreestimación de θ que la subestimación de θ es:

$$L(\theta,\ a) = \begin{cases} (a-\theta)^2 & si \quad a < \theta \\ \\ 10\ (a-\theta)^2 & si \quad a \geq \theta \end{cases}$$

Otra variación del error de pérdida cuadrática es:

$$L(\theta,\ a) = \frac{(a-\theta)^2}{|\theta|+1}$$

En este caso penalizan más los errores en la estimación si θ está cerca del 0 que si $|\theta|$ es grande.

Definición 1.2. Se denomina *regla o función de decisión* a la aplicación $d : \mathscr{X} \to A$, que especifica para cada muestra observada $\mathbf{x}$ la acción que hay que tomar.

Para analizar una regla de decisión $d \in D$, siendo D el conjunto de reglas de decisión que se van a considerar, hay que analizar d sobre todos los posibles valores de la muestra $\mathbf{x}$, y para ello utilizamos la función de riesgo.

Definición 1.3. Denominamos *función de riesgo* a la aplicación $R : \Theta \times D \to \mathbb{R}$ dada por $R(\theta,\ d) = E_\theta\ L(\theta,\ d(\mathbf{X}))$, si existe. Esta función representa la pérdida media dada cuando el estado de la naturaleza es θ y se toma la regla de decisión d.

El problema que queremos resolver mediante la teoría de la decisión estadística es encontrar la regla que minimice el riesgo. Como éste depende de θ, que es desconocido, queremos que el riesgo de la regla sea pequeño para todo valor de $\theta \in \Theta$. Así, a la hora de comparar dos reglas d_1 y d_2, lo haremos mediante sus funciones de riesgo, $R(\theta,\ d_1)$ y $R(\theta,\ d_2)$. Si $R(\theta,\ d_1) \leq R(\theta,\ d_2)$ para todo $\theta \in \Theta$ y $R(\theta,\ d_1) < R(\theta,\ d_2)$ para al menos un $\theta \in \Theta$, entonces d_1 será la regla elegida porque funciona mejor para todo $\theta \in \Theta$. Sin embargo, lo más habitual es que las funciones de riesgo se crucen, y entonces el juicio sobre qué estimador es mejor puede no ser tan sencillo.

A continuación trataremos los problemas de estimación puntual y tests de hipótesis en el marco de la teoría de la decisión. Comenzaremos por la estimación puntual: en un problema de estimación puntual, se busca estimar $h(\theta)$, con $h : \Theta \to \mathbb{R}^m$ una función conocida del parámetro, y por tanto, se toma $A \subseteq \mathbb{R}^m$ (en muchas ocasiones se desea estimar el valor del parámetro θ, por lo que la función h es la identidad). En el caso de que $m = 1$ se usa frecuentemente la función de pérdida cuadrática $L(\theta,\ a) = (a-h(\theta))^2$. A diferencia del enfoque clásico en el que se buscan estimadores con propiedades positivas como insesgadez, consistencia, suficiencia, etc, desde el enfoque decisional se buscan estimadores que minimicen el riesgo. Veamos algunos ejemplos.

Observación 1.4. Podemos destacar el *Error cuadrático medio* (Mean Squared Error) como una función de riesgo común:

$$MSE(d) = E_\theta(d(\mathbf{X}) - h(\theta))^2 = E_\theta L(\theta,\ d(\mathbf{X})) = R(\theta,\ d),$$

donde $L(\theta,\ a) = (a - h(\theta))^2$, la anteriomente definida función de pérdida cuadrática. Además,

$$\begin{aligned} R(\theta,\ d) &= E_\theta(d(\mathbf{X}) - h(\theta))^2 \\ &= E_\theta d(\mathbf{X})^2 - 2h(\theta)E_\theta d(\mathbf{X}) + h(\theta)^2 \\ &= Var_\theta d(\mathbf{X}) + [E_\theta d(\mathbf{X})]^2 - 2h(\theta)E_\theta d(\mathbf{X}) + h(\theta)^2 \\ &= Var_\theta d(\mathbf{X}) + [E_\theta d(\mathbf{X}) - h(\theta)]^2 \\ &= Var_\theta d(\mathbf{X}) + [sesgo_\theta\ d(\mathbf{X})]^2. \end{aligned}$$

Ejemplo 1.5. Riesgo de S^2 para estimar la varianza de una distribución Normal.

Sea $(X_1,\ X_2,\ ...,\ X_n)$ una muestra aleatoria de una población $N(\mu,\ \sigma^2)$. Tenemos $\theta = (\mu,\ \sigma) \in \mathbb{R} \times \mathbb{R}^+$. Procedemos a estimar $h(\theta) = \sigma^2$ utilizando la función de pérdida cuadrática y considerando estimadores de la forma $d_b(\mathbf{x}) = bs^2$, donde s^2 es la varianza muestral y b cualquier constante no negativa. Recordar que para cualquier distribución con variaza σ^2 finita, se tiene que $E_\theta(S^2) = \sigma^2$. Además, en el caso de una distribución Normal, aplicando el lema de Fisher, sabemos que:

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1} \Rightarrow Var_\theta\left((n-1)\frac{S^2}{\sigma^2}\right) = Var_\theta(\chi^2_{n-1}) \Rightarrow Var_\theta(S^2) = \frac{2\sigma^4}{n-1}$$

Usando que $R(\theta,\ d) = Var_\theta d(\mathbf{X}) + [E_\theta d(\mathbf{X}) - h(\theta)]^2$ podemos calcular el riesgo de la función d_b como

$$R(\theta,\ d_b) = Var_\theta(bS^2) + [E_\theta(bS^2) - \sigma^2]^2 = \frac{2b^2\sigma^4}{n-1} + [b\sigma^2 - \sigma^2]^2 = \sigma^4\left[\frac{2b^2}{n-1} + (b-1)^2\right]$$

Observar que esta ecuación es una función cuadrática de σ^2 y que no depende de μ, luego podemos denotar $R(\theta,\ d_b) = c_b\sigma^4$, donde $c_b = \frac{2b^2}{n-1} + (b-1)^2$ es una constante positiva. Vamos ahora a comparar dos estimadores distintos, d_b y d'_b, mediante sus funciones de riesgo, notemos que si $c_b < c'_b$ entonces $R(\theta,\ d_b) = c_b\sigma^4 < c'_b\sigma^4 = R(\theta,\ d'_b)$. Y por lo tanto, d_b será mejor estimador que d'_b. Ahora calcularemos el valor de b que nos da el valor mínimo de c_b. Para ello derivamos c_b e igualamos a 0 su derivada

$$\frac{4b}{n-1} + 2(b-1) = 0 \Rightarrow 2b + (b-1)(n-1) = 0 \Rightarrow b(n+1) = n-1 \Rightarrow b = \frac{n-1}{n+1}$$

Por lo tanto, el estimador $\hat{S}^2 = \frac{n-1}{n+1}\ S^2$ tiene el menor riesgo entre todos los estimadores de la forma bS^2. Luego $\hat{S}^2$ es mejor estimador que S^2, aunque S^2 sea un estimador insesgado.

Ejemplo 1.6. Estimación de la varianza.

De nuevo procedemos a estimar σ^2 mediante estimadores de la forma bS^2. Ahora supongamos que $(X_1,\ X_2,\ ...,\ X_n)$ es una muestra aleatoria de alguna población con varianza σ^2 finita y positiva. En este caso utilizaremos la función de pérdida

$$L(\sigma^2,\ a) = \frac{a}{\sigma^2} - 1 - \log\frac{a}{\sigma^2}$$

Esta función de pérdida es más complicada que la función de pérdida cuadrática, pero tiene buenas propiedades: notar que si $a = \sigma^2$, la pérdida es 0 y que para cualquier valor fijo de σ^2, $L(\sigma^2,\ a) \to +\infty$ cuando $a \to 0$ o $a \to +\infty$. Es decir, demasiada subestimación se penaliza tanto como demasiada sobreestimación. En cambio, en la función de pérdida cuadrática se penaliza la subestimación de manera finita, mientras que sobreestimación se penaliza de manera infinita.

Para el estimador $d_b = bs^2$, la función de riesgo es

$$\begin{aligned} R(\sigma^2,\ d_b) &= E\left(\frac{bS^2}{\sigma^2} - 1 - \log\frac{bS^2}{\sigma^2}\right) \\ &= bE\frac{S^2}{\sigma^2} - 1 - \log b - E\log\frac{S^2}{\sigma^2} \\ &= b - \log b - 1 - E\log\frac{S^2}{\sigma^2} \end{aligned}$$

La cantidad $E\log\frac{S^2}{\sigma^2}$ puede ser una función de σ^2 y otros parámetros de la población, pero no es una función de b. Por lo tanto, $R(\sigma^2,\ d_b)$ se hace mínimo para todo σ^2 en el valor de b que minimiza $b-\log b$, es decir, $1-\frac{1}{b}=0\Rightarrow b=1$. Luego el estimador de la forma bS^2 que tiene el menor riesgo para todo valor de σ^2 es $d_1=s^2$.

En el problema de estimación paramétrica tratado anteriormente, se plantea como objetivo encontrar un estimador puntual del verdadero valor del parámetro θ o de alguna función suya $h(\theta)$. Sin embargo, existen importantes situaciones en las que se tiene alguna intuición de cuál puede ser el valor θ, y lo que se pretende es confirmar o refutar dicha intuición, una vez observados los valores de la muestra $\mathbf{x}$. Y esto constituye el planteamiento de los tests de hipótesis paramétricos.

En un problema de test de hipótesis solo tenemos dos acciones posibles: aceptar H_0 o rechazar H_0, las cuales pueden ser denotadas por a_0 y a_1 respectivamente. Y por lo tanto el espacio de acciones posibles será $A=\{a_0,\ a_1\}$. Para los tests de hipótesis, la función de pérdida ha de ser tal que recoja la importancia de cada error y habrá que encontrar el test que minimice el riesgo. Además, como solo hay dos acciones posibles, la función de pérdida va a estar compuesta por dos partes:

- La función $L(\theta,\ a_0)$, que es la pérdida producida por los valores de θ si aceptamos H_0.

$$L(\theta,\ a_0)=\begin{cases}0 & si \quad \theta\in\Theta_0\\ \\ C_{II} & si \quad \theta\in\Theta_1\end{cases},$$

  siendo C_{II} el coste de cometer un error de tipo 2.

- La función $L(\theta,\ a_1)$, que es la pérdida producida por los valores de θ si rechazamos H_0.

$$L(\theta,\ a_1)=\begin{cases}C_I & si \quad \theta\in\Theta_0\\ \\ 0 & si \quad \theta\in\Theta_1\end{cases},$$

  siendo C_I el coste de cometer un error de tipo 1.

