Capítulo 1: Introducción y Preparaciones

Teoría Estadística Avanzada

Sigla DES124

Prof. Jaime Lincovil

1.1. Introducción

El curso Teoría Estadística Avanzada (DES124) busca entrgar una primera aproximación a la teoría estadística avanzada. Estos apuntes definen la notación que utilizaremos a lo largo del curso y también establecen el nivel conceptual y matemático en el que trabajaremos. Naturalmente, tanto el nivel conceptual como el matemático serán acorde el curso de Probabilidad para el doctorado.

El análisis real y, en particular, la teoría de la medida, son muy importantes en la probabilidad y la estadística. De hecho, la teoría de la medida es la base sobre la cual se construye la probabilidad moderna y, debido a la estrecha conexión entre la probabilidad y la estadística, es natural que la teoría de la medida también impregne la literatura estadística. La teoría de la medida en sí misma puede ser muy abstracta y difícil. En este curso no buscamos convertirnos en expertos en teoria de la medida. Sin embargo, en general, para leer y comprender artículos de investigación en teoría estadística, al menos se debe estar familiarizado con la terminología y los resultados básicos de la teoría de la medida. Mi presentación aquí tiene como objetivo introducir estos conceptos básicos, de modo que tengamos un vocabulario funcional en teoría de la medida para avanzar hacia nuestro enfoque principal en el curso. Además de la teoría de la medida, también proporcionaré una breve introducción a la teoría de grupos y a los conjuntos/funciones convexas. El resto de este primer conjunto de apuntes trata sobre las transiciones de la teoría de la medida a la probabilidad y de la probabilidad a la estadística.

Desde el punto de vista conceptual, además de poder aplicar la teoría a ejemplos particulares, espero comunicar **por qué** se desarrolló dicha teoría; es decir, no solo quiero que estés familiarizado con los resultados y las técnicas, sino que también espero que puedas comprender la motivación detrás de estos desarrollos. A lo largo de esta línea, en este capítulo, discutiré los elementos básicos de un problema de inferencia estadística, junto con algunas reflexiones sobre el razonamiento estadístico, abordando la pregunta fundamental: **¿cómo razonar de una muestra a una población?** Sorprendentemente, **no hay una respuesta completamente satisfactoria a esta cuestión**.

1.2. Preliminares Matemáticos

1.2.1. Medida e Integración

La teoría de la medida es la base sobre la cual se construye la teoría moderna de la probabilidad. Todos los estadísticos deberían, al menos, estar familiarizados con la terminología y los resultados clave (por ejemplo, el teorema de convergencia dominada de Lebesgue). La presentación a continuación está basada en el material de Lehmann y Casella (1998); presentaciones similares se encuentran en Keener (2010).

Una medida es una generalización del concepto de longitud, área, volumen, etc. Más específicamente, una medida μ es una función de conjuntos no negativa, es decir, μ asigna un número no negativo a los subconjuntos A de un conjunto abstracto $\mathbb{X}$, y este número se denota por $\mu(A)$. De manera similar a las longitudes, μ se asume como aditiva:

$$\mu(A \cup B) = \mu(A) + \mu(B), \quad \text{para cada } A \text{ y } B \text{ disjuntos.}$$

Esto se extiende por inducción a cualquier conjunto finito $A_1, \ldots, A_n$ de conjuntos disjuntos. Pero una suposición más fuerte es la **σ-aditividad**:

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i), \quad \text{para todos los } A_1, A_2, \ldots \text{ disjuntos.}$$

Nótese que la aditividad finita no implica σ-aditividad. Todas las medidas (de probabilidad) con las que estamos familiarizados son σ-aditivas. Sin embargo, existen algunas medidas peculiares que son finitamente aditivas pero no σ-aditivas. El ejemplo clásico de esto es el siguiente.

Ejemplo 1.1:
Tomemos $\mathbb{X} = \{1, 2, \ldots\}$ y definamos una medida μ como:

$$\mu(A) = \begin{cases} 0, & \text{si } A \text{ es finito,} \\ 1, & \text{si } A \text{ es co-finito.} \end{cases}$$

Un conjunto A es **co-finito** si es el complemento de un conjunto finito. Es fácil ver que μ es aditiva. Tomando una sucesión disjunta $A_i = \{i\}$, encontramos que $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \mu(\mathbb{X}) = 1$, pero $\sum_{i=1}^{\infty} \mu(A_i) = \sum_{i=1}^{\infty} 0 = 0$. Por lo tanto, μ no es σ-aditiva.

En general, una medida μ no puede definirse para todos los subconjuntos $A \subseteq \mathbb{X}$. Pero la clase de subconjuntos en los que se puede definir la medida es, en general, una **σ-álgebra** o **σ-campo**.

Definición 1.1

Una σ-álgebra $\mathcal{A}$ es una colección de subconjuntos de $\mathbb{X}$ tal que:

- $\mathbb{X}$ está en $\mathcal{A}$;
- Si $A \in \mathcal{A}$, entonces su complemento $A^c \in \mathcal{A}$;
- Si $A_1, A_2, \cdots \in \mathcal{A}$, entonces $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Los conjuntos $A \in \mathcal{A}$ se denominan **medibles**. Nos referimos a $(\mathbb{X}, \mathcal{A})$ como un espacio medible. Si una medida μ está definida en $(\mathbb{X}, \mathcal{A})$, entonces $(\mathbb{X}, \mathcal{A}, \mu)$ es un **espacio de medida**.

Una medida μ es **finita** si $\mu(\mathbb{X})$ es un número finito. Las medidas de probabilidad son ejemplos especiales de medidas finitas donde $\mu(\mathbb{X}) = 1$. Se dice que una medida μ es σ-finita si existe una secuencia de conjuntos $\{A_i\} \subset \mathcal{A}$ tal que $\bigcup_{i=1}^{\infty} A_i = \mathbb{X}$ y $\mu(A_i) < \infty$ para cada i.

Ejemplo 1.2:

Sea $\mathbb{X}$ un conjunto numerable y $\mathcal{A}$ la clase de todos los subconjuntos de $\mathbb{X}$; entonces, claramente $\mathcal{A}$ es una σ-álgebra. Definimos μ de acuerdo con la regla:

$$\mu(A) = \text{número de puntos en } A, \quad A \in \mathcal{A}.$$

Entonces, μ es una medida σ-finita, la cual se conoce como **medida de conteo**.

Ejemplo 1.3:

Sea $\mathbb{X}$ un subconjunto del espacio euclidiano d-dimensional $\mathbb{R}^d$. Tomemos $\mathcal{A}$ como la σ-álgebra más pequeña que contiene la colección de rectángulos abiertos:

$$\mathcal{A} = \{(x_1, \ldots, x_d) : a_i < x_i < b_i, i = 1, \ldots, d\}, \quad a_i < b_i.$$

Entonces, $\mathcal{A}$ es la σ-álgebra de Borel en $\mathbb{X}$, la cual contiene todos los conjuntos abiertos y cerrados en $\mathbb{X}$; sin embargo, hay subconjuntos de $\mathbb{X}$ que no pertenecen a $\mathcal{A}$. La medida μ única, definida por:

$$\mu(A) = \prod_{i=1}^{d} (b_i - a_i), \quad \text{para rectángulos } A \in \mathcal{A},$$

se conoce como **medida de Lebesgue**, y es σ-finita.

A continuación, consideramos la integración de una función real f con respecto a una medida μ en $(\mathbb{X}, \mathcal{A})$. Esta definición más general de integral satisface la mayoría de las propiedades familiares del cálculo, tales como linealidad, monotonía, etc. Sin embargo, la integral de cálculo se define solo para una clase de funciones que generalmente es demasiado pequeña para nuestras aplicaciones.

La clase de funciones de interés son aquellas que son **medibles**. En particular, una función real f es medible si y solo si, para todo número real a, el conjunto $\{x : f(x) \leqslant a\}$ pertenece a $\mathcal{A}$. Si A es un conjunto medible, entonces la función indicadora $I_A(x)$, que vale 1 cuando

$x \in A$ y 0 en otro caso, es medible. Más generalmente, una función simple

$$s(x) = \sum_{k=1}^{K} a_k I_{A_k}(x),$$

es medible siempre que $A_1, \ldots, A_K \in \mathcal{A}$. Las funciones continuas f también suelen ser medibles.

La integral de una función simple no negativa s con respecto a μ se define como:

$$\int s \, d\mu = \sum_{k=1}^{K} a_k \mu(A_k). \tag{1}$$

Tomemos una sucesión no decreciente de funciones simples no negativas $\{s_n\}$ y definamos

$$f(x) = \lim_{n\to\infty} s_n(x). \tag{2}$$

Se puede demostrar que la función f definida en (2) es medible. Entonces, la integral de f con respecto a μ se define como:

$$\int f \, d\mu = \lim_{n\to\infty} \int s_n \, d\mu.$$

El límite de las integrales de funciones simples. Resulta que el lado izquierdo no depende de la secuencia particular $\{s_n\}$, por lo que es único. De hecho, una definición equivalente para la integral de una función no negativa f es:

$$\int f \, d\mu = \sup_{0 \leqslant s \leqslant f, \text{ simple}} \int s \, d\mu. \tag{3}$$

Para una función medible f que puede tomar valores negativos, definimos:

$$f^+(x) = \text{máx}\{f(x), 0\}, \quad f^-(x) = -\,\text{mín}\{f(x), 0\}.$$

Ambas partes, f^+ y f^-, son no negativas, y $f = f^+ - f^-$. La integral de f con respecto a μ se define como:

$$\int f \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu.$$

donde las dos integrales en el lado derecho están definidas a través de (3). En general, una función medible f se dice μ-integrable, o simplemente integrable, si $\int f^+ \, d\mu$ y $\int f^- \, d\mu$ son ambas finitas.

Ejemplo 1.4: Medida de Conteo

Si $\mathbb{X} = \{x_1, x_2, \dots\}$ y μ es la medida de conteo, entonces:

$$\int f \, d\mu = \sum_{i=1}^{\infty} f(x_i).$$

Ejemplo 1.5: Medida de Lebesgue

Si $\mathbb{X}$ es un espacio euclidiano y μ es la medida de Lebesgue, entonces $\int f \, d\mu$ existe y es igual a la integral de Riemann usual de f del cálculo siempre que esta última exista. Sin embargo, la integral de Lebesgue existe para funciones f que no son Riemann-integrables.

A continuación, presentamos algunos resultados importantes del análisis relacionados con las integrales. Los dos primeros tratan sobre la permutación de límites[1] e integración, lo cual es a menudo importante en problemas estadísticos. El primero es relativamente débil, pero se usa en la demostración del segundo.

Teorema 1.1 Lema de Fatou

Dada una sucesión de funciones $\{f_n\}$, no negativas y medibles:

$$\int \left(\liminf_{n\to\infty} f_n \right) d\mu \leqslant \liminf_{n\to\infty} \int f_n \, d\mu.$$

La desigualdad opuesta se cumple para lím sup, siempre que $|f_n| \leqslant g$ para alguna función g integrable.

Teorema 1.2 Convergencia dominada

$$f(x) = \lim_{n\to\infty} f_n(x) \quad \mu\text{-almost everywhere},$$

y que $|f_n(x)| \leqslant g(x)$ para todo n, para todo x, y para alguna función g integrable. Entonces, f_n y f son integrables, y

$$\int f \, d\mu = \lim_{n\to\infty} \int f_n \, d\mu.$$

Demostración. Primero, por definición de f como el límite puntual de f_n, tenemos que $|f_n - f| \leqslant |f_n| + |f| \leqslant 2g$ y que $\lim\sup_n |f_n - f| = 0$.
De un ejercicio previo, obtenemos

$$\left| \int f_n \, d\mu - \int f \, d\mu \right| = \left| \int (f_n - f) \, d\mu \right| \leqslant \int |f_n - f| \, d\mu.$$

Para la cota superior, por el "lema inverso de Fatou", tenemos

[1] Recordemos las nociones de **"lim sup"** y **"lim inf"** del análisis. Por ejemplo, si x_n es una sucesión de números reales, entonces: $\lim\sup_{n\to\infty} x_n = \inf_n \sup_{k\geqslant n} x_k$ intuitivamente, este es el mayor punto de acumulación de la sucesión. De manera similar: $\lim\inf_{n\to\infty} x_n$ es el menor punto de acumulación, y si el mayor y el menor de los puntos de acumulación son iguales, entonces la sucesión converge y el punto de acumulación común es el límite. Además, si f_n es una sucesión de funciones de valores reales, entonces podemos definir $\lim\sup f_n$ y $\lim\inf f_n$ aplicando las definiciones anteriores punto por punto.

$$\limsup_n \int |f_n - f|\, d\mu \leqslant \int \limsup_n |f_n - f|\, d\mu = 0.$$

Por lo tanto,

$$\int f_n\, d\mu \to \int f\, d\mu,$$

lo que completa la demostración. ■

La frase μ-**almost everywhere** usada en el teorema significa que la propiedad se cumple en todos los puntos excepto en un conjunto nulo N, es decir, un conjunto N con $\mu(N) = 0$. Estos conjuntos de medida cero son conjuntos "pequeños" en un sentido de la teoría de la medida, en contraste con los conjuntos de primera categoría que son pequeños en un sentido topológico. En términos generales, los conjuntos de medida cero pueden ignorarse en integración y ciertos tipos de límites, pero siempre se debe ser cuidadoso.

El siguiente teorema es útil para acotar integrales de productos de dos funciones. Puede que estés familiarizado con este nombre de otros cursos, como álgebra lineal. De hecho, ciertas colecciones de funciones integrables se comportan de manera muy similar a los vectores en un espacio vectorial de dimensión finita.

Teorema 1.3 Desigualdad de Cauchy-Schwarz

$$\left(\int fg\, d\mu\right)^2 \leqslant \int f^2\, d\mu \cdot \int g^2\, d\mu.$$

Demostración. Si f^2 o g^2 no son integrables, entonces la desigualdad es trivial; así que asumamos que tanto f^2 como g^2 son integrables. Tomemos cualquier λ; entonces

$$\int (f + \lambda g)^2\, d\mu \geqslant 0.$$

En particular,

$$\int g^2\, d\mu \cdot \lambda^2 + 2\int fg\, d\mu \cdot \lambda + \int f^2\, d\mu \geqslant 0, \quad \forall \lambda.$$

En otras palabras, el polinomio cuadrático (en λ) puede tener a lo sumo una raíz real. Usando la fórmula cuadrática,

$$\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

es claro que la única manera en que puede haber menos de dos raíces reales es si $b^2 - 4ac \leqslant 0$. Operando nos encontramos que

$$4\left(\int fg\, d\mu\right)^2 - 4\int f^2\, d\mu \cdot \int g^2\, d\mu \leqslant 0,$$

y de aquí el resultado se sigue inmediatamente. Una demostración diferente, basada en la desigualdad de Jensen, se presenta en el Ejemplo 8 ■

El siguiente resultado define *integrales dobles* y muestra que, bajo ciertas condiciones, el orden de integración no importa. Sin entrar demasiado en detalles, para dos espacios de medida $(\mathbb{X}, \mathcal{A}, \mu)$ y $(\mathbb{Y}, \mathcal{B}, \nu)$, definimos el espacio producto

$$(\mathbb{X} \times \mathbb{Y}, \mathcal{A} \otimes \mathcal{B}, \mu \times \nu),$$

donde $\mathbb{X}\times\mathbb{Y}$ es el conjunto usual de pares ordenados (x, y), $\mathcal{A}\otimes\mathcal{B}$ es la σ-álgebra más pequeña que contiene todos los conjuntos $A \times B$ para $A \in \mathcal{A}$ y $B \in \mathcal{B}$, y la medida producto $\mu \times \nu$ se define como

$$(\mu \times \nu)(A \times B) = \mu(A)\nu(B).$$

Este concepto es importante porque las distribuciones de probabilidad independientes inducen una medida producto. El teorema de Fubini es un poderoso resultado que permite calcular ciertos integrales sobre un producto de manera unidimensional.

Teorema 1.4 Fubini

$$\int_{\mathbb{X}} \left[\int_{\mathbb{Y}} f(x, y)\, d\nu(y) \right] d\mu(x) = \int_{\mathbb{Y}} \left[\int_{\mathbb{X}} f(x, y)\, d\mu(x) \right] d\nu(y). \tag{4}$$

El valor común anterior es la *integral doble*, escrita como $\int_{\mathbb{X}\times\mathbb{Y}} f\, d(\mu \times \nu)$.
Nuestro último resultado trata sobre la construcción de nuevas medidas a partir de otras. También nos permite generalizar la noción familiar de densidades de probabilidad, lo que facilita la discusión sobre el problema general de inferencia estadística. Supongamos que f es una función medible no negativa[a]. Entonces,

$$\nu(A) = \int_A f\, d\mu \tag{5}$$

define una nueva medida ν en $(\mathbb{X}, \mathcal{A})$. Una propiedad importante es que si $\mu(A) = 0$ implica que $\nu(A) = 0$; la terminología para esto es que ν es *absolutamente continua* con respecto a μ, o que ν está dominada por μ, y se escribe $\nu \ll \mu$. Pero resulta que, si $\nu \ll \mu$, entonces existe f tal que (5) se cumple. Este es el famoso *teorema de Radon-Nikodym.*

[a] f puede tomar valores negativos, pero entonces la medida es una *medida con signo.*

Teorema 1.5 Radon-Nikodym

Suponga que $\nu \ll \mu$. Entonces, existe una función μ-integrable no negativa f, única módulo conjuntos μ-nulos, tal que (5) se cumple. La función f, a menudo escrita como

$$f = \frac{d\nu}{d\mu}$$

es la derivada de Radon-Nikodym de ν con respecto a μ.

Veremos más adelante que, en problemas estadísticos, la derivada de Radon-Nikodym es la densidad familiar o, quizás, la razón de verosimilitud. El teorema de Radon-Nikodym también formaliza la idea del cambio de variable en integración. Por ejemplo, supongamos que μ y ν son medidas σ-finitas definidas en $\mathbb{X}$, de modo que $\nu \ll \mu$, por lo que existe una única derivada de Radon-Nikodym $f = d\nu/d\mu$. Entonces, para una función ν-integrable φ, tenemos

$$\int \varphi\, d\nu = \int \varphi f\, d\mu;$$

simbólicamente, esto tiene sentido como:

$$d\nu = \left(\frac{d\nu}{d\mu}\right) d\mu.$$

1.2.2. Teoría Básica de Grupos

Una definición muy importante para este curso es el de un *grupo*, un conjunto de elementos junto con una cierta operación que tiene una estructura particular. Nuestro interés particular está en los grupos de transformaciones y cómo interactúan con las distribuciones de probabilidad. Aquí establecemos algo de terminología básica y comprensión de los grupos. Un curso de álgebra abstracta cubriría estos conceptos, y mucho más.

Definición 1.2
Un *grupo* es **un conjunto** $\mathcal{G}$ junto con una **operación binaria** $\cdot$, tal que:

- (*clausura*) para cada $g_1, g_2 \in \mathcal{G}$, se cumple que $g_1 \cdot g_2 \in \mathcal{G}$;
- (*identidad*) existe un elemento $e \in \mathcal{G}$ tal que $e \cdot g = g$ para todo $g \in \mathcal{G}$;
- (*inverso*) para cada $g \in \mathcal{G}$, existe $g^{-1} \in \mathcal{G}$ tal que $g^{-1} \cdot g = e$;
- (*asociatividad*) para cada $g_1, g_2, g_3 \in \mathcal{G}$, se cumple que $g_1 \cdot (g_2 \cdot g_3) = (g_1 \cdot g_2) \cdot g_3$.

El elemento e se llama la **identidad**, y el elemento g^{-1} se llama el **inverso** de g. El grupo $\mathcal{G}$ se llama **abeliano**, o **conmutativo**, si $g_1 \cdot g_2 = g_2 \cdot g_1$ para todos $g_1, g_2 \in \mathcal{G}$.

Algunos ejemplos básicos de grupos incluyen $(\mathbb{Z}, +)$, $(\mathbb{R}, +)$ y $(\mathbb{R}\setminus\{0\}, \times)$; este último requiere que se elimine el origen porque el 0 no tiene inverso multiplicativo. Estos tres grupos son abelianos. El grupo lineal general de dimensión m, que consiste en todas las matrices no singulares de $m \times m$, es un grupo bajo la multiplicación de matrices; este no es un grupo abeliano. Algunas propiedades simples de los grupos se dan en el Ejercicio 10.

Nos interesamos principalmente en **grupos de transformaciones**. Sea $\mathbb{X}$ un espacio (por ejemplo, un espacio muestral) y consideremos una colección $\mathcal{G}$ de funciones g, que mapean $\mathbb{X}$ en sí mismo. Consideremos la operación $\circ$ de composición de funciones. El elemento identidad es la función $e(x) = x$ para todo $x \in \mathbb{X}$. Si requerimos que $(\mathcal{G}, \circ)$ sea un grupo con identidad e, entonces cada $g \in \mathcal{G}$ es una función inyectiva. Para ver esto, tomemos cualquier $g \in \mathcal{G}$ y $x_1, x_2 \in \mathbb{X}$ tales que $g(x_1) = g(x_2)$. La composición por g^{-1} da $e(x_1) = e(x_2)$ y, en consecuencia, $x_1 = x_2$; por lo tanto, g es inyectiva. Algunos ejemplos de grupos de transformaciones son:

- Para $\mathbb{X} = \mathbb{R}^m$, definimos el mapeo $g_c(x) = x + c$, donde c es un vector en $\mathbb{R}^m$. Entonces, $\mathcal{G} = \{g_c : c \in \mathbb{R}^m\}$ es un grupo abeliano de transformaciones.
- Para $\mathbb{X} = \mathbb{R}^m$, definimos el mapeo $g_c(x) = cx$, que representa una reescalación del vector x por una constante c. Entonces, $\mathcal{G} = \{g_c : c > 0\}$ es un grupo abeliano de transformaciones.

- Para $\mathbb{X} = \mathbb{R}^m$, definimos $g_{a,b}(x) = ax + b\mathbf{1}_m$, una combinación de un desplazamiento y un escalamiento de x. Entonces, $\mathcal{G} = \{g_{a,b} : a > 0, b \in \mathbb{R}\}$ es un grupo de transformaciones; no es abeliano.
- Para $\mathbb{X} = \mathbb{R}^m$, definimos $g_A(x) = Ax$, donde $A \in GL(m)$. Entonces, $\mathcal{G} = \{g_A : A \in GL(m)\}$ es un grupo de transformaciones; no es abeliano.
- Sea $\mathbb{X} = \{1, 2, \ldots, m\}$ y definimos $g_\pi(x) = (x_{\pi(1)}, \ldots, x_{\pi(m)})$, donde π es una permutación de los índices. Entonces, $\mathcal{G} = \{g_\pi : \text{permutaciones } \pi\}$ es un grupo de transformaciones; no es abeliano.

En la literatura sobre grupos de transformaciones, es común escribir gx en lugar de $g(x)$.

Para un grupo de transformaciones $\mathcal{G}$ en $\mathbb{X}$, existen algunas clases de funciones de interés. Una función α, que mapea $\mathbb{X}$ en sí mismo, se dice *invariante* (con respecto a $\mathcal{G}$) si

$$\alpha(gx) = \alpha(x) \quad \text{para todo } x \in \mathbb{X} \text{ y todo } g \in \mathcal{G}.$$

Una función β, que mapea $\mathbb{X}$ en sí mismo, es *equivariante* (con respecto a $\mathcal{G}$) si

$$\beta(gx) = g\beta(x) \quad \text{para todo } x \in \mathbb{X} \text{ y todo } g \in \mathcal{G}.$$

La idea es que α no es sensible a los cambios inducidos por la aplicación $x \mapsto gx$ para $g \in \mathcal{G}$, mientras que β conserva la estructura de la transformación g aplicada antes o después.

Ejemplo 1.6:
Sea $\mathbb{X} = \mathbb{R}^m$ y definimos $g_c(x) = x + c\mathbf{1}_m$, que representa desplazamientos de localización. La función $\beta(x) = \bar{x}\mathbf{1}_m$ es equivariante con respecto a $\mathcal{G}$, donde $\bar{x}$ es el promedio de las entradas de x. La función $\alpha(x) = x - \bar{x}\mathbf{1}_m$ es invariante con respecto a $\mathcal{G}$.

Un caso ligeramente diferente de invarianza con respecto a un grupo de transformaciones, en un contexto relevante para la estadística y análisis de probabilidad, será considerado mas adelante.

1.2.3. Conjuntos y Funciones Convexas

Existe una propiedad especial que pueden tener las funciones y de la cual ocasionalmente tomaremos ventaja más adelante. Esta propiedad se llama *convexidad*. A lo largo de esta sección, a menos que se indique lo contrario, tomemos $f(x)$ como una función de valores reales definida sobre un espacio euclidiano p-dimensional $\mathbb{X}$. Se dice que la función f es convexa en $\mathbb{X}$ si, para cualquier $x, y \in \mathbb{X}$ y cualquier $\alpha \in [0, 1]$, se cumple la siguiente desigualdad:

$$f(\alpha x + (1 - \alpha)y) \leqslant \alpha f(x) + (1 - \alpha)f(y).$$

Para el caso $p = 1$, esta propiedad es fácil de visualizar. Ejemplos de funciones convexas (univariadas) incluyen e^x, $-\log x$, y x^r para $r > 1$.