Notemos que C_I y C_{II} son funciones de θ, aunque en muchas ocasiones se suponen constantes. En ese caso, a la hora de comparar diferentes tests, lo que nos importa es minimizar la proporción $\frac{C_{II}}{C_I}$, no nos importan los valores individuales C_I y C_{II} como ocurría en el enfoque clásico. Nótese que la función de pérdida más simple en un problema de test de hipótesis es la llamada *pérdida 0-1*, es una situación particular en la que los dos tipos de error tienen la misma consecuencia: $C_I=C_{II}=1$.

Además, como sabemos que una regla de decisión $d(\mathbf{x})$ es una función de $\mathscr{X}$ que toma valores en A, tenemos que $\{\mathbf{x}\in\mathscr{X}:d(\mathbf{x})=a_0\}$ es la *región de aceptación* de H_0 y $\{\mathbf{x}\in\mathscr{X}:d(\mathbf{x})=a_1\}$ es la *región de rechazo* de H_0. A continuación, veremos la función de riesgo de los tests de hipótesis, la cual está relacionada estrechamente con la función potencia. Sea $\beta(\theta)$ la *función potencia* del test basada en una regla de decisión d, y $R=\{\mathbf{x}\in\mathscr{X}:d(\mathbf{x})=a_1\}$ la región de rechazo del test, se tiene que:

$$\beta(\theta)=P_\theta(\mathbf{X}\in R)=P_\theta(d(\mathbf{X})=a_1)$$

Por lo tanto,

$$R(\theta,\ d)=\begin{cases}C_IP_\theta(d(\mathbf{X})=a_1) & si \quad \theta\in\Theta_0\\ \\ C_{II}P_\theta(d(\mathbf{X})=a_0) & si \quad \theta\in\Theta_1\end{cases}$$

$$\Rightarrow R(\theta,\ d)=\begin{cases}C_I\beta(\theta) & si \quad \theta\in\Theta_0\\ \\ C_{II}(1-\beta(\theta)) & si \quad \theta\in\Theta_1\end{cases}$$

Ejemplo 1.7. Riesgo de un test uniformemente más potente (UMP).

Sea $(X_1,\ X_2,\ ...,\ X_n)$ una muestra aleatoria de una población $N(\theta,\ \sigma^2)$, con σ^2 conocido y $\theta \in \mathbb{R}$. Por el lema de Neyman-Pearson el test UMP de nivel α para $H_0 : \theta \geq \theta_0$ frente a $H_1 : \theta < \theta_0$ es el test que rechaza H_0 si $\frac{\bar{x}-\theta_0}{\sigma/\sqrt{n}} < -z_{1-\alpha}$, siendo $z_{1-\alpha}$ el cuantil $1-\alpha$. La función potencia para este test es

$$\beta(\theta) = P_\theta(\overline{X} < -z_{1-\alpha}\frac{\sigma}{\sqrt{n}} + \theta_0) = P(Z < -z_{1-\alpha} - \frac{\theta-\theta_0}{\sigma/\sqrt{n}}),$$

donde Z sigue una distribución $N(0,\ 1)$. Por lo tanto la función de riesgo para el test UMP será

$$R(\theta,\ d) = \begin{cases} C_I P\left(Z < -z_{1-\alpha} - \frac{\theta-\theta_0}{\sigma/\sqrt{n}}\right) & si \quad \theta \geq \theta_0 \\ \\ C_{II} P\left(Z > -z_{1-\alpha} - \frac{\theta-\theta_0}{\sigma/\sqrt{n}}\right) & si \quad \theta < \theta_0 \end{cases}$$

Capítulo 2

Teoría de la decisión Bayesiana

2.1. Introducción

En el capítulo anterior hemos planteado el problema de teoría de la decisión estadística como la búsqueda de una regla d que minimice el riesgo. Como el riesgo depende del parámetro θ, habrá que encontrar una regla que tenga riesgo mínimo para todo valor de θ, algo que, en la mayoría de los casos no existe. Por lo tanto, a la hora de buscar la "mejor" regla hay que utilizar algún criterio adicional. Una posibilidad es restringir la clase de reglas a utilizar, como en los Ejemplos 1.5. y 1.6. anteriores. Otras posibilidades son las que se plantean en este capítulo y los siguientes. En este capítulo trataremos la reglas Bayes en la que, en vez de buscar una regla que minimice el riesgo para todo valor de $\theta \in \Theta$, se busca aquella que minimice el riesgo ponderado por los valores de θ. Antes de entrar en la definición y las propiedades de la regla Bayes, haremos una breve introducción al enfoque Bayesiano de la estadística, que nos será util para entender mejor esta regla.

2.2. Enfoque Bayesiano de la estadística

En el enfoque Bayesiano θ es considerado como una cantidad aleatoria cuya variación es descrita por una distribución de probabilidad subjetiva y formulada antes de ver los datos. A esta distribución la denominamos *distribución a priori* y la denotamos por $\pi(\theta)$. Tras definir la distribución a priori se toma una muestra $\mathbf{x}$ de una población indexada por θ, con su distribución muestral denotada por $f(\mathbf{x}|\theta)$, y se actualiza la distribución a priori con la información muestral. Esta actualización es llamada *distribución a posteriori* que es la distribución condicional de θ dada la muestra $\mathbf{x}$ y se calcula de la siguiente manera:

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

donde $m(\mathbf{x})$ es la distibución marginal de $\mathbf{X}$ en el punto $\mathbf{x}$.

La distribución a posteriori es la que recoge toda la información que se tiene sobre el parámetro, resumiendo la que se tenía antes de realizar el experimento y la que proporciona la muestra. Debido a esto, la media de la distribución a posteriori se considera un buen estimador puntual de θ.

Definición 2.1. Sea $\mathscr{F}$ una clase de funciones de densidad o de funciones de masa de probabilidad $f(x|\theta)$, indexadas por θ. Una clase $\sqcap$ de distribuciones a priori es una *familia conjugada* de $\mathscr{F}$ si la distribución a posteriori está en la clase $\sqcap$ para todo $f \in \mathscr{F}$, para todas las distribuciones a priori en $\sqcap$ y para todo $\mathbf{x} \in \mathscr{X}$.

La justificación de la utilización de las familias conjugadas se basa en que simplifican notablemente los cálculos y en que recogen la idea de que, si una distribución de probabilidad es adecuada para el parámetro antes de observar los datos, sería logico que lo siguiera siendo una vez observados dichos datos. A continuación, podemos observar algunos ejemplos de familias conjugadas en la Tabla (2.1).

Distribución muestral: $f(x\|\theta)$	**A priori:** $\pi(\theta)$	**A posteriori:** $\pi(\theta\|x)$
$N(\theta,\ \sigma^2)$	$N(\mu,\ \tau^2)$	$N(\frac{\sigma^2\mu+x\tau^2}{\sigma^2+\tau^2},\ \frac{\sigma^2\tau^2}{\sigma^2+\tau^2})$
$G(\nu,\ \theta)$	$G(\alpha,\ \beta)$	$G(\alpha+\nu,\ \beta+x)$
$Bin(n,\ \theta)$	$B(\alpha,\ \beta)$	$B(\alpha+x,\ \beta+n-x)$
$P(\theta)$	$G(\alpha,\ \beta)$	$G(\alpha+x,\ \beta+1)$
$BinNeg(r,\ \theta)$	$B(\alpha,\ \beta)$	$B(\alpha+r,\ \beta+x)$

Tabla 2.1: Tabla familias conjugadas.

Ejemplo 2.2. Estimación Binomial de Bayes.

Sean $X_1,\ X_2,\ ...,\ X_n$ variables aleatorias independientes e identicamente distribuidas (iid) con distribución $Bernoulli(p)$ y queremos estimar p. Como la suma de variables aleatorias iid con distribución $Bernoulli(p)$ sigue una distribución $Binomial(n,\ p)$, tenemos $Y=\sum_{i=1}^{n} X_i \sim Bin(n,\ p)$. Suponemos que la distribución a priori de p es $Beta(\alpha,\ \beta)$. De toda esta información se deduce que

$$\pi(p|y)=\frac{f(y|p)\pi(p)}{m(y)}=\frac{\binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{y+\alpha-1}(1-p)^{n-y+\beta-1}}{m(y)}=K(y)p^{y+\alpha-1}(1-p)^{n-y+\beta-1},$$

con $p\in(0,1)$ y donde $K(y)$ no depende de p. Como $\pi(p|y)$ es una función de densidad sobre $(0,1)$, se tiene que

$$K(y)=\frac{1}{Beta(y+\alpha,n-y+\beta)}=\frac{\Gamma(\alpha+n+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}$$

Por lo tanto,

$$\pi(p|y)=\frac{\Gamma(\alpha+n+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}p^{y+\alpha-1}(1-p)^{n-y+\beta-1},$$

Luego la distribución a posteriori sigue una distribución $Beta(y+\alpha,\ n-y+\beta)$. Como un estimador de Bayes de p es la media de la distribución a porteriori, se tiene que

$$\overline{p_b}=\frac{y+\alpha}{\alpha+n+\beta}$$

Notar que la distribución a priori tiene media $\frac{\alpha}{\alpha+\beta}$, que puede ser el mejor estimador de p sin ver los datos y, si ignoramos la información a priori, seguramente tomaríamos $p=\frac{y}{n}$ como un estimador de p. Observar que el estimador de Bayes de p que hemos obtenido anteriormente combina toda esta información:

$$\overline{p_b}=\frac{y+\alpha}{\alpha+n+\beta}=\frac{n}{\alpha+n+\beta}\frac{y}{n}+\frac{\alpha+\beta}{\alpha+n+\beta}\frac{\alpha}{\alpha+\beta},$$

con lo que $\overline{p_b}$ es una combinación lineal convexa de la media a priori y la media muestral. Además, podemos observar en este ejemplo que la distribución Beta es una familia conjugada respecto a la distribución Binomial.

Ejemplo 2.3. Estimación Normal de Bayes.

Sea $(X_1,\ X_2,\ ...,\ X_n)$ una muestra aleatoria de una distribución $N(\theta,\sigma^2)$ y supongamos que la distribución a priori de θ es $N(\mu,\tau^2)$, con σ^2, μ y τ^2 conocidos. Sabemos que $\overline{\mathbf{x}}\sim N(\theta,\frac{\sigma^2}{n})$ y queremos estimar θ. De toda esta información se deduce que

$$\pi(\theta|\overline{\mathbf{x}})=\frac{f(\overline{\mathbf{x}}|\theta)\pi(\theta)}{m(\overline{\mathbf{x}})}=\frac{\frac{\sqrt{n}}{2\pi\sigma\tau}e^{-\frac{1}{2}\left[\frac{n(\overline{\mathbf{x}}-\theta)^2}{\sigma^2}+\frac{(\theta-\mu)^2}{\tau^2}\right]}}{m(\overline{\mathbf{x}})}=K(\overline{\mathbf{x}})e^{-\frac{1}{2}\left[\frac{\left(\theta-\frac{\overline{\mathbf{x}}n\tau^2+\mu\sigma^2}{n\tau^2+\sigma^2}\right)^2}{\frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}}\right]},$$

con $\theta \in (-\infty, +\infty)$ y donde $K(\overline{\mathbf{x}})$ no depende de θ. Obtenemos que $E(\theta|\overline{\mathbf{x}}) = \frac{\overline{\mathbf{x}}n\tau^2+\mu\sigma^2}{n\tau^2+\sigma^2}$ y que $Var(\theta|\overline{\mathbf{x}}) = \frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}$. Es decir, la distribución a posteriori es $N(\frac{\overline{\mathbf{x}}n\tau^2+\mu\sigma^2}{n\tau^2+\sigma^2}, \frac{\tau^2\sigma^2}{n\tau^2+\sigma^2})$. Y como un estimador de θ es la media de la distribución a posteriori se tiene que

$$\overline{\theta_b} = E(\theta|\overline{\mathbf{x}}) = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}}\overline{\mathbf{x}} + \frac{\frac{\sigma^2}{n}}{\tau^2 + \frac{\sigma^2}{n}}\mu.$$

2.3. Regla Bayes

En este caso, el criterio para minimizar el riesgo se obtiene a través de la asignación de pesos π a los distintos valores del parámetro θ, de forma que se busca la regla que minimiza el riesgo ponderado. A esta regla la denominamos regla Bayes. La distribución π puede verse como una a priori, es decir, considerando θ como una variable aleatoria, pero esto no es necesario y puede pensarse que θ es una constante desconocida a la que se da unos pesos π que ponderen la importancia de un error en los distintos valores de θ.

Definición 2.4. Sea L una función de pérdida y R la función de riesgo asociada, denominamos *riesgo Bayes* de la función decisión $d \in D$ con respecto a la distribución a priori definida por $\pi(\theta)$ a:

$$R_B(\pi,\, d) = E_\pi R(\theta,\, d) = \int_\Theta R(\theta,\, d)\pi(\theta)\, d\theta$$

- Si la distribución a priori y la distribución muestral son continuas se tiene que:

$$R_B(\pi,\, d) = \int_\Theta R(\theta,\, d)\pi(\theta)\, d\theta = \int_\Theta \int_{\mathscr{X}} L(\theta,\, d(\mathbf{x}))f(\mathbf{x}|\theta)\pi(\theta)\, d\mathbf{x}\, d\theta =$$
$$= \int_\Theta \int_{\mathscr{X}} L(\theta,\, d(\mathbf{x}))f(\mathbf{x},\, \theta)\, d\mathbf{x}\, d\theta,$$

  donde $f(\mathbf{x},\, \theta)$ es la función de densidad conjunta de la muestra y el parámetro.

- Si la distribución a priori y la distribución muestral son discretas se tiene que:

$$R_B(\pi,\, d) = \sum_\Theta [R(\theta,\, d)\pi(\theta)] = \sum_\Theta \sum_{\mathscr{X}} [L(\theta,\, d(\mathbf{x}))f(\mathbf{x}|\theta)\pi(\theta)] = \sum_\theta \sum_{\mathscr{X}} [L(\theta,\, d(\mathbf{x}))f(\mathbf{x},\, \theta)]$$

También puede haber una distribución continua y otra discreta.

Definición 2.5. Una regla de decisión $d^* \in D$ se dice que es una *regla Bayes* con respecto a la distribución a priori $\pi(\theta)$ si:

$$R_B(\pi,\, d^*) = inf_{d\in D}\, R_B(\pi,\, d)$$

Además, denominamos *riesgo Bayes mínimo* a $inf_{d\in D}\, R_B(\pi,\, d)$.

Teorema 2.6. *Para calcular la regla Bayes basta tomar la regla que consiste en escoger, para cada valor de la muestra* $\mathbf{x}$*, la acción que minimice la esperanza de la función de pérdida con respecto a la distribución a posteriori.*

Demostración: Lo hacemos en el caso continuo, el caso discreto es igual pero cambiando las integrales por sumas.

$$\begin{aligned} R_B(\pi,\, d) &= \int_\Theta \int_{\mathscr{X}} L(\theta,\, d(\mathbf{x}))f(\mathbf{x},\, \theta)\, d\mathbf{x}\, d\theta \\ &= \int_\Theta \int_{\mathscr{X}} L(\theta,\, d(\mathbf{x}))\pi(\theta|\mathbf{x})m(\mathbf{x})\, d\mathbf{x}\, d\theta \\ &= \int_{\mathscr{X}} m(\mathbf{x}) \int_\Theta L(\theta,\, d(\mathbf{x}))\pi(\theta|\mathbf{x})\, d\theta\, d\mathbf{x} \end{aligned}$$

donde $m(\mathbf{x})$ es la distribución marginal de $\mathbf{X}$ y $\pi(\theta|\mathbf{x})$ es la distribución a posteriori de θ. Por lo tanto, la regla Bayes es la regla que minimiza la cantidad $\int_\Theta L(\theta,\, d(\mathbf{x}))\pi(\theta|\mathbf{x})\, d\theta$ para cada valor de $\mathbf{x} \in \mathscr{X}$.

□

2.4. Estimación puntual

Teorema 2.7. *Consideramos el problema de estimación de un parámetro $\theta \in \Theta \subseteq \mathbb{R}$ con respecto a la función de pérdida cuadrátia $L(\theta,\ d) = (\theta - d)^2$. La regla $d(\mathbf{x}) = E(\theta|X = \mathbf{x})$ es una regla Bayes.*

Demostración: Lo hacemos en el caso continuo. Por el Teorema 2.6. sabemos que la minimización de $R_B(\pi,\ d)$ es lo mismo que la minimización de $\int_\Theta (\theta - d(\mathbf{x}))^2 \pi(\theta|\mathbf{x})\, d\theta$. Se tiene:

$$\begin{aligned}\int_\Theta (\theta - d(\mathbf{x}))^2 \pi(\theta|\mathbf{x})\, d\theta &= \int_\Theta (\theta - E(\theta|\mathbf{x}) + E(\theta|\mathbf{x}) - d(\mathbf{x}))^2 \pi(\theta|\mathbf{x})\, d\theta \\ &= \int_\Theta (\theta - E(\theta|\mathbf{x}))^2 \pi(\theta|\mathbf{x})\, d\theta + (E(\theta|\mathbf{x}) - d(\mathbf{x}))^2\end{aligned}$$

Por lo tanto, $\int_\Theta (\theta - d(\mathbf{x}))^2 \pi(\theta|\mathbf{x})\, d\theta$ es mínimo si y solo si $E(\theta|X = \mathbf{x}) = d(\mathbf{x})$, como queríamos demostrar.

□

Ejemplo 2.8. Sean $X_1, X_2, ..., X_n$ variables aleatorias iid con distribución $Bernoulli(p)$. Sabiendo que $Y = \sum_{i=1}^n X_i \sim Bin(n,\ p)$ queremos estimar p aplicando la función de pérdida cuadrática $L(p,\ d(y)) = (p - d(y))^2$ y tomando como distribución a priori de p, $\pi(p) = 1$, para $0 < p < 1$.

Observamos que es un caso particular del Ejemplo 2.2., luego la distribución a posteriori sigue una distribución $Beta(y+1,\ n-y+1)$. Por lo tanto el estimador de Bayes es $d^*(y) = \frac{y+1}{n+2}$. El riesgo Bayes es:

$$\begin{aligned}R_B(\pi,\ d^*) &= \int_0^1 R(p,\ d^*)\pi(p)\, dp \\ &= \int_0^1 \sum_{y=0}^n \left[(d^*(y) - p)^2 f(y|p)\right] dp \\ &= \int_0^1 E_p\left(\frac{Y+1}{n+2} - p\right)^2 dp \\ &= \int_0^1 \left[\frac{E_p(Y+1)^2}{(n+2)^2} - \frac{2pE_p(Y+1)}{n+2} + p^2\right] dp \\ &= \int_0^1 \left[\frac{1}{(n+2)^2}[np(1-p) + n^2p^2 + 2np + 1] - \frac{2p}{n+2}(np+1) + p^2\right] dp \\ &= \frac{1}{(n+2)^2}\int_0^1 [np - np^2 + 1 - 4p + 4p^2]\, dp \\ &= \frac{1}{6(n+2)}\end{aligned}$$

Ejemplo 2.9. Sea $(X_1,\ X_2,\ ...,\ X_n)$ una muestra aleatoria de una distribución $N(\mu,\ 1)$ y supongamos que la distribución a priori de μ es una $N(0,\ 1)$. Procedemos a estimar μ aplicando la función de pérdida cuadrática $L(\mu,\ d(\mathbf{x})) = (\mu - d(\mathbf{x}))^2$, para $\mu \in (-\infty,\ +\infty)$. Entonces, sabemos por el Ejemplo 2.3. que la distribución a posteriori de μ será $N(\frac{\sum_{i=1}^n x_i}{n+1},\ \frac{1}{n+1})$. Por lo tanto el estimador de Bayes es $d^*(\mathbf{x}) = \frac{\sum_{i=1}^n x_i}{n+1}$.

El riesgo Bayes es:

$$\begin{aligned} R_B(\pi,\, d^*) &= \int_{-\infty}^{+\infty} R(\mu,\, d^*)\pi(\mu)\, d\mu \\ &= \int_{-\infty}^{+\infty} \int_{\mathbb{R}^n} \left[(d^*(\mathbf{x}) - \mu)^2 f(\mathbf{x}|\mu)\right] \pi(\mu)\, d\mathbf{x}\, d\mu \\ &= \int_{-\infty}^{+\infty} E_\mu \left(\frac{\sum_{i=1}^n X_i}{n+1} - \mu \right)^2 \pi(\mu)\, d\mu \\ &= \int_{-\infty}^{+\infty} \frac{n+\mu^2}{(n+1)^2} \pi(\mu)\, d\mu \\ &= \int_{-\infty}^{+\infty} \frac{n+\mu^2}{(n+1)^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\mu^2}\, d\mu \\ &= \frac{1}{n+1} \end{aligned}$$

Teorema 2.10. *Consideramos el problema de estimación de un parámetro $\theta \in \Theta \subseteq \mathbb{R}$ con respecto a la función de pérdida cuadrátia $L(\theta,\, d) = w(\theta)(\theta - d)^2$, con $w(\theta) > 0$, para todo $\theta \in \Theta \subseteq \mathbb{R}$. Entonces la regla*

$$d(\mathbf{x}) = \frac{E(\theta w(\theta)|X = \mathbf{x})}{E(w(\theta)|X = \mathbf{x})}$$

es una regla Bayes.