En el caso en que f sea dos veces diferenciable, existe una caracterización alternativa de convexidad. Esto es algo que se cubre en la mayoría de los cursos intermedios de cálculo.

Proposición 1.1

Una función f dos veces diferenciable, definida en un espacio p-dimensional, es convexa si y solo si

$$\nabla^2 f(x) = \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right)_{i,j=1,\ldots,p}$$

es una matriz de segundas derivadas que es *semidefinida positiva* para cada x.

La convexidad es importante en problemas de optimización (máxima verosimilitud, mínimos cuadrados, etc.) ya que está relacionada con la existencia y unicidad de óptimos globales. Por ejemplo, si la función criterio (de pérdida) a minimizar es convexa y existe un mínimo local, entonces la convexidad garantiza que dicho mínimo es global.

El término *convexo* puede usarse como adjetivo para conjuntos, no solo para funciones. Un conjunto C, en un espacio lineal, es convexo si, para cualquier par de puntos x y y en C, la combinación convexa

$$ax + (1-a)y, \quad \text{para } a \in [0,1],$$

también pertenece a C. En otras palabras, un conjunto convexo C contiene los segmentos de línea que conectan todos los pares de puntos en C. Ejemplos de conjuntos convexos incluyen intervalos de números, círculos en el plano y bolas/elipses en dimensiones superiores.

Existe una conexión entre conjuntos convexos y funciones convexas: si f es una función convexa de valores reales, entonces, para cualquier t real, el conjunto

$$C_t = \{x : f(x) \leqslant t\}$$

es convexo. Habrá algunas aplicaciones de conjuntos convexos en los capítulos posteriores[2].

1.3. Probabilidad

1.3.1. Formulación en Teoría de la Medida

Resulta que la probabilidad matemática es solo un caso especial de la teoría de la medida presentada anteriormente. Nuestras probabilidades son medidas finitas, nuestras variables aleatorias son funciones medibles y los valores esperados son integrales.

Comenzamos con un espacio medible esencialmente arbitrario $(\Omega, \mathcal{F})$ e introducimos una

[2]Por ejemplo, el espacio de parámetros para las familias exponenciales naturales es convexo; el lema de Anderson, que se utiliza para demostrar minimaxidad en problemas de media normal, entre otras cosas, involucra conjuntos convexos, etc.

medida de probabilidad P; es decir, $P(\Omega) = 1$. Entonces, $(\Omega, \mathcal{F}, P)$ se llama un *espacio de probabilidad*. La idea es que Ω contiene todos los posibles resultados del experimento aleatorio.

Consideremos, por ejemplo, el caso de las alturas en el ejemplo de la Sección 4. Supongamos que planeamos seleccionar un único estudiante de la UNI al azar de la población de estudiantes. Entonces, Ω está compuesto por todos los estudiantes, y exactamente uno de ellos será el que se observe. La medida P codificará el esquema de muestreo subyacente. Sin embargo, en este ejemplo, no nos interesa qué estudiante en particular ha sido elegido, sino su altura, que es una medición o característica del estudiante seleccionado. ¿Cómo representamos esto matemáticamente?

Una variable aleatoria X no es más que una función medible de Ω a otro espacio $\mathbb{X}$. Es importante entender que X, como mapeo, no es aleatorio; en su lugar, X es una función de un elemento ω seleccionado aleatoriamente en Ω. Por lo tanto, cuando discutimos probabilidades de que X cumpla ciertas propiedades, en realidad estamos considerando la probabilidad (o medida) del conjunto de ω en los cuales $X(\omega)$ satisface la propiedad dada.

Para hacer esto más preciso, escribimos:

$$P(X \in A) = P\{\omega : X(\omega) \in A\} = PX^{-1}(A).$$

Para simplificar la notación, a menudo ignoramos el operador de preimagen e indicamos simplemente la medida de probabilidad de X, escribiendo:

$$P_X(\cdot) = PX^{-1}(\cdot).$$

Esta es la formulación que conocemos en la probabilidad básica y estadística; por ejemplo, la expresión $X \sim N(0, 1)$ describe esta medida de probabilidad inducida en $\mathbb{R}$ por la aplicación X es una distribución normal estándar. Cuando no haya posibilidad de confusión, omitiremos el subíndice "X" y simplemente escribiremos P en lugar de P_X.

Cuando P_X, una medida en el espacio X, está dominada por una medida σ-finita μ, el *teorema de Radon-Nikodym* establece que existe una densidad $dP_X/d\mu = p_X$, y

$$P_X(A) = \int_A p_X \, d\mu.$$

Este es el caso familiar al que estamos acostumbrados; cuando μ es la medida de conteo, p_X es una función de masa de probabilidad, y cuando μ es la medida de Lebesgue, p_X es una función de densidad de probabilidad. Uno de los beneficios de la formulación en teoría de la medida es que no tenemos que tratar estos dos casos importantes por separado.

Sea φ una función medible de valores reales definida en $\mathbb{X}$. Entonces, el valor esperado de $\varphi(X)$ es:

$$E_X\{\varphi(X)\} = \int_{\mathbb{X}} \varphi(x) \, dP_X(x) = \int_{\mathbb{X}} \varphi(x) p_X(x) \, d\mu(x),$$

donde la última expresión se cumple solo cuando $P_X \ll \mu$ para una medida σ-finita μ en $\mathbb{X}$. Las propiedades usuales del valor esperado (por ejemplo, linealidad) se mantienen en este caso más general; las mismas herramientas que usamos en la teoría de la medida para estudiar propiedades de integrales de funciones medibles son útiles para derivar este tipo de resultados.

En estas notas, se asumirá que estás familiarizado con todos los cálculos básicos de probabilidad definidos y utilizados en cursos básicos de probabilidad y estadística. En particular, se espera que conozcas las distribuciones más comunes (por ejemplo, normal, binomial, Poisson, gamma, uniforme, etc.) y cómo calcular valores esperados para estas y otras distribuciones. Además, se asumirá que estás familiarizado con algunas operaciones básicas que involucran vectores aleatorios (por ejemplo, matrices de covarianza) y algunos conceptos básicos de álgebra lineal.

En probabilidad y estadística, los espacios producto son especialmente importantes. La razón, como se insinuó anteriormente, es que la independencia de variables aleatorias está relacionada con los espacios producto y, en particular, con medidas producto. Si $X_1, \ldots, X_n$ son iid P_X, entonces su distribución conjunta es la medida producto:

$$P_{X_1} \times P_{X_2} \times \cdots \times P_{X_n} = P_X \times P_X \times \cdots \times P_X = P_X^n.$$

El primer término solo expresa "independencia"; el segundo requiere "idénticamente distribuidas"; el último término es solo una notación abreviada para el término intermedio.

Cuando hablamos de teoremas de convergencia, como la *ley de los grandes números*, decimos algo como: para una secuencia infinita de variables aleatorias $X_1, X_2, \ldots$, algún evento ocurre con probabilidad 1. Pero, ¿cuál es la medida que se está considerando aquí? En el caso iid, resulta ser una *medida producto infinita*, escrita como P_X^∞. Más adelante hablaremos más sobre esto.

1.3.2. Distribuciones condicionales

Las distribuciones condicionales, en general, son bastante abstractas. Cuando las variables aleatorias en cuestión son discretas (μ es la medida de conteo), sin embargo, las cosas son más simples; la razón es que los eventos donde la variable aleatoria toma un valor fijo tienen probabilidad positiva, por lo que la fórmula de probabilidad condicional ordinaria que involucra cocientes puede aplicarse.

Cuando una o más de las variables aleatorias en cuestión son continuas (dominadas por la medida de Lebesgue), se debe tener más cuidado. Supongamos que las variables aleatorias X y Y tienen una distribución conjunta con función de densidad $p_{X,Y}(x, y)$, con respecto a alguna medida dominante (producto) $\mu \times \nu$. Entonces, las distribuciones marginales correspondientes tienen densidades con respecto a μ y ν, respectivamente, dadas por

$$p_X(x) = \int p_{X,Y}(x, y)\, d\nu(y) \quad \text{y} \quad p_Y(y) = \int p_{X,Y}(x, y)\, d\mu(x).$$

Además, la distribución condicional de Y, dado $X = x$, también tiene una densidad con respecto a ν, y está dada por la razón:

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x,y)}{p_X(x)}.$$

Como función de x, para un y dado, esto es claramente μ-medible, ya que las densidades conjuntas y marginales son medibles. Además, para un x dado, $p_{Y|X}(y \mid x)$ define una medida de probabilidad Q_x, llamada *distribución condicional de* Y, *dado* $X = x$, a través de la integral:

$$Q_x(B) = \int_B p_{Y|X}(y \mid x)\, d\nu(y).$$

Es decir, $p_{Y|X}(y \mid x)$ es la derivada de Radon-Nikodym para la distribución condicional Q_x. Para nuestros propósitos, la distribución condicional siempre puede definirse a través de esta densidad condicional, aunque, en general, una densidad condicional puede no existir incluso si la distribución condicional Q_x sí existe. Existen casos raros donde se requiere la definición más general de distribución condicional, por ejemplo, en la demostración de la factorización de Neyman-Fisher y en la prueba del teorema general de Bayes.

También vale la pena mencionar que las distribuciones condicionales no son únicas: el punto clave es que la densidad condicional puede redefinirse arbitrariamente en un conjunto de μ-medida cero, sin afectar la integral que define $Q_x(B)$ arriba. No profundizaremos en este punto aquí, pero los estudiantes deben ser conscientes de las sutilezas de las distribuciones condicionales; ver la página de Wikipedia[3] sobre la *Borel paradox* para obtener una explicación de estas dificultades, junto con referencias como Jaynes (2003), Capítulo 15.

Dada una distribución condicional bien definida $p_{Y|X}(y \mid x)$, podemos definir las probabilidades condicionales y los valores esperados. Es decir,

$$P(Y \in B \mid X = x) = \int_B p_{Y|X}(y \mid x)\, d\nu(y).$$

Aquí uso la notación más estándar para la probabilidad condicional. La ley de la probabilidad total nos permite escribir

$$P(Y \in B) = \int P(Y \in B \mid X = x) p_X(x)\, d\mu(x),$$

en otras palabras, las probabilidades marginales de Y pueden obtenerse tomando la esperanza de las probabilidades condicionales. Más generalmente, para cualquier función ν-integrable φ, podemos escribir la *esperanza condicional* como

$$E\{\varphi(Y) \mid X = x\} = \int \varphi(y) p_{Y|X}(y \mid x)\, d\nu(y).$$

Podemos evaluar la esperanza anterior para cualquier x, por lo que hemos definido una función μ-medible, digamos, $g(x) = E(Y \mid X = x)$; aquí tomé $\varphi(y) = y$ por simplicidad. Ahora, $g(X)$

[3] https://en.wikipedia.org/wiki/BorelKolmogorov_paradox

es una variable aleatoria, que denotaremos por $E(Y \mid X)$, y podemos preguntarnos sobre su media, varianza, etc. La versión correspondiente de la *ley de la probabilidad total para esperanzas condicionales* es

$$E(Y) = E\{E(Y \mid X)\}. \tag{6}$$

Esta fórmula es llamada *suavización* en Keener (2010), pero probablemente la llamaría una *ley de la esperanza iterada*. Este es, en realidad, un resultado muy poderoso que puede simplificar muchos cálculos; Keener (2010) la usa con frecuencia. Existen versiones de la esperanza iterada para momentos superiores, por ejemplo,

$$V(Y) = V\{E(Y \mid X)\} + E\{V(Y \mid X)\}, \tag{7}$$

$$C(X, Y) = E\{C(X, Y \mid Z)\} + C(E(X \mid Z), E(Y \mid Z)), \tag{8}$$

donde $V(Y \mid X)$ es la *varianza condicional*, es decir, la varianza de Y relativa a su distribución condicional y, de manera similar, $C(X, Y \mid Z)$ es la *covarianza condicional* de X y Y.

Como una observación final sobre distribuciones condicionales, vale la pena mencionar que las distribuciones condicionales son particularmente útiles en la especificación de modelos complejos. De hecho, puede ser difícil especificar directamente una distribución conjunta significativa para una colección de variables aleatorias en una aplicación dada. Sin embargo, a menudo es posible escribir una serie de distribuciones condicionales que, en conjunto, especifican una distribución conjunta significativa. Es decir, podemos construir el modelo paso a paso con distribuciones condicionales de menor dimensión. Esto es particularmente útil para modelos gráficos probabilísticos conocidos como análisis bayesiano.

1.3.3. Desigualdad de Jensen

Los conjuntos y funciones convexas aparecen con bastante frecuencia en aplicaciones de estadística y probabilidad, por lo que puede ser útil ver algunas aplicaciones. El primer resultado, que relaciona la esperanza de una función convexa con la función de la esperanza, debería ser familiar.

Teorema 1.6 Desigualdad de Jensen

$$\varphi[E(X)] \leqslant E[\varphi(X)].$$

Si φ es estrictamente convexa, entonces la igualdad se cumple si y solo si X es constante.
Demostración. Primero, tomemos x_0 como cualquier punto fijo en $\mathbb{X}$. Entonces, existe una función lineal $\ell(x) = c(x - x_0) + \varphi(x_0)$, que pasa por el punto $(x_0, \varphi(x_0))$, tal que $\ell(x) \leqslant \varphi(x)$ para todo x. Para probar nuestra afirmación, tomamos $x_0 = E(X)$, y notamos que

$$\varphi(X) \geqslant c[X - E(X)] + \varphi[E(X)].$$

Tomando esperanzas en ambos lados obtenemos el resultado. ■

La desigualdad de Jensen puede utilizarse para confirmar: $E(1/X) \geqslant 1/E(X)$, $E(X^2) \geqslant E(X)^2$, y $E[\log X] \leqslant \log E(X)$. Una consecuencia interesante es la siguiente:

Ejemplo 1.7: Divergencia de Kullback–Leibler

Sean f y g dos funciones de densidad de probabilidad dominadas por una medida σ-finita μ. La divergencia de Kullback–Leibler de g respecto de f se define como

$$E_f\{\log[f(X)/g(X)]\} = \int \log(f/g)\, d\mu.$$

Se sigue de la desigualdad de Jensen que

$$\begin{aligned} E_f\left\{\log \frac{f(X)}{g(X)}\right\} &= -E_f\left\{\log\left[\frac{g(X)}{f(X)}\right]\right\} \\ &\geqslant -\log E_f\left[\frac{g(X)}{f(X)}\right] \\ &= -\log \int \left(\frac{g}{f}\right) d\mu = 0. \end{aligned}$$

Es decir, la divergencia de Kullback–Leibler es no negativa para toda f y g. Además, es igual a cero si y solo si $f = g$ (μ-almost everywhere). Por lo tanto, la divergencia de Kullback–Leibler actúa como una medida de distancia entre funciones de densidad. Aunque no es una métrica en un sentido formal matemático[a], tiene muchas aplicaciones en estadística.

[a] No es simetrica y no satisface la desiguladad del triangulo

Ejemplo 1.8: Otra prueba de Cauchy–Schwarz

Recordemos que f^2 y g^2 son funciones μ-medibles. Si $\int g^2 d\mu$ es infinito, entonces no hay nada que probar, así que supongamos lo contrario. Definiendo $p = g^2/\int g^2 d\mu$ como una densidad de probabilidad en $\mathbb{X}$, tenemos

$$\begin{aligned} \frac{\left(\int fg\, d\mu\right)^2}{\int g^2 d\mu} &= \left(\int (f/g) p\, d\mu\right)^2 \\ &\leqslant \int (f/g)^2 p\, d\mu = \int \frac{f^2 d\mu.}{g^2 d\mu.} \end{aligned}$$

Donde la desigualdad se sigue del Teorema 1.6. Reordenando los términos obtenemos

$$\left(\int fg\, d\mu\right)^2 \leqslant \int f^2 d\mu \cdot \int g^2 d\mu,$$

que es el resultado deseado.

Otro caso de aplicación de la convexidad y la desigualdad de Jensen aparecerá en el contexto de la teoría de decisiones, que se discutirá más adelante. En particular, cuando la función de pérdida es convexa, se seguirá de la desigualdad de Jensen que las reglas de decisión aleatorias son inadmisibles y, por lo tanto, pueden ignorarse.

1.3.4. Una desigualdad de concentración

Sabemos que las medias muestrales de variables aleatorias iid, para tamaños de muestra grandes, se "concentran" alrededor de la media poblacional. Una desigualdad de concentración proporciona una cota para la probabilidad de que la media muestral se encuentre fuera de un vecindario de la media poblacional. *La desigualdad de Chebyshev* (Ejercicio 25) es un ejemplo de una desigualdad de concentración y, a menudo, estas herramientas son clave para demostrar teoremas límite e incluso algunos resultados de muestra finita en estadística y aprendizaje automático.

Aquí probamos una desigualdad de concentración famosa pero relativamente simple para sumas de variables aleatorias acotadas independientes. Por "variables aleatorias acotadas" entendemos aquellas X_i tales que

$$P(a_i \leqslant X_i \leqslant b_i) = 1.$$

Por un lado, el hecho de estar acotadas implica la existencia de funciones generatrices de momentos. Comenzamos con un resultado simple para una variable aleatoria acotada con media cero; la demostración utiliza algunas propiedades de funciones convexas. Parte de lo que sigue se basa en notas preparadas por Larry Wasserman.

Lema 1.1

Sea X una variable aleatoria con media cero, acotada en el intervalo $[a, b]$. Entonces, la función generatriz de momentos $M_X(t) = \mathbb{E}(e^{tX})$ satisface

$$M_X(t) \leqslant e^{t^2(b-a)^2/8}.$$

Demostración. Escribimos $X = Wa + (1 - W)b$, donde $W = \frac{X-a}{b-a}$. La función $z \mapsto e^{tz}$ es convexa, por lo que obtenemos:

$$e^{tX} \leqslant We^{ta} + (1 - W)e^{tb}.$$

Tomando la esperanza y usando el hecho de que $\mathbb{E}(X) = 0$, se obtiene

$$M_X(t) \leqslant \frac{a}{b-a}e^{ta} + \frac{b}{b-a}e^{tb}.$$

El lado derecho puede reescribirse como $e^{h(\zeta)}$, donde

$$\zeta = t(b - a) > 0, \quad h(z) = -cz + \log(1 - c + ce^z), \quad c = -\frac{a}{b-a} \in (0, 1).$$

Claramente, $h(0) = 0$; de manera similar,

$$h'(z) = -c + \frac{ce^z}{1 - c + ce^z}, \quad \text{por lo que} \quad h'(0) = 0.$$

Además,

$$h''(z) = \frac{c(1 - c)e^z}{(1 - c + ce^z)^2}, \quad h'''(z) = \frac{c(1 - c)e^z(1 - c - ce^z)}{(1 - c + ce^z)^3}.$$

Es fácil verificar que $h'''(z) = 0$ si y solo si $z = \log\left(\frac{1-c}{c}\right)$. Sustituyendo este valor de z en h'', se obtiene $\frac{1}{4}$, que es el máximo global. Por lo tanto, $h''(z) \leqslant \frac{1}{4}$ para todo $z > 0$. Ahora, para algún $z_0 \in (0, \zeta)$, existe una aproximación de Taylor de segundo orden de $h(\zeta)$ alrededor de 0:

$$h(\zeta) = h(0) + h'(0)\zeta + h''(z_0)\frac{\zeta^2}{2} \leqslant \frac{\zeta^2}{8} = \frac{t^2(b-a)^2}{8}.$$

Sustituyendo esta cota, se obtiene $M_X(t) \leqslant e^{h(\zeta)} \leqslant e^{t^2(b-a)^2/8}$. ■

Lema 1.2 Chernoff

Para cualquier variable aleatoria X,

$$P(X > \varepsilon) \leqslant \inf_{t>0} e^{-t\varepsilon}\mathbb{E}(e^{tX}).$$

Ahora estamos listos para el resultado principal, la desigualdad de Hoeffding. La demostración combina los resultados de los dos lemas anteriores.

Teorema 1.7 Desigualdad de Hoeffding

$$P(a \leqslant Y_i \leqslant b) = 1$$

y media μ. Entonces, se cumple que

$$P(|\bar{Y}_n - \mu| > \varepsilon) \leqslant 2e^{-2n\varepsilon^2/(b-a)^2}.$$

Demostración. Podemos tomar $\mu = 0$ sin pérdida de generalidad, trabajando con $X_i = Y_i - \mu$. Por supuesto, X_i sigue estando acotada y la longitud del intervalo de acotamiento sigue siendo $b - a$. Escribimos

$$P(|\bar{X}_n| > \varepsilon) = P(\bar{X}_n > \varepsilon) + P(-\bar{X}_n > \varepsilon).$$

Comenzamos con el primer término del lado derecho. Usando el Lema 2,

$$P(\bar{X}_n > \varepsilon) = P(X_1 + \cdots + X_n > n\varepsilon) \leqslant \inf_{t>0} e^{-tn\varepsilon} M_X(t)^n,$$

donde $M_X(t)$ es la función generatriz de momentos de X_1. Por el Lema 1.1, tenemos

$$P(\bar{X}_n > \varepsilon) \leqslant \inf_{t>0} e^{-tn\varepsilon} e^{nt^2(b-a)^2/8}.$$

El minimizador, sobre $t > 0$, del lado derecho es $t = \frac{4\varepsilon}{(b-a)^2}$, por lo que obtenemos

$$P(\bar{X}_n > \varepsilon) \leqslant e^{-2n\varepsilon^2/(b-a)^2}.$$

Para completar la demostración, aplicamos el mismo argumento a $P(-\bar{X}_n > \varepsilon)$, obtenemos la misma cota que arriba y luego sumamos ambas cotas. ■

Existen muchas otras desigualdades de concentración, la mayoría más generales que la desigualdad de Hoeffding presentada anteriormente. El Ejercicio 28 presenta una desigualdad de concentración para variables aleatorias normales y una ley fuerte correspondiente. El trabajo moderno en desigualdades de concentración trata con tipos más avanzados de cantidades aleatorias, por ejemplo, funciones aleatorias o procesos estocásticos. La siguiente subsección presenta un caso especial de uno de estos resultados.

1.3.5. El "teorema fundamental de la estadística"

Consideremos el problema en el que $X_1, \ldots, X_n$ son iid con función de distribución común F en la recta real; por simplicidad, asumimos que F es continua en todas partes.

Por supuesto, si conociéramos F, entonces, al menos en principio, sabríamos todo sobre la distribución de las variables aleatorias. También debería ser claro, al menos intuitivamente, que si n es grande, habríamos observado "todos los valores posibles" de una variable aleatoria $X \sim F$ en sus frecuencias relativas, por lo que debería ser posible aprender F a partir de una secuencia suficientemente larga de datos.

El resultado siguiente, conocido como el *teorema de Glivenko-Cantelli* o, para algunos, el *teorema fundamental de la estadística*, demuestra que nuestra intuición es correcta.

Primero necesitamos una definición. Dados $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, queremos construir un estimador $\hat{F}_n$ de F. Una elección natural es la *función de distribución empírica:*

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(X_i), \quad x \in \mathbb{R},$$

es decir, $\hat{F}_n(x)$ es simplemente la proporción de la muestra con valores que no exceden x. Es una consecuencia directa de la desigualdad de Hoeffding (combinada con el lema de Borel-Cantelli) que $\hat{F}_n(x)$ converge casi seguramente a $F(x)$ para cada x. El *teorema de Glivenko-Cantelli* establece que $\hat{F}_n$ converge a F no solo puntualmente, sino *uniformemente*.

Teorema 1.8 Glivenko-Cantelli

$$\|\hat{F}_n - F\|_\infty := \sup_x |\hat{F}_n(x) - F(x)|.$$

Entonces, $\|\hat{F}_n - F\|_\infty$ *converge a cero casi seguramente.*
Demostración. Nuestro objetivo es demostrar que, para cualquier $\varepsilon > 0$,

$$\lim_n \sup_x |\hat{F}_n(x) - F(x)| \leqslant \varepsilon, \quad \text{casi seguramente.}$$

Para comenzar, dado un $\varepsilon > 0$ arbitrario, sea $-\infty = t_1 < t_2 < \ldots < t_J = \infty$ una partición de $\mathbb{R}$ tal que

$$F(t_{j+1}^-) - F(t_j) \leqslant \varepsilon, \quad j = 1, \ldots, J-1.$$

El Ejercicio 29 demuestra la existencia de tal partición. Luego, para cualquier x, existe j tal que $t_j \leqslant x < t_{j+1}$ y, por monotonía,

$$\hat{F}_n(t_j) \leqslant \hat{F}_n(x) \leqslant \hat{F}_n(t_{j+1}^-) \quad \text{y} \quad F(t_j) \leqslant F(x) \leqslant F(t_{j+1}).$$

Esto implica que

$$\hat{F}_n(t_j) - F(t_{j+1}^-) \leqslant \hat{F}_n(x) - F(x) \leqslant \hat{F}_n(t_{j+1}^-) - F(t_j).$$

Sumando y restando términos apropiados en las cotas superior e inferior, obtenemos

$$\hat{F}_n(x) - F(x) \geqslant \hat{F}_n(t_j) - F(t_j) + F(t_j) - F(t_{j+1}^-),$$

$$\hat{F}_n(x) - F(x) \leqslant \hat{F}_n(t_{j+1}^-) - F(t_{j+1}^-) + F(t_{j+1}^-) - F(t_j).$$

Dado que la partición fue definida de esta manera, se tiene que

$$\hat{F}_n(t_j) - F(t_j) - \varepsilon \leqslant \hat{F}_n(x) - F(x) \leqslant \hat{F}_n(t_{j+1}^-) - F(t_{j+1}^-) + \varepsilon.$$

Si aplicamos la ley de los grandes números para cada uno de los J valores finitos de j, entonces las cotas superior e inferior convergen a $\pm\varepsilon/2$, uniformemente en x, lo que completa la demostración. ■

Se conocen resultados aún más fuertes sobre la convergencia de la función de distribución empírica. En particular, Dvoretzky et al. (1956) demostraron que

$$P(\|\hat{F}_n - F\|_\infty > \varepsilon) \leqslant 2e^{-2n\varepsilon^2},$$

lo que implica que la tasa de convergencia es $n^{-1/2}$, es decir, $\|\hat{F}_n - F\|_\infty = O_P(n^{-1/2})$.