Demostración: Lo hacemos en el caso continuo. Por el Teorema 2.6. sabemos que la minimización de $R_B(\pi,\, d)$ es lo mismo que la minimización de $\int_\Theta w(\theta)(\theta - d(\mathbf{x}))^2 \pi(\theta|\mathbf{x})\, d\theta$. Se tiene:

$$\begin{aligned} \int_\Theta W(\theta)(\theta - d(\mathbf{x}))^2 \pi(\theta|\mathbf{x})\, d\theta &= \int_\Theta \left(\theta - \frac{E(\theta w(\theta)|\mathbf{x})}{E(w(\theta)|\mathbf{x})} + \frac{E(\theta w(\theta)|\mathbf{x})}{E(w(\theta)|\mathbf{x})} - d(\mathbf{x}) \right)^2 \pi(\theta|\mathbf{x})\, d\theta \\ &= \int_\Theta \left(\theta - \frac{E(\theta w(\theta)|\mathbf{x})}{E(w(\theta)|\mathbf{x})} \right)^2 \pi(\theta|\mathbf{x})\, d\theta + \left(\frac{E(\theta w(\theta)|\mathbf{x})}{E(w(\theta)|\mathbf{x})} - d(\mathbf{x}) \right)^2 \end{aligned}$$

Por lo tanto, $\int_\Theta w(\theta)(\theta - d(\mathbf{x}))^2 \pi(\theta|\mathbf{x})\, d\theta$ es mínimo si y solo si $d(\mathbf{x}) = \frac{E(\theta\ w(\theta)|X=\mathbf{x})}{E(w(\theta)|X=\mathbf{x})}$, como queríamos demostrar.

□

2.5. Test de Hipótesis

Es útil observar la relación entre los tests de hipótesis Bayesianos y los tests de hipótesis clásicos. En los tests clásicos, el valor crítico de la región de rechazo es función de la probabilidad de cometer el error de tipo I, mientras que en los tests Bayesianos el valor crítico dependerá de la función de pérdida y la distribución a priori. Se puede considerar que el método Bayesiano proporciona una elección racional del tamaño del test. La estadística clásica no proporciona tales pautas, utiliza con frecuencia tamaños estándar. Sin embargo, lo que hace el enfoque Bayesiano es elegir el tamaño de acuerdo con factores subjetivos.

Teorema 2.11. *Sea $\mathbf{X} = (X_1,\, X_2,\, \ldots,\, X_n)$ una variable aleatoria discreta (continua) con pmf (pdf) $f(\mathbf{x}|\theta), \theta \in \Theta = \{\theta_0,\, \theta_1\}$. Sean $\pi(\theta_0) = \pi_0$ y $\pi(\theta_1) = 1 - \pi_0 = \pi_1$ la función de probabilidad a priori en Θ. La regla Bayes para el test $H_0 : \mathbf{X} \sim f(\mathbf{x}|\theta_0)$ frente a $H_1 : \mathbf{X} \sim f(\mathbf{x}|\theta_1)$ usando la función de*

pérdida:

$$L(\theta,\, a_0) = \begin{cases} 0 & si \quad \theta = \theta_0 \\ \\ C_{II} & si \quad \theta = \theta_1 \end{cases}$$
$$L(\theta,\, a_1) = \begin{cases} C_I & si \quad \theta = \theta_0 \\ \\ 0 & si \quad \theta = \theta_1 \end{cases} \tag{2.1}$$

será rechazar H_0 *si* $\frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} \geq \frac{C_I \pi_0}{C_{II}\pi_1}$.

Demostración: Por el Teorema 2.6. sabemos que es suficiente encontrar un d que, minimice para cada $\mathbf{x} \in \mathscr{X}$, $E(L(\theta, d(\mathbf{x}))|\mathbf{x})$. La distribución a posteriori de θ es:

$$\pi(\theta|\mathbf{x}) = \frac{\pi(\theta) f(\mathbf{x}|\theta)}{\sum_\theta \pi(\theta) f(\mathbf{x}|\theta)} = \frac{\pi(\theta) f(\mathbf{x}|\theta)}{\pi_0 f(\mathbf{x}|\theta_0) + \pi_1 f(\mathbf{x}|\theta_1)} = \begin{cases} \frac{\pi(\theta_0) f(\mathbf{x}|\theta_0)}{\pi(\theta_0) f(\mathbf{x}|\theta_0) + \pi(\theta_1) f(\mathbf{x}|\theta_1)} & si \quad \theta = \theta_0 \\ \\ \frac{\pi(\theta_1) f(\mathbf{x}|\theta_1)}{\pi(\theta_0) f(\mathbf{x}|\theta_0) + \pi(\theta_1) f(\mathbf{x}|\theta_1)} & si \quad \theta = \theta_1 \end{cases}$$

Por lo tanto,

$$E(L(\theta,\, d(\mathbf{x}))|\mathbf{x}) = \begin{cases} C_I \pi(\theta_0|\mathbf{x}) & si \quad d(\mathbf{x}) = a_1 \\ \\ C_{II} \pi(\theta_1|\mathbf{x}) & si \quad d(\mathbf{x}) = a_0 \end{cases}$$

Luego, rechazamos H_0, es decir, $d(\mathbf{x}) = a_1$ si

$$C_I \pi(\theta_0|\mathbf{x}) \leq C_{II} \pi(\theta_1|\mathbf{x}) \Leftrightarrow C_I \pi_0 f(\mathbf{x}|\theta_0) \leq C_{II} \pi_1 f(\mathbf{x}|\theta_1) \Leftrightarrow \frac{C_I \pi_0}{C_{II}\pi_1} \leq \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)}.$$

□

Ejemplo 2.12. Sean $X_1,\, X_2,\, ...,\, X_n$ variables aleatorias iid con distribución $N(\mu,\, 1)$. Para el test $H_0 : \mu = \mu_0$ frente a $H_1 : \mu = \mu_1 (> \mu_0)$, tomando $C_I = C_{II}$ en la función de pérdida (2.1), el Teorema 2.11. nos asegura que la regla de Bayes es la que rechaza H_0 si

$$\frac{f(\mathbf{x}|\mu_1)}{f(\mathbf{x}|\mu_0)} \geq \frac{\pi_0}{1-\pi_0}$$

es decir,

$$e^{\left[-\frac{1}{2}\sum_{i=1}^n (x_i-\mu_1)^2 + \frac{1}{2}\sum_{i=1}^n (x_i-\mu_0)^2\right]} \geq \frac{\pi_0}{1-\pi_0}$$
$$\Rightarrow e^{\left[(\mu_1-\mu_0)\sum_{i=1}^n x_i + \frac{n}{2}(\mu_0^2-\mu_1^2)\right]} \geq \frac{\pi_0}{1-\pi_0}$$

Esto sucede si y solo si

$$(\mu_1-\mu_0)\sum_{i=1}^n x_i + \frac{n(\mu_0^2-\mu_1^2)}{2} \geq \log\frac{\pi_0}{1-\pi_0}$$
$$\sum_{i=1}^n x_i \geq \frac{\log(\frac{\pi_0}{1-\pi_0})}{\mu_1-\mu_0} + \frac{n(\mu_1+\mu_0)}{2}$$
$$\frac{\sum_{i=1}^n x_i}{n} \geq \frac{\log(\frac{\pi_0}{1-\pi_0})}{n(\mu_1-\mu_0)} + \frac{\mu_1+\mu_0}{2},$$

donde el logaritmo tienen base e.

En particular, si $\pi_0 = \frac{1}{2}$, la región de rechazo consiste en $\bar{x} \geq \frac{\mu_1+\mu_0}{2}$. A continuación mostramos un gráfico tomando $\mu_0 = 0$, $\mu_1 = 1$, $n = 1$ y variando π_0:

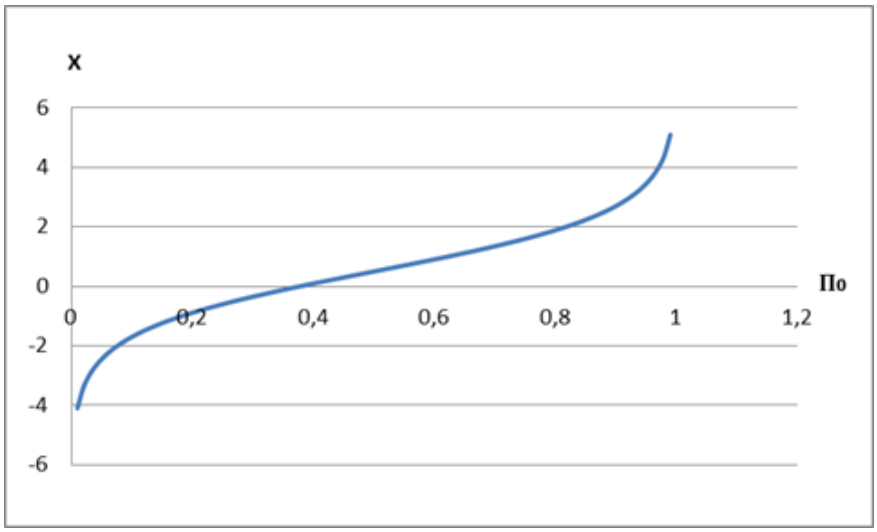


Figura 2.1: Gráfica para $\mu_0 = 0$ y $\mu_1 = 1$.

Observación 2.13. Generalización del Teorema 2.11. Supongamos $\Theta = \{\theta_1,\ \theta_2,\ ...,\ \theta_k\}$, donde θ puede tomar cualquiera de los k valores $\theta_1,\ \theta_2,\ ...,\ \theta_k$. El problema es observar $\mathbf{x}$ y decidir cuál de los θ_i es el valor correcto de θ. Sea $H_i : \theta = \theta_i$, $i = 1,\ 2,\ ...,\ k$, y supongamos que $\pi(\theta_i) = \pi_i$, $i = 1,\ 2,\ ...,\ k$, $\sum_1^k \pi_i = 1$, es la distribución a priori de θ. Sea

$$L(\theta_i,\ d) = \begin{cases} 1 & si \quad d\ elige\ \theta_j,\ j \neq i \\ \\ 0 & si \qquad d\ elige\ \theta_i \end{cases}$$

El problema es encontrar una regla d que minimice $R_B(\pi,\ d)$. La regla de Bayes es aceptar $H_i : \theta = \theta_i$ $(i = 1,\ 2,\ ...,\ k)$ si

$$\pi_i f(\mathbf{x}|\theta_i) \geq \pi_j f(\mathbf{x}|\theta_j)$$

para todo $j \neq i$, $j = 1,\ 2,\ ...,\ k$, donde cualquier punto que situado en más de una de esas regiones se asigna a cualquiera de ellas.

Ejemplo 2.14. Ilustración de la observación 2.12.

Sea $(X_1,\ X_2,\ ...,\ X_n)$ una muestra de una distribución $N(\mu,\ 1)$ y supongamos que μ puede ser $\mu_1,\ \mu_2$ ó μ_3, siendo $\mu_1 < \mu_2 < \mu_3$. Supongamos, por simplicidad, que $\pi_1 = \pi_2 = \pi_3$. Entonces aceptaremos $H_i : \mu = \mu_i,\ i = 1,\ 2,\ 3$, si

$$\pi_i e^{-\frac{1}{2}\sum_{k=1}^n (x_k - \mu_i)^2} \geq \pi_j e^{-\frac{1}{2}\sum_{k=1}^n (x_k - \mu_j)^2},$$

para cada $j \neq i,\ j = 1,\ 2,\ 3$. Por lo tanto, aceptaremos H_i si

$$(\mu_i - \mu_j)\bar{x} + \frac{\mu_j^2 - \mu_i^2}{2} \geq 0$$
$$\Rightarrow \bar{x}(\mu_i - \mu_j) \geq \frac{(\mu_i - \mu_j)(\mu_i + \mu_j)}{2},$$

con $j = 1,\ 2,\ 3$ y $j \neq i$.

Así, la región de aceptación de H_1 es:

$$\bar{x} \leq \tfrac{1}{2}(\mu_1 + \mu_2)$$

La región de aceptación de H_2 es:

$$\bar{x} \geq \tfrac{1}{2}(\mu_1 + \mu_2) \ \text{ y } \ \bar{x} \leq \tfrac{1}{2}(\mu_2 + \mu_3)$$

La región de aceptación de H_3 es:

$$\bar{x} \geq \tfrac{1}{2}(\mu_2 + \mu_3)$$

En particular, si $\mu_1 = 0,\ \mu_2 = 2$ y $\mu_3 = 4$, aceptaremos H_1 si $\bar{x} \leq 1,\ H_2$ si $1 \leq \bar{x} \leq 3$ y H_3 si $\bar{x} \geq 3$.

Capítulo 3

Minimax

Otro criterio para la búsqueda de la decisión óptima es el criterio minimax. En este criterio, se plantea el problema de decisión como un juego contra un contrincante genérico (la naturaleza), en el que se tiene como objetivo encontrar la regla que minimice el máximo riesgo en función de θ. Es decir, se intenta minimizar el riesgo en la peor situación posible. Se piensa que la naturaleza es un contrincante que va a poner las cosas de la peor manera posible para nuestros intereses. A la hora de comparar este enfoque con la regla Bayes, se puede destacar que aquí no se necesita la distribución a priori, es decir, no tenemos que suponer conocimiento previo del parámetro o una función de peso a la hora de calcular el riesgo.

Definición 3.1. Se denomina *estimador minimax* a aquel $d^* \in D$ que verifica que:

$$sup_{\theta\in\Theta}\, R(\theta,\, d^*) = inf_{d\in D}\, sup_{\theta\in\Theta}\, R(\theta,\, d)$$

Observación 3.2. Para cualquier $d \in D$ y cualquier distribución a priori $\pi(\theta)$ sobre Θ, se tiene que

$$R_B(\pi,\, d) = \int_\Theta R(\theta,\, d)\pi(\theta)\, d\theta \leq sup_{\theta\in\Theta}\, R(\theta,\, d) \tag{3.1}$$

3.1. Estimación puntual. Relación con las reglas de Bayes

Desde el punto de vista intuitivo, parece que las reglas que aparecen como solución de un problema minimax van a ser demasiado conservadoras, porque consideran la peor situación posible. Sin embargo, en muchas ocasiones el criterio minimax da lugar a reglas con buenas propiedades. Además, bajo ciertas condiciones, el estimador minimax coincide con una regla Bayes.

Observación 3.3. Sea $R_B : \mathscr{F} \times D \to \mathbb{R}$, notemos que se verifica

$$\underline{V} = sup_{\pi\in\mathscr{F}}\, inf_{d\in D}\, R_B(\pi,\, d) \leq inf_{d\in D}\, sup_{\pi\in\mathscr{F}}\, R_B(\pi,\, d) = \overline{V}; \tag{3.2}$$

donde denominamos $\mathscr{F}$ a la familia de todas las distribuciones a priori sobre Θ, a $\underline{V}$ lo denominamos valor inferior y a $\overline{V}$ valor superior. Efectivamente, dada $d \in D$ resulta claro que

$$inf_{d\in D}\, R_B(\pi,\, d) \leq inf_{d\in D}\, sup_{\pi\in\mathscr{F}}\, R_B(\pi,\, d),$$

de donde se obtiene inmediatamente la desigualdad (3.2).

Definición 3.4. Se denomina *distribución a priori menos favorable (o más desfavorable)* a la distribución a priori π^* tal que

$$inf_{d\in D}\, R_B(\pi^*,\, d) = sup_{\pi\in\mathscr{F}}\, inf_{d\in D}\, R_B(\pi,\, d)$$

Teorema 3.5. *Si $\underline{V} = \overline{V}$, d^* es un estimador minimax y π^* una distribución menos favorable, entonces d^* es un estimador Bayes con distribución a priori π^*.*

Demostración: Como d^* es un estimador minimax

$$R_B(\pi,\ d^*) \leq sup_{\theta\in\Theta}\ R(\theta,\ d^*) = inf_{d\in D}\ sup_{\theta\in\Theta}\ R(\theta,\ d) \leq inf_{d\in D}\ sup_{\pi\in\mathscr{F}}\ R_B(\pi,\ d)$$

donde la primera desigualdad es por (3.1), la igualdad por ser d^* minimax y la última desigualdad porque las distribuciones con masa 1 en un punto θ son un subconjunto de las distribuciones a priori para todo $\pi \in \mathscr{F}$. En particular para π^*.

$$R_B(\pi^*,\ d^*) \leq inf_{d\in D}\ sup_{\pi\in\mathscr{F}}\ R_B(\pi,\ d)$$

Por tanto, como $\underline{V} = \overline{V}$ se tiene que

$$R_B(\pi^*,\ d^*) \leq inf_{d\in D}\ sup_{\pi\in\mathscr{F}}\ R_B(\pi,\ d) = sup_{\pi\in\mathscr{F}}\ inf_{d\in D}\ R_B(\pi,\ d) = inf_{d\in D}\ R_B(\pi^*,\ d)$$

donde en la última igualdad se utiliza que π^* es la distribución menos favorable.

Por tanto, d^* es un estimador Bayes con distribución a priori π^*.

□

Teorema 3.6. *Si hay una distribución π^* sobre Θ, de modo que es d^* un estimador Bayes de θ respecto a π^*, y la función de riesgo $R(\theta,\ d^*)$ es constante en Θ, entonces d^* es un estimador minimax de θ y π^* es la distribución menos favorable.*

Demostración: Como $R(\theta,\ d^*) = r^*$ para todo $\theta \in \Theta$, entonces

$$\begin{aligned} r^* &= R_B(\pi^*,\ d^*) \\ &= inf_{d\in D}\ R_B(\pi^*,\ d) \\ &\leq sup_{\pi\in\mathscr{F}}\ inf_{d\in D}\ R_B(\pi,\ d) \\ &\leq inf_{d\in D}\ sup_{\pi\in\mathscr{F}}\ R_B(\pi,\ d) \\ &\leq sup_{\theta\in\Theta}\ R(\theta,\ d^*) \\ &= r^* \end{aligned}$$

luego todas las desigualdades son igualdades, por lo que π^* es una distribución a priori menos favorable y $V = \overline{V}$, lo que implica que d^* es un estimador minimax por el Teorema 3.5.

□

Observación 3.7. En el Teorema 3.6., si d^* es el único estimador de Bayes con distribución a priori π^*, entonces es también el único estimador minimax, puesto que para todo $d \neq d^*$

$$sup_{\theta\in\Theta}\ R(\theta,\ d) \geq R_B(\pi^*,\ d) > R_B(\pi^*,\ d^*) = r^* = sup_{\theta\in\Theta}\ R(\theta,\ d^*)$$

y entonces d no puede ser un estimador minimax.

Ejemplo 3.8. Sean $X_1, X_2, ..., X_n$ variables aleatorias iid con distribución $Bernoulli(\theta)$. Sabiendo que $Y = \sum_{i=1}^{n} X_i \sim Bin(n,\ \theta)$ queremos estimar $\theta \in (0,\ 1)$ con una muestra de tamaño 1 de Y. Tomemos la distribución a priori de θ como una $B(\alpha,\ \beta)$, entonces la distribución a posteriori es una $B(\alpha + y,\ \beta + n - y)$. Por lo tanto, por el Ejemplo 2.2. el estimador de Bayes de θ es:

$$\overline{\theta_b} = \frac{\alpha + y}{\alpha + \beta + n} = d^*,$$

siendo lineal en y, ya que: $d^* = ay + b$, con $a = \frac{1}{\alpha+\beta+n}$ y $b = \frac{\alpha}{\alpha+\beta+n}$. La función de riesgo del estimador de la forma $ay + b$ es:

$$\begin{aligned} R(\theta,\ d^*) &= E_\theta[L(\theta,\ d^*)] \\ &= E_\theta[(aY + b - \theta)^2] \\ &= E_\theta[(aY + b)^2 - 2\theta(aY + b) + \theta^2] \\ &= a^2 n\theta(1-\theta) + a^2 n^2\theta^2 + 2abn\theta + b^2 - 2an\theta^2 - 2b\theta + \theta^2 \\ &= \theta^2[(an-1)^2 - a^2 n] + \theta[2b(an-1) + a^2 n] + b^2 \end{aligned}$$

Deseamos que el riesgo sea constante y para ello igualamos los coeficientes que multiplican a θ^2 y θ a 0:

$$\begin{cases} (an-1)^2 - a^2 n = 0 \\ \\ 2b(an-1) + a^2 n = 0 \end{cases}$$

Y obtenemos que:

$$a = \frac{\sqrt{n}}{n(1+\sqrt{n})} \qquad\qquad b = \frac{1}{2(1+\sqrt{n})}$$

Por lo tanto, como $a = \frac{1}{\alpha+\beta+n}$ y $b = \frac{\alpha}{\alpha+\beta+n}$, se tiene que $\alpha = \beta = \frac{\sqrt{n}}{2}$. En conclusión,

$$d^* = \frac{\sqrt{n}}{n(1+\sqrt{n})} y + \frac{1}{2(1+\sqrt{n})}$$

es un estimador de Bayes frente a una distribución a priori π^* que es una $B(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2})$ y su riesgo es:

$$R(\theta, d^*) = b^2 = \frac{1}{4(1+\sqrt{n})^2}$$

que, al no depender de θ es constante. Entonces por el Teorema 3.6. deducimos que d^* es un estimador minimax.