¿Qué implicaciones tiene este resultado en estadística? Es decir, ¿por qué se le llama el "teorema fundamental de la estadística"? Esto significa que cualquier cantidad que pueda expresarse en términos de la función de distribución F puede estimarse a partir de los datos. En la mayoría de los casos, el "parámetro" de interés (ver más abajo) es una función(al) de la función de distribución F.

Por ejemplo, la media de una distribución puede expresarse como

$$\theta = \theta(F) = \int x\, dF(x),$$

la mediana como

$$\theta = \theta(F) = F^{-1}(0{,}5),$$

etc. El teorema de Glivenko-Cantelli establece que cualquier $\theta(F)$ puede estimarse con $\theta(\hat{F}_n)$ y, además, se puede esperar que estos estimadores tipo plug-in tengan buenas propiedades.

Como veremos en la Sección 3.6, la distribución estará indexada por un parámetro θ de interés, es decir, escribimos F_θ en lugar de F y $\theta(F)$. El teorema de Glivenko-Cantelli garantiza que es posible aprender sobre F_θ a partir de la muestra; para poder aprender sobre el parámetro de interés, requerimos que θ sea **identificable**, es decir, que $\theta \mapsto F_\theta$ sea una función inyectiva.

Es importante enfatizar que la convergencia puntual, $\hat{F}_n(x) \to F(x)$ para cada x, es una consecuencia automática de la ley de los grandes números (que, en el caso de variables aleatorias acotadas, es una consecuencia de la desigualdad de Hoeffding). El esfuerzo que se requiere aquí es fortalecer la conclusión de convergencia puntual a convergencia uniforme.

Este problema es bastante general—convertir convergencia puntual en convergencia uniforme—y hay un considerable y muy técnico desarrollo sobre este tema. Una buena introducción

se encuentra en van der Vaart (1998, Capítulo 19), donde también se presentan versiones más generales del teorema de Glivenko-Cantelli junto con extensiones (por ejemplo, los "teoremas de Donsker") y una introducción a las herramientas necesarias para demostrar tales teoremas.

1.3.6. Familias paramétricas de distribuciones

Como discutiremos en la seccion 4.1, en un problema estadístico no hay solo una medida de probabilidad en cuestión, sino toda una familia de medidas P_θ indexadas[4] por un parámetro $\theta \in \Theta$. Ya estás familiarizado con esta configuración; $X_1, \ldots, X_n$ iid $N(\theta, 1)$ es un ejemplo común.

Una clase muy importante y amplia de distribuciones es la *familia exponencial*. Es decir, para una medida dominante dada μ, una familia exponencial tiene una función de densidad (derivada de Radon-Nikodym con respecto a μ) de la forma

$$p_\theta(x) = e^{\langle \eta(\theta), T(x) \rangle + A(\theta)} h(x),$$

donde $\eta(\theta)$, $T(x)$, $A(\theta)$ y $h(x)$ son algunas funciones, y $\langle \cdot, \cdot \rangle$ es el producto interno euclidiano.

Debes estar familiarizado con estas distribuciones de un curso previo. Discutiremos las **familias exponenciales** con mayor detalle más adelante.

En esta sección, consideraremos otra familia especial de medidas de probabilidad que se caracterizan por una "medida base" y un grupo de transformaciones. Comenzamos con un caso especial importante.

Ejemplo 1.9:
Sea P_0 una medida de probabilidad con densidad simétrica p_0 respecto a la medida de Lebesgue en $\mathbb{R}$.
La simetría implica que la mediana es 0; si el valor esperado existe, entonces también es 0. Para $X \sim P_0$, definimos $X' = X + \theta$ para algún número real θ.
Entonces, la distribución de X' es

$$P_\theta(A) := P_0(X + \theta \in A).$$

Realizando esto para todos los valores de θ, se genera la familia $\{P_\theta : \theta \in \mathbb{R}\}$. La familia normal $N(\theta, 1)$ es un caso especial.

La familia de distribuciones en el Ejemplo 9 se genera a partir de una única distribución, centrada en 0, y una colección de "desplazamientos de ubicación".

Existen cuatro propiedades clave de estos desplazamientos de ubicación: 1. Desplazar por cero no cambia nada. 2. El resultado de dos desplazamientos consecutivos puede lograrse mediante un solo desplazamiento. 3. El orden en que se realizan los desplazamientos es irrelevante. 4. Para cualquier ubicación dada, existe un desplazamiento que devuelve la ubicación a 0.

[4] Nótese que el subíndice en P_θ tiene un propósito diferente al del subíndice en P_X descrito en la Sección 3.1.

Resulta que estas propiedades caracterizan lo que se conoce como un *grupo de transformaciones*, discutido en la Sección 1.2.2. Keener (2010, Cap. 10) proporciona algunos detalles sobre modelos de transformaciones de grupos, y Eaton (1989) es una referencia completa sobre el tema.

Para generalizar el ejemplo de desplazamiento de ubicación, comenzamos con una medida de probabilidad fija P sobre $(\mathbb{X}, \mathcal{A})$. Ahora introducimos un grupo $\mathcal{G}$ de transformaciones sobre $\mathbb{X}$, y tomamos $P_e = P$; aquí el subíndice "e" se refiere a la identidad del grupo e. Definimos la familia $\{P_g : g \in \mathcal{G}\}$ como

$$P_g(A) = P_e(g^{-1}A), \quad A \in \mathcal{A}.$$

Es decir, $P_g(A)$ es la probabilidad, bajo $X \sim P_e$, de que gX caiga en A. En el caso en que P_e tenga densidad p_e con respecto a la medida de Lebesgue, se tiene que

$$p_g(x) = p_e(g^{-1}x) \left| \frac{dg^{-1}x}{dx} \right|,$$

que es simplemente la fórmula usual de cambio de variable de la probabilidad introductoria; por supuesto, la fórmula anterior supone que cada $g \in \mathcal{G}$ es diferenciable.

La explicación en el párrafo anterior trata sobre la construcción de una familia de distribuciones que, en cierto sentido, es invariante con respecto a $\mathcal{G}$. En muchos casos, como el ejemplo normal anterior, ya existe una familia $\mathbb{P} = \{P_\theta : \theta \in \Theta\}$ de distribuciones en $\mathbb{X}$, indexada por Θ.

Si $\mathcal{G}$ es un grupo de transformaciones sobre $\mathbb{X}$, entonces podríamos preguntarnos si la familia es invariante con respecto a $\mathcal{G}$. Es decir, si $X \sim P_\theta$, ¿es posible que no exista $\theta' \in \Theta$ tal que $gX \sim P_{\theta'}$? En resumen, ¿se cumple que $\mathcal{G}\mathbb{P} = \mathbb{P}$?

1.4. Preliminares conceptuales

1.4.1. Ingredientes de un problema de inferencia estadística

La estadística, en general, se ocupa de la recopilación y el análisis de datos. La etapa de recopilación de datos es importante, pero no será considerada aquí—supondremos que los datos están dados y nos enfocaremos únicamente en cómo deben analizarse estos datos.

En nuestro caso, el problema estadístico general que enfrentaremos consiste en datos X, posiblemente vectoriales, que toman valores en $\mathbb{X}$, junto con un modelo que describe el mecanismo que produjo estos datos.

Por ejemplo, si $X = (X_1, \ldots, X_n)$ es un vector que consiste en las alturas registradas de n estudiantes de la UNI, entonces el modelo podría indicar que estos individuos fueron muestreados completamente al azar de la población total de estudiantes de la UNI y que las alturas de los estudiantes en la población siguen una distribución normal. En resumen, escribiríamos

algo como $X_1, \ldots, X_n$ son iid $N(\mu, \sigma^2)$; aquí, "iid" significa independiente e idénticamente distribuido.

No habría nada que analizar si la población en cuestión fuera completamente conocida. En el ejemplo de las alturas, se supone que al menos uno de los parámetros μ y σ^2 es desconocido y queremos usar los datos observados X para aprender algo sobre estas cantidades desconocidas.

Así, en cierto sentido, la población en cuestión es en realidad solo una clase o familia de distribuciones—en el ejemplo de las alturas, esta es la colección de todas las distribuciones normales (univariadas). De manera más general, especificamos una familia paramétrica $\{P_\theta : \theta \in \Theta\}$, discutida en la Sección 1.3.6, como el *modelo* para los datos observables X; en otras palabras, $X \sim P_\theta$ para algún $\theta \in \Theta$, aunque el θ específico que corresponde a la observación $X = x$ es desconocido.

El desafío del estadístico es aprender algo sobre el verdadero θ a partir de las observaciones. Sin embargo, el significado de "aprender algo" no es tan fácil de explicar; intentaré aclararlo en la siguiente sección.

Los datos y el modelo son ingredientes familiares en el problema de inferencia estadística. Sin embargo, hay un elemento importante pero menos familiar en la inferencia estadística: la *función de pérdida*, la cual no recibe mucha atención en los cursos introductorios de inferencia.

Para facilitar esta discusión, consideremos el problema de intentar estimar θ basándonos en datos $X \sim P_\theta$. La función de pérdida L registra cuánto "pierdo" al adivinar que θ es igual a algún valor particular a en Θ. En otras palabras, la función $(\theta, a) \mapsto L(\theta, a)$ es simplemente una función real definida sobre $\Theta \times \Theta$.

En los cursos introductorios, usualmente se toma

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

la llamada *pérdida cuadrática*, sin mucha explicación. En este curso, consideraremos funciones de pérdida más generales en problemas de inferencia más amplios, particularmente cuando discutamos la teoría de decisión.

Para resumir, el problema de inferencia estadística consiste en datos X que toman valores en un espacio muestral Θ y una familia de distribuciones de probabilidad $\{P_\theta : \theta \in \Theta\}$. En algunos casos, será necesario considerar la función de pérdida $L(\cdot, \cdot)$, y en otros casos habrá una distribución de probabilidad conocida Π definida sobre el espacio de parámetros Θ, que representa algún conocimiento previo sobre el parámetro desconocido, el cual deberá incorporarse de alguna manera.

En cualquier caso, el objetivo es identificar la distribución particular P_θ que produjo los datos observados X.

1.4.2. Razonamiento de la muestra a la población

Se cree generalmente que la estadística y la probabilidad están estrechamente relacionadas. Aunque esta afirmación es cierta en cierto sentido, la conexión no es inmediata ni obvia.

Claramente, el modelo de muestreo general "$X \sim P_\theta$" es una afirmación probabilística. Por ejemplo, si $X \sim N(\theta, 1)$ con θ conocido, entonces podemos calcular

$$P_\theta(X \leqslant c) = \Phi(c - \theta)$$

para cualquier c, donde Φ es la función de distribución normal estándar. Cálculos similares pueden realizarse para otras distribuciones dependiendo de un θ conocido. Pero este ejercicio consiste en hacer afirmaciones probabilísticas sobre un valor aún no observado de una variable aleatoria X con un parámetro θ conocido. Es decir, la probabilidad está diseñada para describir la incertidumbre sobre una muestra que será tomada de una población fija y conocida. El problema estadístico, por otro lado, es uno en el que la muestra es dada, pero alguna característica de la población es desconocida. Básicamente, esto es lo opuesto al problema de probabilidad y, visto bajo esta luz, parece muy difícil. Además, no está claro cómo usar la probabilidad o incluso si debería usarse en absoluto.

Un problema crucial es que no está claro cómo interpretar afirmaciones probabilísticas sobre $X \sim P_\theta$ después de haber observado[5] X. Una ilustración de esta idea se encuentra en el contexto de los valores p en pruebas de hipótesis. Si el valor p es pequeño, entonces el valor observado es un "atípico" con respecto a la distribución hipotetizada. Es común interpretar tal resultado como evidencia en contra de la hipótesis, pero esto es una *elección* que el estadístico está haciendo—no hay una base matemática o de otro tipo para manejar el problema de esta manera. El punto clave aquí es que el modelo de muestreo, por sí solo, es insuficiente para la inferencia estadística; se necesita algo más. Para ilustrar aún más este punto, consideremos el argumento *fiducial* de Fisher para la inferencia estadística. Supongamos que los datos X y el parámetro θ son ambos escalares, y sea $F_\theta(x)$ la función de distribución. Tomemos cualquier $p \in (0, 1]$ y supongamos que la ecuación $p = F_\theta(x)$ puede resolverse de manera única para x, dado θ, y para θ, dado x. Es decir, existen funciones $x_p(\theta)$ y $\theta_p(x)$ tales que

$$p = F_\theta(x_p(\theta)) = F_{\theta_p(x)}(x), \quad \forall (x, \theta).$$

Si el modelo de muestreo es "monótono" en el sentido de que, para todo (p, x, θ),

$$x_p(\theta) \geqslant x \iff \theta_p(x) \leqslant \theta,$$

entonces es fácil demostrar que

$$p = P_\theta\{X \leqslant x_p(\theta)\} = P_\theta\{\theta_p(X) \leqslant \theta\}.$$

La idea de Fisher fue tomar la última expresión y darle una interpretación después de observar $X = x$. Es decir, definió

$$P\{\theta \geqslant \theta_p(x)\} = p, \quad \forall p \in [0, 1], \text{ dado que } x \text{ es el observado } X.$$

[5]Los estudiantes probablemente se hayan encontrado con esta dificultad en su primer contacto con la *fórmula de Bayes*, donde una probabilidad condicional se invierte y se intenta usar la probabilidad para explicar la incertidumbre sobre el resultado de un experimento que ya ha sido realizado pero aún no ha sido observado.

La colección $\{\theta_p(x) : p \in [0,1]\}$ define los cuantiles de una distribución y, por lo tanto, una distribución en sí misma. Fisher llamó a esto la *distribución fiducial* y formuló la controvertida afirmación de que había llevado a cabo la tarea bayesiana de obtener una especie de "distribución posterior" para el parámetro sin necesidad de una distribución previa ni de invocar el teorema de Bayes; ver Zabell (1992) para más detalles.

Nuestro objetivo aquí no es discutir la validez de las afirmaciones de Fisher, sino simplemente señalar que la construcción de Fisher de una distribución fiducial, aunque intuitiva, requiere una especie de "salto de fe"—de hecho, la palabra *fiducial* significa literalmente "basado en la creencia o fe".

Por lo tanto, el argumento fiducial no es una derivación matemática de una solución al problema de inferencia estadística basada únicamente en el modelo de muestreo.

En general, evitaremos preocupaciones filosóficas en este curso, pero los estudiantes deben ser conscientes de que: (i) la inferencia estadística es difícil, y (ii) no existe un enfoque ampliamente aceptado.

El problema es que la inferencia estadística está mal formulada desde un punto de vista matemático, por lo que no se puede deducir, desde principios fundamentales, una "respuesta correcta". (Por esta razón, nadie puede afirmar que un enfoque es "correcto" o mejor que otro; el pequeño poema en ingles en la Figura 1 es relevante aquí.)

El profesor Ronald Fisher reflexionó cuidadosamente sobre estos temas y, aunque su argumento fiducial no es completamente satisfactorio, iba en la dirección correcta. El argumento fiducial estaba, en esencia, destinado a facilitar

> *la conversión de la información en los datos observados en un resumen significativo de la evidencia que respalda la veracidad de diversas hipótesis relacionadas con el parámetro de interés.*

Esta es **mi definición de inferencia estadística que emplearemos en este curso**. Siguiendo esta idea, ha habido intentos de extender o mejorar el argumento original de Fisher, incluyendo la *inferencia fiducial generalizada* (Hannig 2009), la *inferencia estructural* (Fraser 1968) y la *teoría de Dempster-Shafer* (Dempster 2008; Shafer 1976).

Un punto importante que falta en estos enfoques existentes es una declaración de qué hace que sus resúmenes sean "significativos". El nuevo *marco de modelos inferenciales* (Martin y Liu 2013, 2015b) aclara qué significa "significativo", pero no entraremos en este punto aquí.

Aunque estos enfoques alternativos discutidos anteriormente, que no son ni frecuentistas ni bayesianos, aún no han alcanzado la corriente principal, los avances son prometedores y tengo esperanzas en ellos.

1.5. Tipos de modelos Estadísticos. Definiciones formales

Sea $(X_1, \ldots, X_n)$ un vector aleatorio con una familia de probabilidad $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Adicionalmente, sea Λ^n el espacio muestral del vector, T una funcion sobre Λ^n y I_Θ la clase de todos los intervalos contenidos en Θ.

- (Estimación puntual) Encontrar una función $T : \Lambda^n \to \Theta$ con propiedades óptimas para asi determinar la población $P_{T(x)}$.
- (Estimación intervalar) Encontrar una función $T : \Lambda^n \to I_\Theta$ con propiedades óptimas.
- (Prueba de hipótesis) Sea $\mathcal{P}_0$ y $\mathcal{P}_1$ dos subfamilias disjuntas de $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. y las hipótesis $H_0 : P \in \mathcal{P}_0$ contra $H_1 : P \in \mathcal{P}_1$. Encontrar una función $T : \Lambda^n \to \{0, 1\}$ tal que $T(x) = 1$ implica rechazar H_0 y $T(x) = 0$ implica no rechazar H_0.

Definición 1.3 **Definición 2. Modelo Estadístico Clásico Una clase $\{P_\theta\}$ en $(\Lambda, \mathcal{G})$ indexada por un parámetro $\theta \in \Theta$ es una familia paramétrica si y solo si $\Theta \subset \mathcal{R}^d$ para $d \in \mathcal{N}_+$ y cada P_θ es una medida de probabilidad totalmente especificada. Θ es llamado espacio paramétrico de dimensión d. Un modelo estadístico paramétrico para los eventos en $(\Lambda, \mathcal{G})$ es la tripla**

$$\Big(\Lambda, \mathcal{G}, \{P_\theta : \theta \in \Theta\}\Big). \tag{9}$$

- **Un modelo paramétrico asume que la población P pertenece a una familia de medidas de probabilidad paramétrizada o indexada por $\theta \in \Theta$.**
- **Una familia paramétrica $\{P_\theta : \theta \in \Theta\}$ es identificable si y solo si $\theta_1 \neq \theta_2$ y $\theta_i \in \Theta$ implican $P_{\theta_1} \neq P_{\theta_2}$.**
- **Si la familia está dominada por ν, entonces esta puede ser identificada por la familia de densidades $\{f_\theta\} = \left\{\frac{dP}{d\nu} : P_\theta \in \mathcal{P}\right\}$ de P con respecto a ν.**

Definición 1.4 Familia Exponencial

Una familia paramétrica $\{P_\theta : \theta \in \Theta\}$ dominada por una medida σ-finita ν en $(\Lambda, \mathcal{G})$ se llama **familia exponencial** si y solo si

$$\frac{dP_\theta}{d\nu}(x) = \exp\left\{\eta^\top(\theta)T(x) - \xi(\theta)\right\} h(x), \quad x \in \Lambda, \tag{10}$$

donde $\exp\{x\} = e^x$, T y $\eta(\theta)$ son vectores de dimensión p, η es una función de θ y T función de X, h es una función Boreliana no negativa y $\xi(\theta)$ es llamada de constante de normalización de la densidad[a].

[a] T y h son funciones de x. η y ξ son funciones de θ.

Definición 1.5 Modelo Bayesiano a posteriori

Un modelo de probabilidad **a priori** para θ es $(\Theta, \mathcal{F}_\Theta, \Pi_\theta)$, en que $\Pi_\theta(B) = \int_B \pi(\theta)d\theta$ y $\pi(\theta)$ es la densidad a priori de θ (o probabilidad cambiando la integral por una sumatoria).
Luego de observar $\mathbf{X} = \mathbf{x}$, el **modelo Bayesiano a posteriori** de $\theta|\mathbf{X} = \mathbf{x}$ es la tripla $(\Theta, \mathcal{F}_\Theta, P_{\theta|\mathbf{X}=\mathbf{x}})$ en que

$$P_{\theta|\mathbf{X}=\mathbf{x}}(B) = \int_B \pi(\theta|\mathbf{X} = \mathbf{x})d\theta \text{ para } B \in \mathcal{F}_\Theta. \tag{11}$$

y $\pi(\theta|\mathbf{X} = \mathbf{x})$ es la densidad a posteriori de θ dado la muestra observada $\mathbf{X} = \mathbf{x}$.

Por el teorema de Bayes, la densidad a posteriori se rescribe:

$$\pi(\theta|\mathbf{X} = \mathbf{x}) = \frac{\pi(\theta)f(\mathbf{x}|\theta)}{\int_\Theta \pi(\theta)f(\mathbf{x}|\theta)d\theta}, \quad g(\mathbf{x}) = \int_\Theta \pi(\theta)f(\mathbf{x}|\theta)d\theta \text{ marginal de } \mathbf{x}.$$

1.5.1. Ejemplo de modelo Bernoulli y Poison

- Consideremos que el parametro de interes es la proporción $\theta \in (0, 1)$ y que a priori (antes de observar los datos) le provemos de un modelo Beta(a, b). Supongamos, que $X_i|\theta = \theta_0 \sim$ Bernoulli(θ_0). El modelo de probabilidad a posteriori de $\theta|\mathbf{X} = \mathbf{x}$ es la medida de probabilidad Beta$(\sum_{i=1}^n x_i + a; n - \sum_{i=1}^n + b)$.
- Supongamos que el parametro de interes es la taza de ocurrencia $\theta > 0$ y que a priori (antes de observar los datos) le provemos de un modelo Gama(a, b). Supongamos, que $X_i|\theta = \theta_0 \sim$ Poisson(θ_0). El modelo de probabilidad a posteriori de $\theta|\mathbf{X} = \mathbf{x}$ es la medida de probabilidad Gama $(\sum_{i=1}^n x_i + a; n + b)$.
- Un modelo a priori para θ, representado por π, es **conjugado** con $f(\mathbf{x}|\theta)$ cuando $\pi(\theta)$ y $\pi(\theta|\mathbf{X} = \mathbf{x})$

1.5.2. Ejemplo de modelo Normal

- Si el parametro de interes es la media $\theta \in \mathcal{R}$ y que a priori (antes de observar los datos) le provemos de un modelo Normal(a, b). Supongamos, que $X_i|\theta = \theta_0 \sim$ Normal(θ_0, σ_0^2) (σ_0^2 es conocido). El modelo de probabilidad a posteriori de $\theta|\mathbf{X} = \mathbf{x}$ es la medida de probabilidad

$$\text{Normal}\left\{\frac{\sum_{i=1}^n \frac{x_i}{\sigma_o^2} + \frac{a}{b^2}}{\frac{n}{\sigma_0^2} + \frac{1}{b^2}}; \frac{1}{\frac{n}{\sigma_0^2} + \frac{1}{b^2}}\right\}.$$

Definición 1.6 Modelo no parametrico

Un **modelo estadístico no paramétrico** es una tripla

$$\Big(\Lambda, \mathcal{G}, \{P_s : s \in S\}\Big),$$

en que s es una **función** que pertenece a una clase de funciones S bien definida. Por

ejemplo, el espacio L^2 en los reales o en el intervalo unitario.

- Sea $\epsilon \sim N(0, \sigma^2)$, una variable X de modelo desconocido e $Y = s(X) + \epsilon$, en que s es una función Boreleana en L^2. El modelo para $Y|X \sim N(s(X), \sigma^2)$ es no paramétrico. El problema estadístico consiste en estimar la función $s(X)$.
- Sea $H : \mathcal{R} \to [0, 1]$ conocida (por ejemplo Φ o la Logística) y s antes definida. Un modelo no paramétrico para una variable Bernoulli $Y|X$ considera una función de probabilidad
$$p_s(y) = H(s(x))^y[1 - H(s(x))]^{1-y}.$$

1.5.3. Ejemplos de modelo estadístico no paramétrico 1

- Consideremos una familia de funciones de probabilidad acumulada (f.d.a) $\mathcal{L} = \left\{F : \int_{-\infty}^{\infty} xdF(x) < \infty\right.$
El modelo de probabilidad indexado por $\mathcal{L}$ es la tripla $(\mathcal{R}, \mathcal{B}, \{P_F : F \in \mathcal{L}\})$, en que
$$\left\{P_F : P_F(A) = \int_A dF(x) \quad A \subset \mathcal{R} \text{ y } F \in \mathcal{L}\right\}.$$
Podemos estimar $\theta = \int_{-\infty}^{\infty} xdF(x)$ por $\widehat{\theta} = n^{-1}\sum_{i=1}^n X_i$.
- Consideremos una familia de f.d.a $\mathcal{L} = \left\{F : \int |x|^2 dF(x) < \infty\right\}$. El modelo de probabilidad indexado por $\mathcal{L}$ es la tripla $(\mathcal{R}, \mathcal{B}, \{P_F : F \in \mathcal{L}\})$. Podemos estimar el funcional
$$\theta = \int\int 2^{-1}(x_1 - x_2)^2 dF(x_1)dF(x_2) \quad \text{(varianza de } F\text{)}.$$
por $\widehat{\theta} = \binom{n}{2}^{-1}\sum_{1\leqslant i<j\leqslant n}^n 2^{-1}(X_i - X_j)^2$.