Teorema 3.9. *Sea $\{\pi_k(\theta) : k \geq 1\}$ una sucesión de distribuciones a priori sobre Θ, y $\{d_k : k \geq 1\}$ la correspondiente sucesión de estimadores de Bayes con riesgo de Bayes $R_B(\pi_k,\ d_k)$. Si existe un estimador d^* tal que:*

$$sup_{\theta\in\Theta}\, R(\theta,\ d^*) \leq \overline{lim}_{k\to\infty}\, R_B(\pi_k,\ d_k)$$

Entonces d^* es un estimador minimax.

Demostración: Supongamos que d^* no es un estimador minimax, entonces existe otro estimador $\overline{d}$ tal que:

$$sup_{\theta\in\Theta}\, R(\theta,\ \overline{d}) < sup_{\theta\in\Theta}\, R(\theta,\ d^*)$$

Pero como los $\{d_k\}$ son los estimadores de Bayes correspondientes a las distribuciones a priori $\{\pi_k(\theta) : k \geq 1\}$ se tiene que:

$$R_B(\pi_k,\ d_k) = \int_\Theta R(\theta,\ d_k)\pi_k(\theta)\, d\theta \leq \int_\Theta R(\theta,\ \overline{d})\pi_k(\theta)\, d\theta \leq sup_{\theta\in\Theta}\, R(\theta,\ \overline{d}),$$

para todo $k \geq 1$. Por lo tanto,

$$\overline{lim}_{k\to\infty}\, R_B(\pi_k,\ d_k) \leq sup_{\theta\in\Theta}\, R(\theta,\ \overline{d}) < sup_{\theta\in\Theta}\, R(\theta,\ d^*)$$

Y con esto llegamos a una contradicción, por lo tanto d^* es un estimador minimax.

□

Corolario 3.10. Si d^* es un estimador con función de riesgo constante, $R(\theta,\ d^*) = r^*$ para todo $\theta \in \Theta$ y existe una sucesión de distribuciones a priori $\{\pi_k(\theta) : k \geq 1\}$ tal que $lim_{k\to\infty}\, R_B(\pi_k,\ d_k) = r^*$, donde $\{d_k : k \geq 1\}$ son los estimadores de Bayes respecto a ellas, entonces d^* es un estimador minimax.

Demostración: Inmediata del Teorema 3.9.

Ejemplo 3.11. Sea $(X_1,\ X_2,\ ...,\ X_n)$ una muestra aleatoria de una distribución $N(\theta, \sigma^2)$ con σ conocida, y se queremos estimar θ con una muestra de tamaño n, bajo el error de pérdida cuadrática. El estimador $\overline{X}$ tiene riesgo $\frac{\sigma^2}{n}$ constante. Por otra parte, si la distribución a priori π_k es una distribución $N(0,\ k^2)$, aplicando el Ejemplo 2.3. deducimos que el estimador Bayes es:

$$d_k = \frac{nk^2\overline{\mathbf{x}}}{\sigma^2 + nk^2}$$

Observamos que,

$$\begin{aligned}
R_B(\pi_k,\ d_k) &= \int_{-\infty}^{+\infty} R(\theta,\ d_k)\pi_k(\theta)\,d\theta \\
&= \int_{-\infty}^{+\infty} \int_{\mathbb{R}^n} \left[(d_k(\mathbf{x}) - \theta)^2 f(\mathbf{x}|\theta)\right] \pi_k(\theta)\,d\mathbf{x}\,d\theta \\
&= \int_{-\infty}^{+\infty} E_\theta \left(\frac{nk^2\overline{X}}{\sigma^2 + nk^2} - \theta\right)^2 \pi_k(\theta)\,d\theta \\
&= \int_{-\infty}^{+\infty} \frac{nk^4\sigma^2 + \theta^2\sigma^4}{(\sigma^2 + nk^2)^2}\pi_k(\theta)\,d\theta \\
&= \int_{-\infty}^{+\infty} \frac{nk^4\sigma^2 + \theta^2\sigma^4}{(\sigma^2 + nk^2)^2} \frac{1}{k\sqrt{2\pi}} e^{-\frac{\theta^2}{2k^2}}\,d\theta \\
&= \frac{\sigma^2 k^2}{\sigma^2 + nk^2},
\end{aligned}$$

que converge a σ^2/n cuando k tiende a infinito. Por consiguiente, $\overline{X}$ es minimax.

3.2. Test de Hipótesis

Para encontrar una solución minimax al problema del test de hipótesis $H_0 : \theta \in \Theta_0$ frente a $H_1 : \theta \in \Theta_1$ usando la función de pérdida (2.1) debemos encontrar una regla d que minimice $sup_{\theta\in\Theta}\, R(\theta,\ d)$, es decir, encontrar d que minimice:

$$max_\theta\ [C_{II}P_\theta(d(\mathbf{X}) = a_0)1_{\{\theta\in\Theta_1\}},\ C_I P_\theta(d(\mathbf{X}) = a_1)1_{\{\theta\in\Theta_0\}}], \tag{3.3}$$

a diferencia de una soulción de Bayes que requiere encontrar una regla d que minimice $R_B(\pi,\ d) = E_\pi[R(\theta,\ d)]$, donde π es la distribución a priori en Θ.

Teorema 3.12. *Una regla minimax para contrastar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$, usando la función de pérdida (2.1) viene dada por la región de rechazo $\{\mathbf{x} \in \mathscr{X} : \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} \geq K$, donde la constante K verifica*

$$C_{II}P_{\theta_1}\left(\frac{f(\mathbf{X}|\theta_1)}{f(\mathbf{X}|\theta_0)} < K\right) = C_I P_{\theta_0}\left(\frac{f(\mathbf{X}|\theta_1)}{f(\mathbf{X}|\theta_0)} \geq K\right)$$

Demostración: Notemos que debemos encontrar d que minimice (3.3). Sea d la regla del enunciado y notemos que $R(\theta_0,\ d) = R(\theta_1,\ d)$. Supongamos que no es minimax, esto es, que existe otra regla d^* tal que

$$sup_{\theta\in\Theta}\, R(\theta,\ d^*) < R(\theta_0,\ d) = R(\theta_1,\ d)$$

Es decir, $R(\theta_0,\ d^*) < R(\theta_0,\ d)$ y $R(\theta_1,\ d^*) < R(\theta_1,\ d)$. La primera desigualdad implica que el tamaño del test d^* es menor que el de d, por lo que, por el lema de Neyman-Pearson, la potencia de d debe ser al menos la de d^*. Por tanto, $R(\theta_1,\ d) \leq R(\theta_1,\ d^*)$ lo que es una contradicción y el resultado está probado. □

Ejemplo 3.13. Sean $X_1,\ X_2,\ ...,\ X_n$ variables aleatorias iid de distribución $N(\mu,\ 1)$. Para el test H_0 : $\mu = \mu_0$ frente a $H_1 : \mu = \mu_1\ (> \mu_0)$, tendremos que elegir K tal que

$$\begin{aligned} C_{II}P_{\mu_1}(\frac{f(\mathbf{X}|\mu_1)}{f(\mathbf{X}|\mu_0)} < K) &= C_I P_{\mu_0}(\frac{f(\mathbf{X}|\mu_1)}{f(\mathbf{X}|\mu_0)} \geq K) \\ \Rightarrow C_{II}P_{\mu_1}(\overline{X} < K) &= C_I P_{\mu_0}(\overline{X} \geq K) \\ \Rightarrow C_{II}P_{\mu_1}(\frac{\overline{X}-\mu_1}{\frac{1}{\sqrt{n}}} < \frac{K-\mu_1}{\frac{1}{\sqrt{n}}}) &= C_I P_{\mu_0}(\frac{\overline{X}-\mu_0}{\frac{1}{\sqrt{n}}} \geq \frac{K-\mu_0}{\frac{1}{\sqrt{n}}}) \\ \Rightarrow C_{II}\ \Phi(\sqrt{n}(K-\mu_1)) &= C_I\ \{1-\Phi(\sqrt{n}(K-\mu_0))\}, \end{aligned}$$

donde Φ es la función de distribución de una variable aleatoria $N(0,\ 1)$. Esto se resuelve numéricamente, una vez conozcamos $C_I,\ C_{II},\ \mu_0,\ \mu_1$ y n.

Capítulo 4

Admisibilidad y Completitud

Se puede dar un orden parcial llamado *dominancia* dentro de la clase D de estimadores que se estudia en términos de la función de riesgo.

Definición 4.1. Dado el conjunto D de reglas de decisión, se definen las siguientes relaciones:

- La regla de decisión d_1, se dice que es *tan buena como* la regla d_2, si $R(\theta,\ d_1) \leq R(\theta,\ d_2)$ para todo $\theta \in \Theta$.
- La regla de decisión d_1, se dice que es *mejor que* la regla d_2, si $R(\theta,\ d_1) \leq R(\theta,\ d_2)$ para todo $\theta \in \Theta$ y $R(\theta,\ d_1) < R(\theta,\ d_2)$ para al menos un $\theta \in \Theta$.
- La regla de decisión d_1, se dice que es *equivalente a* la regla d_2, si $R(\theta,\ d_1) = R(\theta,\ d_2)$ para todo $\theta \in \Theta$.

Definición 4.2. Se dice que una regla de decisión d es *admisible* si no existe una regla mejor que d. Se dice que una regla es *inadmisible* si no es admisible.

Es importante reflexionar sobre el concepto de admisibilidad. Sobre todo, entender que no puede tomarse como una garantía de optimo sino como la ausencia de una propiedad negativa, ya que no queremos utilizar una regla sabiendo que hay otra mejor que ella. En el Ejemplo 1.5. observamos que el estimador $\hat{S}^2 = \frac{n-1}{n+1}\ S^2$ tiene el menor riesgo entre todos los estimadores de la forma bS^2. Luego $\hat{S}^2$ es mejor estimador que S^2, y por lo tanto S^2 es un estimador inadmisible.

Parece lógico restringir la búsqueda de una regla a algunas clases que contengan todas las reglas admisibles. Para ello vamos a introducir el concepto de completitud.

Definición 4.3. Se dice que una clase $C \subset D$ de reglas de decisión es *completa*, si para toda $d \in D$ tal que $d \notin C$, existe una regla $d_0 \in C$ que es mejor que d.

Se dice que una clase $C \subset D$ de reglas de decisión es *esencialmente completa*, si para toda $d \in D$ tal que $d \notin C$, existe una regla $d_0 \in C$ que es tan buena como d.