1.5.4. Ejemplos de modelo estadístico no paramétrico 2

- Consideremos una familia de f.d.a bivariada
$$\mathcal{L} = \left\{F : \int |xy|^2 dF(x, y) < \infty\right\}.$$
El modelo de probabilidad indexado por $\mathcal{L}$ es la tripla $(\mathcal{R}, \mathcal{B}, \{P_F : F \in \mathcal{L}\})$. Podemos estimar el funcional
$$\theta = \int_{\mathcal{R}}\int_{\mathcal{R}} 2^{-1}(x_1 - x_2)(y_1 - y_2)dF(x_1, y_2)dF(x_2, y_2) \quad \text{(cov. de F)}$$
por $\widehat{\theta} = \binom{n}{2}^{-1}\sum_{1\leqslant i<j\leqslant n}^n 2^{-1}(X_i - X_j)(Y_i - Y_j)$.

1.5.5. Ejemplo de modelos estadísticos

Modelo logistico paramétrico

Sea $(Y_1, \ldots, Y_n)$ un conjunto de variables aleatorias independientes en el espacio de probabilidad $(\Omega, \mathcal{F})$. En un GLM, cada Y_i pertenece a la *familia exponencial* y su densidad (o masa) viene dada por

$$f(y_i; \theta_i, \phi) \;=\; \exp\big[\, \phi\,\{\, y_i\,\theta_i \;-\; b(\theta_i)\}\;+\; c(y_i, \phi)\big], \tag{12}$$

donde

- θ_i es el *parámetro natural*.
- $b(\theta)$ es la *función cumulante*, dos veces diferenciable y convexa.
- $\phi > 0$ es el *parámetro de precisión* (su inverso, ϕ^{-1}, es la dispersión).
- $c(y, \phi)$ es el término de normalización, independiente de θ.

La *densidad conjunta* de la muestra $\mathbf{Y} = (Y_1, \ldots, Y_n)$ es

$$f(\mathbf{y}; \boldsymbol{\theta}, \phi) \;=\; \prod_{i=1}^{n} \exp\big[\, \phi\,(y_i\,\theta_i - b(\theta_i)) + c(y_i, \phi)\big] \;=\; \exp\Big[\phi \sum_{i=1}^{n} (y_i\,\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(y_i, \phi)\Big].$$

Propiedades: A partir de (12) se deducen

$$\mathbb{E}[Y_i] = b'(\theta_i) =: \mu_i, \tag{13}$$

$$\operatorname{Var}(Y_i) = \frac{1}{\phi}\, b''(\theta_i) = \phi^{-1}\, V(\mu_i), \quad V(\mu) = b''\big((b')^{-1}(\mu)\big). \tag{14}$$

La *parte sistemática* une la media a un predictor lineal mediante un *enlace* g:

$$g(\mu_i) = \eta_i, \quad \eta_i = x_i^{\top}\beta, \quad \beta \in \mathbb{R}^p.$$

De este modo, el GLM se expresa como

$$\big(\,\mathcal{X},\, \mathcal{F},\, \{\, P_{\beta,\phi} : \beta \in \mathbb{R}^p,\ \phi > 0\}\big),$$

en que $\mathcal{X} = \{0, 1\}^n$.

Regresión logística como caso particular: Para regresión logística:

- $Y_i \sim \text{Bernoulli}(\mu_i)$, con $\mu_i = \Pr(Y_i = 1 \mid x_i)$.
- $\phi = 1$, $a(\phi) = 1$.
- Parámetro natural: $\theta_i = \log \dfrac{\mu_i}{1 - \mu_i}$.
- Función cumulante: $b(\theta) = \log\big(1 + e^{\theta}\big)$.

- Término de normalización: $c(y_i, 1) = 0$.

La densidad de cada observación es

$$f(y_i; \theta_i) = \exp\Big[y_i\,\theta_i \;-\; \log\big(1 + e^{\theta_i}\big)\Big]. \tag{15}$$

El enlace canónico (logit) y su inversa son

$$g(\mu_i) = \log\frac{\mu_i}{1-\mu_i}, \quad \mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}, \quad \eta_i = x_i^\top \beta.$$

La Media explicada: En regresión logística, la media condicionada

$$\mathbb{E}[Y_i \mid x_i] = \mu_i$$

es precisamente la probabilidad de éxito $\Pr(Y_i = 1 \mid x_i)$, que el modelo busca explicar a través del predictor lineal $\eta_i = x_i^\top \beta$.

5. Regresión Logística Bayesiana

Aprovechando la notación del GLM logístico anterior, añadimos una perspectiva bayesiana sobre los parámetros β.

Modelo

$$Y_i \mid x_i, \beta \;\sim\; \text{Bernoulli}(\mu_i), \quad \mu_i = \frac{\exp(x_i^\top \beta)}{1+\exp(x_i^\top \beta)}, \tag{16}$$

$$\beta \;\sim\; \mathcal{N}_p\big(0,\, \Sigma_0\big), \tag{17}$$

donde:

- $Y_i \in \{0, 1\}$ es la respuesta observada para el vector de predictores $x_i \in \mathbb{R}^p$.
- μ_i es la probabilidad de éxito descrita por el enlace logit: $\mu_i = g^{-1}(x_i^\top \beta)$ con $g^{-1}(u) = e^u/(1+e^u)$.
- La *prior* de los coeficientes β es una gaussiana multivariada de media cero y covarianza Σ_0 (p. ej. $\Sigma_0 = \tau^2 I_p$).

Modelo logisitco posterior

La *densidad posterior* de β dado los datos $\{(y_i, x_i)\}_{i=1}^n$ es

$$p(\beta \mid y, X) \;\propto\; \Big[\prod_{i=1}^{n} \mu_i^{y_i}(1-\mu_i)^{1-y_i}\Big]\; \exp\Big(-\tfrac{1}{2}\beta^\top \Sigma_0^{-1}\beta\Big).$$

No existe una forma analítica cerrada para esta posterior; se explora mediante:

- **Aproximación de Laplace:** aproximar con una normal en el modo de la posterior.
- **Muestreo MCMC:** p. ej. algoritmo de Metropolis–Hastings o Hamiltonian Monte Carlo.

Ejemplo básico

Supongamos $p = 2$, $\Sigma_0 = 10\, I_2$, y datos $\{(y_i, x_i)\}_{i=1}^n$. Entonces:

$$\ell(\beta) = \sum_{i=1}^{n} \big[y_i\, x_i^\top \beta - \log(1 + e^{x_i^\top \beta}) \big] \; - \; \tfrac{1}{2}\, \beta^\top (10\, I_2)^{-1} \beta$$

y la aproximación de Laplace calcula $\hat{\beta} = \arg\max_\beta \ell(\beta), \quad \mathrm{Cov}(\beta \mid y, X) \approx -\big[\nabla^2 \ell(\hat{\beta})\big]^{-1}$.

De este modo obtenemos una distribución aproximada $\beta \mid y, X \approx \mathcal{N}(\hat{\beta},\, \widehat{V})$, útil para inferencia y predicción bayesiana.

Modelo Temporal de Ruido Blanco Gaussiano con Drift

Definimos el proceso $X(t)$, $t \in \mathbb{R}$, como

$$X(t) \;=\; m(t)dt \;+\; \epsilon dW(t),$$

donde:

- $m(t)$ es la función determinística de deriva (drift), $m \in L^2(\mathbb{R})$.
- $W(t)$ es ruido blanco gaussiano en tiempo continuo, caracterizado por

$$\mathbb{E}[W(t)] = 0, \quad \mathrm{Cov}\big(W(s), W(t)\big) = \sigma^2\, \delta(t - s).$$

Propiedades

$$\mathbb{E}[X(t)] = m(t), \tag{18}$$

$$\mathrm{Cov}\big(X(s), X(t)\big) = \mathrm{Cov}\big(\xi(s), \xi(t)\big) = \sigma^2\, \delta(t - s). \tag{19}$$

Comentarios de análisis funcional

- El proceso $X(t)$ no tiene trayectorias clásicamente continuas; se interpreta como un *campo generalizado* por la presencia de $\delta(t - s)$.

- La covarianza singular $K(s,t) = \sigma^2\delta(t-s)$ refleja independencia instantánea de los incrementos.
- La deriva $m(t)$ aporta la componente determinística, mientras que $\xi(t)$ introduce una variabilidad de alta frecuencia, estudiable en espacios de Sobolev fraccionarios.

1. K-means como modelo no paramétrico

Sea

$$\Omega = \mathbb{R}^p, \quad \mathcal{F} = \text{Borel}(\mathbb{R}^p), \quad \mathcal{P} = \{\, P : P \text{ es medida de probabilidad en } (\Omega, \mathcal{F})\}.$$

Definimos el espacio de parámetros

$$\Theta = \Big\{ \big(c, \mu_1, \ldots, \mu_K\big) : c : \Omega \to \{1, \ldots, K\} \text{ medible}, \ \mu_j \in \mathbb{R}^p \Big\}.$$

Los parámetros $(c, \mu_{1:K})$ se estiman imponiendo la condición de minimización:

$$(c^*, \mu^*_{1:K}) = \arg \min_{(c,\mu_{1:K})} \int \big\|x - \mu_{c(x)}\big\|^2 \, dP(x), \tag{20}$$

$$\text{sujeto a:} \quad c^*(x) \in \arg \min_{1\le j\le K} \|x - \mu_j\|, \quad \mu^*_j = \frac{1}{P\{x : c^*(x) = j\}} \int_{c^*(x)=j} x \, dP(x). \tag{21}$$

Suposiciones no paramétricas sobre los datos:

- P puede ser cualquier distribución con momento segundo finito: $\displaystyle\int \|x\|^2 \, dP(x) < \infty$.
- No se impone forma de densidad ni número fijo de parámetros: la complejidad crece con n.
- La medida empírica $\widehat{P}_n$ sustituye a P en la práctica, llevando al algoritmo iterativo clásico.

2. Bosques Aleatorios como modelo no paramétrico

Sea

$$\Omega = \mathcal{X} \times \mathcal{Y}, \quad \mathcal{F} = \text{Borel}(\mathcal{X} \times \mathcal{Y}), \quad \mathcal{P} = \{\, P : P \text{ es medida de probabilidad en } (\Omega, \mathcal{F})\}.$$

El espacio de parámetros es el conjunto de funciones de predicción:

$$\Theta = \big\{\, f : \mathcal{X} \to \mathcal{Y} \text{ medible}\big\}.$$

La función poblacional buscada depende de la pérdida L:

$$\text{Regresión (p. cuadrático):} \quad f^*(x) = \arg\min_f \mathrm{E}_{(X,Y)\sim P}\big[(Y - f(X))^2\big] = \mathrm{E}[Y \mid X = x], \tag{22}$$

$$\text{Clasificación (0–1 loss):} \quad f^*(x) = \arg\min_f \Pr\big\{Y \neq f(X)\big\} = \arg\max_{y\in\mathcal{Y}} \Pr(Y = y \mid X = x). \tag{23}$$

El estimador de Random Forest combina M árboles $\{f_m\}$:

$$\widehat{f}_{\mathrm{RF}}(x) = \begin{cases} \dfrac{1}{M}\displaystyle\sum_{m=1}^{M} f_m(x), & \text{(regresión)}, \\ \text{modo}\{f_m(x)\}_{m=1}^{M}, & \text{(clasificación)}. \end{cases}$$

Cada f_m se ajusta sobre un *bootstrap* de la muestra y una selección aleatoria de variables por nodo.

Suposiciones no paramétricas:

- Los datos $\{(X_i, Y_i)\}$ son iid con distribución P.
- No se asume forma paramétrica en P; basta cierta regularidad (p. ej., momentos finitos).
- La complejidad del modelo crece libremente con el tamaño de los datos y la profundidad de los árboles.

1.6. Ejercicios

1. Mostrar que si $A_1, A_2, \ldots$ son miembros de una σ-álgebra $\mathcal{A}$, entonces también lo es $\bigcap_{i=1}^{\infty} A_i$.

2. Para $A_1, A_2, \cdots \in \mathcal{A}$, definimos

$$\lim\sup A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m = \{x : x \text{ está en } A_n \text{ para infinitos } n\}.$$

   Mostrar que $\lim\sup A_n$ también está en $\mathcal{A}$.