A continuación vamos a ilustrar en los siguientes lemas la diferencia entre las clases completas y las esencialmente completas.

Lema 4.4. *Si C es una clase completa y A denota la clase de todas las reglas admisibles, entonces $A \subset C$.*

Demostración: Inmediata aplicando las definiciones de clase completa y regla admisible.

Lema 4.5. *Si C es una clase esencialmente completa y existe una regla de decisión d admisible tal que $d \notin C$, entonces existe $d' \in C$ tal que d' es equivalente a d.*

Demostración: Como C es una clase esencialmente completa, dada una regla $d \in D$ tal que $d \notin C$, existe una regla $d' \in C$ tan buena como d. Es decir, $R(\theta,\ d') \leq R(\theta,\ d)$ para todo $\theta \in \Theta$ y no puede ocurrir $R(\theta,\ d') < R(\theta,\ d)$ para ningún θ porque entonces d' será mejor que d y esta no sería admisible. En conclusión, $R(\theta,\ d') = R(\theta,\ d)$ para todo $\theta \in \Theta$, es decir d' es equivalente a d.

Definición 4.6. Se dice que una clase C de reglas de decisión es *completa minimal*, si $C' \subset C$ con C' clase completa implica que $C = C'$.

Se dice que una clase C de reglas de decisión es *esencialmente completa minimal*, si $C' \subset C$ con C' clase esencialmente completa implica que $C = C'$.

Notemos que no es necesario que exista ni la clase completa minimal ni la clase esencialmente completa minimal. Si existe la clase (esencialmente) completa minimal, nos permite la máxima simplificación del problema. El siguiente teorema aclara la relación entre reglas admisibles y clases completas minimales.

Teorema 4.7. *Sea C una clase completa minimal y A el conjunto de las reglas admisibles. Se tiene $C = A$.*

Demostración: Como una clase completa minimal es a su vez una clase completa, por el Lema 4.4. sabemos que $A \subset C$. Nos falta probar que $C \subset A$, y lo haremos suponiendo que es falso y llegando a una contadicción. Sea $d \in C$ y $d \notin A$, por no ser d admisible existirá un d' mejor que d. Si $d' \notin C$ por ser C una clase completa existirá un $d^* \in C$ mejor que d'. Es decir, existe algún $\bar{d}$ mejor que d (ya sea d' o d^*) que está en C.

Veamos ahora que $C - \{d\}$ es una clase completa. Sea $d_1 \notin C - \{d\}$, como C es una clase completa existe $d_2 \in C$ tal que d_2 es mejor que d_1. Si $d_2 = d$, entonces $\bar{d} \in C - \{d\}$ es mejor que d y a su vez mejor que d_1. En cambio, si $d_2 \neq d$, entonces $d_2 \in C - \{d\}$ y mejor es que d_1. Luego $C - \{d\}$ es una clase completa y esto contradice que C sea una clase completa minimal.

□

4.1. Estimación puntual. Relación con las reglas de Bayes y Minimax

Veamos a continuación algunos teoremas que caracterizan la admisibilidad en las reglas de Bayes y minimax.

Teorema 4.8. *Sea $\Theta = \{\theta_1,\ \theta_2,\ ...,\ \theta_n\}$ y d un estimador de Bayes frente a la distribución a priori $(\pi_1,\ \pi_2,\ ...,\ \pi_n)$. Si $\pi_j > 0$ para $j = 1,\ 2,\ ...,\ n$ entonces d es admisible.*

Demostración: Supongamos que d no es admisible, entonces existirá un $d^* \in D$ mejor que d, es decir, $R(\theta_j,\ d^*) \leq R(\theta_j,\ d)$, para todo j y $R(\theta_j,\ d^*) < R(\theta_j,\ d)$, para algún j. Como $\pi_j > 0$ para $j = 1,\ 2,\ ...,\ n$, tenemos que

$$\sum_{j=i}^{n} \pi_j R(\theta_j,\ d^*) < \sum_{j=i}^{n} \pi_j R(\theta_j,\ d),$$

lo que contradice que d sea una regla de Bayes con respecto a $(\pi_1,\ \pi_2,\ ...,\ \pi_n)$.

□

Teorema 4.9. *Si para una distribución a priori π sobre Θ hay una única regla de Bayes d esta es admisible.*

Demostración: Supongamos que d no es admisible, entonces existirá un d' mejor que d, es decir, $R(\theta,\ d') \leq R(\theta,\ d)$, para todo $\theta \in \Theta$ y $R(\theta,\ d') < R(\theta,\ d)$, para algún $\theta \in \Theta$. Y esto implica que

$$\int_{\Theta} R(\theta,\ d')\pi(\theta)\,d\theta \leq \int_{\Theta} R(\theta,\ d)\pi(\theta)\,d\theta,$$

en contradicción con la hipótesis de que d es una regla de Bayes y es única.

□

Observación 4.10. Notemos que si se suprime la condición de unicidad del Teorema 4.9., no se deduce que d sea admisible.

Teorema 4.11. *Supongamos que $R(\theta,\ d)$ es una función continua de θ para cada d. Si d^* es una regla de Bayes frente a una distribución a priori π con soporte en Θ, entonces d^* es admisible.*

Demostración: Supongamos que d^* no es admisible, entonces existirá un d' mejor que d^*, es decir, $R(\theta,\ d') \leq R(\theta,\ d^*)$, para todo $\theta \in \Theta$ y $R(\theta,\ d') < R(\theta,\ d^*)$, para algún $\theta_0 \in \Theta$. Entonces existe $\varepsilon > 0$ de modo que

$$R(\theta,\ d') \leq R(\theta,\ d^*) - \varepsilon,$$

para todo $\theta \in B(\theta_0,\ \varepsilon)$, donde $B(\theta_0,\ \varepsilon)$ es un entorno de θ_0. Se tiene así que

$$\begin{aligned} R_B(\pi,\ d') &= \int_{B(\theta_0,\ \varepsilon)} R(\theta,\ d')\pi(\theta)\,d\theta + \int_{\mathbb{R}-B(\theta_0,\ \varepsilon)} R(\theta,\ d')\pi(\theta)\,d\theta \\ &\leq \int_{\mathbb{R}} R(\theta,\ d^*)\pi(\theta)\,d\theta - \varepsilon \int_{B(\theta_0,\ \varepsilon)} \pi(\theta)\,d\theta \\ &= R_B(\pi,\ d^*) - \varepsilon \int_{B(\theta_0,\ \varepsilon)} \pi(\theta)\,d\theta \\ &< R_B(\pi,\ d^*) \end{aligned}$$

Como d^* es una regla de Bayes frente a π se llega a una contradición y, por lo tanto, d^* es admisible.

□

Teorema 4.12. *Si d^* es la única regla minimax entonces d^* es admisible.*

Demostración: Supongamos que d^* no es admisible, entonces existirá un d' mejor que d^*, es decir, $R(\theta,\ d') \leq R(\theta,\ d^*)$, para todo $\theta \in \Theta$ y $R(\theta,\ d') < R(\theta,\ d^*)$, para algún $\theta \in \Theta$. Por lo tanto,

$$sup_{\theta\in\Theta}\, R(\theta,\ d') \leq sup_{\theta\in\Theta}\, R(\theta,\ d^*)$$

Luego d' debería ser también minimax, lo que contradice que d^* es el único estimador minimax.

□

Teorema 4.13. *Si d^* es admisible y tiene riesgo constante sobre Θ, entonces d^* es minimax.*

Demostración: Supongamos que d^* no es minimax, entonces existirá un d' tal que

$$sup_{\theta\in\Theta}\, R(\theta,\ d') < sup_{\theta\in\Theta}\, R(\theta,\ d^*),$$

para algún $\theta \in \Theta$. Pero como el riesgo de d^* es constante, tenemos que:

$$sup_{\theta\in\Theta}\, R(\theta,\ d^*) = R(\theta,\ d^*),$$

para todo $\theta \in \Theta$. Luego $R(\theta,\ d') < R(\theta,\ d^*)$ para todo $\theta \in \Theta$ y por lo tanto d' es mejor que d^*, y entonces d^* no es admisible, lo que nos lleva a una contradicción.

□

Ejemplo 4.14. Sea $(X_1,\ X_2,\ ...,\ X_n)$ una muestra aleatoria de una distribución $N(\theta,\ 1)$, veamos que $\overline{X}$ es un estimador admisible de θ con la función de pérdida cuadrática. Sabemos que $\overline{X} \sim N(\theta,\ \frac{1}{n})$.

Supongamos que $\overline{X}$ no es admisible, luego existirá una regla d^* mejor que $\overline{X}$, es decir, $R(\theta,\ d^*) \leq R(\theta,\ \overline{X})$, para todo $\theta \in \Theta$ y $R(\theta_0,\ d^*) < R(\theta_0,\ \overline{X})$, para algún $\theta_0 \in \Theta$. Como el riesgo es una función continua de θ, existe $\varepsilon,\ \delta > 0$ tales que

$$R(\theta, d^*) < R(\theta, \bar{X}) - \varepsilon,$$

para $|\theta - \theta_0| < \delta$. Si consideramos como distribución a priori una $N(0, \tau^2)$, por el Ejemplo 2.3. sabemos que el estimador de Bayes es

$$d(\mathbf{x}) = \frac{\tau^2}{\tau^2 + \frac{1}{n}}\bar{x}$$

Con riesgo Bayes

$$R_B(\pi,\, d) = \frac{\frac{\tau^2}{n}}{\tau^2 + \frac{1}{n}} = \frac{1}{n}\left[\frac{n\tau^2}{n\tau^2+1}\right]$$

Por lo tanto,

$$R_B(\pi,\, \bar{x}) - R_B(\pi,\, d) = \frac{1}{n} - \frac{1}{n}\left(\frac{n\tau^2}{n\tau^2+1}\right) = \frac{1}{n(n\tau^2+1)}$$

$$\begin{aligned}\tau\,[R_B(\pi,\, d^*) - R_B(\pi,\, \bar{x})] &= \tau \int_\Theta [R(\theta,\, d^*) - R(\theta,\, \bar{x})] \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{\theta^2}{2\tau^2}}\, d\theta \\ &< -\frac{\varepsilon}{\sqrt{2\pi}} \int_{\theta_0-\delta}^{\theta_0+\delta} e^{-\frac{\theta^2}{2\tau^2}}\, d\theta\end{aligned}$$

Y así,

$$\begin{aligned}0 &\leq \tau\,[R_B(\pi,\, d^*) - R_B(\pi,\, d)] \\ &= \tau\,[R_B(\pi,\, d^*) - R_B(\pi,\, \bar{x})] + \tau\,[R_B(\pi,\, \bar{x}) - R_B(\pi,\, d)] \\ &< -\frac{\varepsilon}{\sqrt{2\pi}} \int_{\theta_0-\delta}^{\theta_0+\delta} e^{-\frac{\theta^2}{2\tau^2}}\, d\theta + \frac{\tau}{n(n\tau^2+1)} \to_{\tau\to\infty} -\frac{\varepsilon}{\sqrt{2\pi}} \int_{\theta_0-\delta}^{\theta_0+\delta} 1\, d\theta = -\frac{2\varepsilon\delta}{\sqrt{2\pi}}.\end{aligned}$$

Esto es imposible, luego $\overline{X}$ es un estimador admisible de la media de la distribución $N(\theta, 1)$. Además, como $\overline{X}$ tiene riesgo constante ($\frac{1}{n}$ para todo $\theta \in \Theta$), entonces por el Teorema 4.13. concluimos que $\overline{X}$ es minimax. Por lo tanto, $\overline{X}$ es un estimador admisible y minimax de θ.

En el Teorema 3.6. vimos que una regla d de Bayes con riesgo constante es minimax, y en el Teorema 4.13. hemos visto que una regla d admisible con riesgo constante también es minimax. A continuación veremos un resultado para que un estimador d minimax tenga riesgo constante.

Teorema 4.15. *Supongamos que la función de riesgo $R(\theta,\, d)$ es continua en θ y que los valores inferior ($\underline{V}$) y superior ($\overline{V}$) definidos en la Observación 3.3. coinciden, $\underline{V} = \overline{V} = V$. Entonces, si d^* es minimax y π^* una distribución menos favorable, se tiene*

$$R(\theta,\, d^*) = V$$

para todo θ en el soporte de π^*.

Demostración: Como d^* es minimax, se tiene que

$$R(\theta,\, d^*) \leq V,$$

para todo $\theta \in \Theta$. Supongamos que existe $\theta_0 \in \Theta$ tal que $R(\theta_0,\, d^*) < V$ y θ_0 está en el soporte de π^*. Como $R(\theta,\, d^*)$ es continua en θ_0, habrá un entorno $B(\theta_0)$ dentro del cual

$$R(\theta,\, d^*) < V,$$

para todo $\theta \in B(\theta_0)$ y como θ_0 está en el soporte de π^*, entonces:

$$\int_{B(\theta_0)} \pi(\theta)\, d\theta > 0$$

Ahora bien

$$V = R_B(\pi^*,\, d^*) = \int_\Theta R(\theta,\, d^*)\pi^*(\theta)\, d\theta < V \int_{\Theta - B(\theta_0)} \pi^*(\theta)\, d\theta + V \int_{B(\theta_0)} \pi^*(\theta)\, d\theta = V$$

hemos llegado a una contradicción.

□

Notemos que tanto en el Teorema 4.11. como en el Teorema 4.15. se impone la hipótesis de que la función de riesgo $R(\theta,\ d)$ sea continua en θ para cada estimador d, en un caso para probar la admisibilidad de un estimador de Bayes, y en el otro para probar que el riesgo de un estimador minimax es constante.

Un ejemplo de clase esencialmente completa lo da el teorema de Rao-Blackwell que anunciamos a continuación.

Teorema 4.16. ***Rao-Blakwell.***

Sea $g : \mathbb{R} \to \mathbb{R}$ una función convexa y T un estadístico suficiente unidimensional para $\theta \in \Theta \subseteq \mathbb{R}$. Para cualquier estadístico suficiente unidimensional T' con esperanza finita, se tiene

$$E_\theta g(T' - E_\theta(T')) \geq E_\theta g(T'' - E_\theta(T''))$$

con $T'' = E_\theta(T'|T)$. En particular, la clase de estimadores basados en T es esencialmente completa dentro de los estimadores insesgados de θ cuando se usa la función de pérdida cuadrática.

4.2. Test de Hipótesis

En el caso del test con H_0 y H_1 simples y la función de pérdida (2.1), el Lema de Neyman-Pearson, que afirma que el test basado en el cociente de verosimilitudes es el uniformemente más potente de su tamaño nos proporciona una clase completa.

Teorema 4.17. ***Neyman-Pearson.***

Todo tests de la forma

$$\phi(\mathbf{x}) = \begin{cases} 1 & si \quad f(\mathbf{x}|\theta_1) > Kf(\mathbf{x}|\theta_0) \\ \gamma(\mathbf{x}) & si \quad f(\mathbf{x}|\theta_1) = Kf(\mathbf{x}|\theta_0) \\ 0 & si \quad f(\mathbf{x}|\theta_1) < Kf(\mathbf{x}|\theta_0) \end{cases}$$

para cualquier $K \geq 0$ y cualquier $0 \leq \gamma(\mathbf{x}) \leq 1$, es el test más potente de su tamaño para contrastar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$.

Si $K = \infty$, el test:

$$\phi(\mathbf{x}) = \begin{cases} 1 & si \quad f(\mathbf{x}|\theta_0) = 0 \\ 0 & si \quad f(\mathbf{x}|\theta_0) > 0 \end{cases}$$

es el más potente para contrastar H_0 frente a H_1.

Estos dos test forma una clase completa minimal de reglas de decisión. La subclase de tales tests con $\gamma(\mathbf{x}) \equiv \gamma$ (una constante) es una clase esencialmente completa.

Ejemplo 4.18. Sean $X1,\ X2,\ ...,\ Xn$ variables aleatorias iid con distribución $Bernoulli(\theta)$. Sea $Y = \sum_{i=1}^{n} Xi \sim Bin(n, \theta)$ con una muestra de tamaño 1 de Y. Queremos hacer el test $H_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$, donde $\theta_0 < \theta_1$. Notar que $f(y|\theta)$ tiene una razón de verosimilitud monótona, es decir, si $\theta < \theta'$ la razón de verosimilitud $f(y|\theta')/f(y|\theta)$ es una función no decreciente de x. Por lo tanto el test puede ser reescrito así

$$\phi(y) = \begin{cases} 1 & si \quad y > j \\ \gamma & si \quad y = j \\ 0 & si \quad y < j \end{cases} \tag{4.1}$$

donde j es algún entero (que depende de $C_I,\ C_{II},\ \pi_0\ \pi_1$), puesto que ϕ se puede escribir así

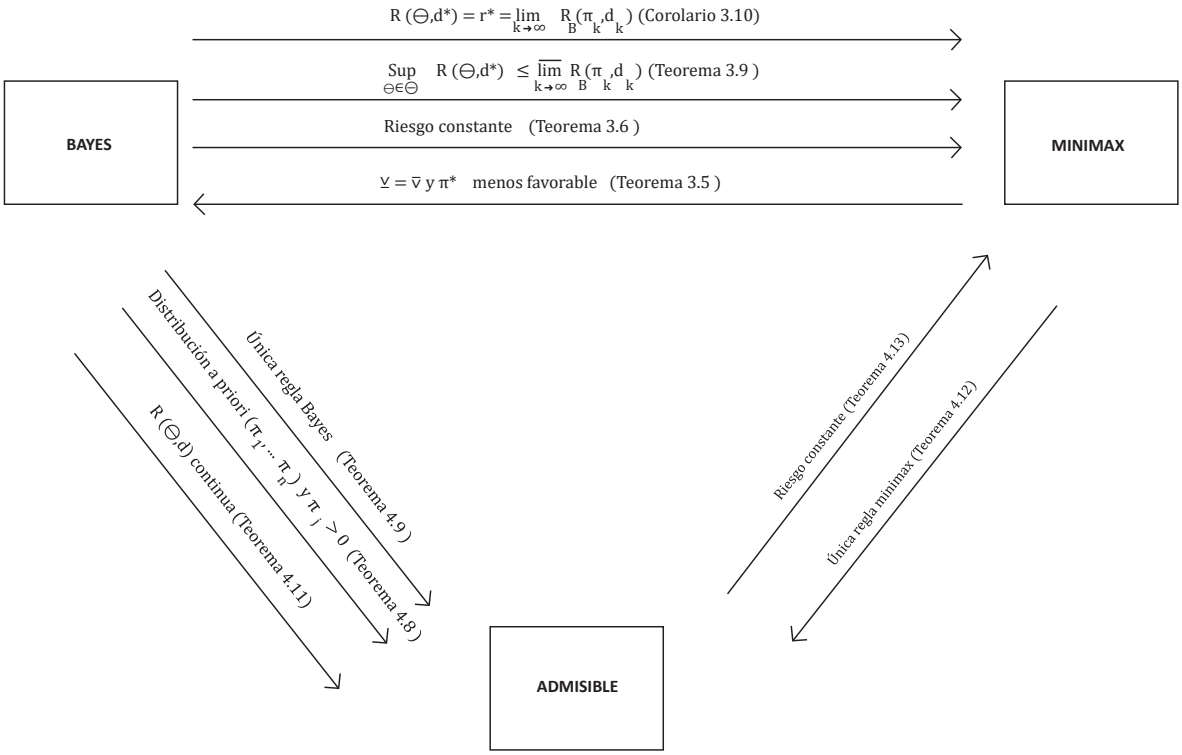
$$\phi(y) = \begin{cases} 1 & si \quad C_{II}\pi_1 f(y|\theta_1) > C_I\pi_0 f(y|\theta_0) \\ \gamma & si \quad C_{II}\pi_1 f(y|\theta_1) = C_I\pi_0 f(y|\theta_0) \\ 0 & si \quad C_{II}\pi_1 f(y|\theta_1) < C_I\pi_0 f(y|\theta_0) \end{cases}$$

Para un test de esta forma (4.1),

$$\begin{aligned} \beta(\theta) &= P_{\theta_0}(Y > j) + \gamma P_{\theta_0}(Y = j) \\ 1 - \beta(\theta) &= P_{\theta_1}(Y < j) + (1-\gamma)P_{\theta_1}(Y = j). \end{aligned}$$

Buscamos j y γ tales que $C_I\beta(\theta) = C_{II}(1-\beta(\theta))$.

Para finalizar, tenemos un esquema en el que se relacionan los conceptos de regla Bayes, minimax y admisibilidad que se han tratado a lo largo del trabajo.



Bibliografía

[1] GEORGE CASELLA Y ROGER L. BERGER, *Statistical Inference*, International Student Edition.

[2] VIJAY K. ROHATGI Y A. K. MD. EHSANES SALEH, *An Introduction to Probability and Statistics*, Wiley series in probability and Statistics.

[3] THOMAS S. FERGUSON, *Mathematical Statistics. A Decision Theoretic Approach*, a Series of Monographs and Textbooks, New York 1967.

[4] JOSÉ ANTONIO CRISTÓBAL CRISTÓBAL, *Inferencia Estadística*, Prensas Universitarias de Zaragoza, 2000.

[5] JAMES O. BERGER, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag New York, Inc. 1980.