3. Demuestre el **Lema de Borel-Cantelli**: Si μ es una medida finita (es decir, $\mu(\mathbb{X}) < \infty$) y

$$\sum_{n=1}^{\infty} \mu(A_n) < \infty,$$

   entonces $\mu(\lim\sup A_n) = 0$.

4. Demuestre que si f y g son funciones medibles, entonces también lo son $f+g$ y $f \vee g = \max\{f, g\}$. *[Sugerencia: Para demostrar que $f+g$ es medible, observe que si $f(x) \leqslant a - g(x)$, entonces existe un número racional r que está entre $f(x)$ y $a - g(x)$].*

5. Mostrar que si f es μ-integrable, entonces

$$\left| \int f \, d\mu \right| \leqslant \int |f| \, d\mu.$$

   [Sugerencia: Escriba $|f|$ en términos de f^+ y f^-].

6. *a*) Use el **teorema de Fubini** para demostrar que, para una variable aleatoria X no negativa con función de distribución F, se tiene

$$\mathbb{E}(X) = \int_0^{\infty} (1 - F(x)) \, dx.$$

   b) Use este resultado para derivar la media de una distribución exponencial con parámetro de escala θ.

7. Sea $(\mathcal{G}, \cdot)$ un grupo. Demuestre que:

   a) $g \cdot e = g$ para todo $g \in \mathcal{G}$;

   b) $g \cdot g^{-1} = e$ para todo $g \in \mathcal{G}$;

   c) La identidad e es única;

   d) Para cada g, el inverso g^{-1} es único;

   e) Para cada g, se cumple $(g^{-1})^{-1} = g$.

8. Mostrar que $\mathbb{P} = \{N(0, \theta) : \theta > 0\}$ es invariante con respecto al grupo $\mathcal{G} = \{g_a(x) = ax : a > 0\}$.

9. Suponga que φ es convexa en (a, b) y ψ es convexa y no decreciente en el rango de φ. Demuestre que $\psi \circ \varphi$ es convexa en (a, b), donde $\circ$ denota composición de funciones.

10. Suponga que $\varphi_1, \ldots, \varphi_n$ son funciones convexas y que $a_1, \ldots, a_n$ son constantes positivas. Demuestre que $\varphi(x) = \sum_{i=1}^{n} a_i \varphi_i(x)$ es convexa.

11. Sea $\{C_t : t \in T\}$ una colección de conjuntos convexos. Demuestre que $\bigcap_{t \in T} C_t$ también es convexo.

12. Sea f una función convexa de valores reales y, para cualquier t, defina $C_t = \{x : f(x) \leqslant t\}$. Demuestre que C_t es convexo.

13. Sea $X \sim N(\mu, \sigma^2)$ y, dado $X = x$, $Y \sim N(x, \tau^2)$. Encuentre la distribución condicional de X dado $Y = y$.

14. Demuestre las fórmulas de esperanza condicional (6), (7), (8) que estan en la pagina 14.

15. Sea X una variable aleatoria.

    a) Si $\mathbb{E}(X^2) < \infty$, encuentre c que minimice $\mathbb{E}\{(X - c)^2\}$.

    b) Si $\mathbb{E}|X| < \infty$, encuentre c que minimice $\mathbb{E}|X - c|$.

16. **Una versión inversa de la desigualdad de Jensen.** Sea X una variable aleatoria acotada, es decir, $\mathbb{P}(X \in [a, b]) = 1$. Si f es una función creciente, entonces

$$\mathbb{E}f(X) \leqslant f(\mathbb{E}(X) + d),$$

    donde $d = b - a$.

17. Sean f y g funciones de densidad correspondientes a $N(\theta, 1)$ y $N(\mu, 1)$. Calcule la divergencia de Kullback-Leibler $K(f, g)$.

18. **Desigualdad de Markov.**

    a) Sea X una variable aleatoria positiva con media $\mathbb{E}(X)$. Demuestre que

$$\mathbb{P}(X > \varepsilon) \leqslant e^{-1}\mathbb{E}(X), \quad \forall \varepsilon > 0.$$

    b) Considere un espacio de medida $(\mathbb{X}, \mathcal{A}, \mu)$, donde $\mu(\mathbb{X}) < \infty$, y una función μ-integrable f. Enuncie y demuestre una versión general de la desigualdad de Markov en el contexto de teoría de medida.

19. **Desigualdad de Chebyshev.** Sea X una variable aleatoria con media μ y varianza σ^2. Use la desigualdad de Markov para demostrar que

$$\mathbb{P}(|X - \mu| > \varepsilon) \leqslant \varepsilon^{-2}\sigma^2, \quad \forall \varepsilon > 0.$$

20. Demuestre la **cota de Chernoff**, Lema 2.

21. *a*) Especialice la **desigualdad de Hoeffding** (Teorema 7) para el caso en que $X_1, \ldots, X_n$ son variables aleatorias $\text{Ber}(\mu)$.

    b) Dado un η pequeño, encuentre $\varepsilon = \varepsilon(n, \eta)$ tal que $\mathbb{P}(|\bar{X}_n - \mu| \leqslant \varepsilon) \geqslant 1 - \eta$.

22. *a*) Sea $Z \sim N(0, 1)$. Demuestre que $\mathbb{P}(|Z| > \varepsilon) \leqslant e^{-1}e^{-\varepsilon^2/2}$.

b) Sea $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ y $\bar{X}_n$ la media muestral. Dé una cota similar a la anterior para $\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon)$.

c) Use su desigualdad en (b), junto con el **lema de Borel-Cantelli**, para demostrar la **ley fuerte de los grandes números** para normales: Si $X_1, X_2, \cdots \sim N(\mu, \sigma^2)$, entonces $\bar{X}_n \to \mu$ casi seguramente.

23. Sea F una función de distribución en la recta real, y sea $\varepsilon > 0$ un número fijo. Defina $t_1 = -\infty$ y, para $j > 1$,

$$t_{j+1} = \sup\{t : F(t) \leqslant F(t_j) + \varepsilon\}.$$

a) Demuestre que esta secuencia es finita y define la partición usada en la demostración del Teorema 8.

b) ¿Cuántos puntos t_j se necesitan para la partición?

24. Muestre que si X está distribuida según una **familia de escala**, entonces $Y = \log X$ está distribuida según una **familia de ubicación**.

25. Sea X una variable aleatoria positiva y considere la familia $\mathbb{P}$ de distribuciones generadas por X y las transformaciones $\mathcal{G} = \{g_{b,c} : b > 0, c > 0\}$, dada por

$$g_{b,c}(x) = bx^{1/c}.$$

a) Demuestre que $\mathcal{G}$ es un grupo bajo composición de funciones.

b) Si X sigue una distribución exponencial con tasa unitaria, demuestre que la familia $\mathbb{P}$ generada por $\{g_{b,c}\}$ es la **familia de Weibull**, con densidad

$$\frac{c}{b}\left(\frac{x}{b}\right)^{c-1} \exp\left\{-\left(\frac{x}{b}\right)^{c}\right\}, \quad x > 0.$$

26. El intervalo estándar de confianza al $100(1-\alpha)\,\%$ para la media normal con varianza conocida $\sigma^2 = 1$ es

$$\bar{X} \pm z_{1-\alpha/2}\sigma n^{-1/2},$$

donde $\Phi(z_{1-\alpha/2}) = 1 - \alpha/2$.

a) ¿Qué significa que la probabilidad de cobertura sea $1 - \alpha$?

b) Explique cómo debe interpretarse este intervalo después de observar los datos.

27. Un concepto fundamental en la teoría frecuentista de la estadística es la **distribución muestral**. Para una muestra observable $X_1, \ldots, X_n$ de una distribución que depende de algún parámetro θ, sea $T = T(X_1, \ldots, X_n)$ una estadística.

a) ¿Qué entendemos por la **distribución muestral** de T?

b) Explique cómo se usa la distribución muestral para razonar sobre la inferencia estadística. Si es útil, puede usar un ejemplo para explicarlo.

Capítulo 2: Familias Exponenciales, Suficiencia e Información

Teoría Estadística Avanzada

Sigla DES124

Prof. Jaime Lincovil

2.1. Introducción

En estadística, la suficiencia, ancilaridad e información son conceptos fundamentales, sin importar el enfoque que se adopte: bayesiano, frecuentista u otro. La idea básica es que, para un modelo estadístico dado $\{P_\theta : \theta \in \Theta\}$, indexado por un espacio de parámetros (finito-dimensional) Θ, existen funciones de los datos observables $X = (X_1, \ldots, X_n)$ que contienen toda la información disponible en X sobre el parámetro desconocido θ. Tales funciones se denominan **estadísticos suficientes**, y la idea es que, en general, es suficiente, por ejemplo, para la estimación puntual, restringir la atención a funciones de estadísticos suficientes. La noción de "información" introducida anteriormente es informal; sin embargo, para mayor rigor, debemos formular una noción precisa de información en un problema estadístico. En particular, nos enfocaremos en la información de Fisher, atribuida a R. A. Fisher. El resultado clave es:

> la información de Fisher para θ en una función $T = T(X)$ de los datos observables X no es mayor que la información de Fisher para θ en X mismo, y ambas medidas de información son iguales si y solo si T es un estadístico suficiente.

La definición de suficiencia no es útil para encontrar un estadístico suficiente en un problema dado. Afortunadamente, el **Teorema de factorización de Neyman-Fisher** facilita bastante esta tarea. La idea es que, con un poco de álgebra sobre la función de verosimilitud, se puede obtener un estadístico suficiente fácilmente. Sin embargo, los estadísticos suficientes no son únicos. Por lo tanto, existe interés en tratar de encontrar el "mejor" estadístico suficiente en un problema dado. Este mejor estadístico suficiente se denomina **mínimo** o **mínimal**, y discutimos algunas técnicas para encontrar el estadístico suficiente mínimo. Sin embargo, puede ocurrir que incluso un estadístico suficiente mínimo T proporcione una reducción ineficiente de los datos X, ya sea porque la dimensión de T es mayor que la de θ, o porque hay información redundante en T. En tales casos, tiene sentido considerar la posibilidad de condicionar sobre un **estadístico auxiliar**, una especie de complemento de un estadístico suficiente que no contiene información sobre θ. Existen casos especiales en los que el estadístico suficiente mínimo T es **completo**. Esto significa que T no contiene información redundante sobre θ, por lo que condicionar sobre un estadístico auxiliar es innecesario (teorema de Basu).

Iniciamos este capítulo con una introducción a las familias exponenciales. Esta es una clase amplia que contiene casi todas las distribuciones que se encuentran en un curso intermedio de estadística. De manera general, lo que hace que las familias exponenciales sean tan útiles es que sus funciones de verosimilitud correspondientes tienen propiedades convenientes que permiten aplicar muchas técnicas. Por ejemplo, las condiciones de regularidad necesarias para la normalidad asintótica de los estimadores de máxima verosimilitud o la desigualdad de Cramér-Rao se cumplen para las familias exponenciales (regulares). Un resultado clave es el **Teorema 1**, que es una aplicación interesante del teorema de convergencia dominada de Lebesgue.

De particular importancia aquí es que, esencialmente, solo las familias exponenciales (regulares) permiten una reducción adecuada de la dimensión mediante la suficiencia (**Teorema 4**). Los detalles sobre las familias exponenciales, incluyendo todo lo presentado aquí, se discuten en la monografía técnica de Brown (1986).

2.2. Familias exponenciales de distribuciones

Anteriormente discutimos el concepto general de una familia paramétrica de medidas de probabilidad, con cierto detalle sobre un caso especial con una estructura inducida por un grupo de transformaciones. En esta sección discutimos una clase muy importante de distribuciones que contiene muchos de los modelos estadísticos comunes, tales como la distribución Binomial, de Poisson, Normal, etc.

Definición 2.1

Una colección de medidas de probabilidad $\{P_\theta : \theta \in \Theta\}$ en $(\mathbb{X}, \mathcal{A})$, cada una dominada por una medida σ-finita μ, se llama **familia exponencial** si las derivadas de Radon-Nikodym $p_\theta(x) = \frac{dP_\theta}{d\mu}(x)$ satisfacen:

$$p_\theta(x) = h(x)e^{\langle \eta(\theta), T(x)\rangle - A(\theta)} \tag{1}$$

para algunas funciones h, A, η y T, donde

$$\eta(\theta) = (\eta_1(\theta), \ldots, \eta_d(\theta))^\top \quad \text{y} \quad T(x) = (T_1(x), \ldots, T_d(x))^\top.$$

Aquí, $\langle x, y\rangle$ denota el producto interno euclidiano usual entre vectores en $\mathbb{R}^d$, es decir,

$$\langle x, y\rangle = \sum_{i=1}^{d} x_i y_i.$$

Cuando es conveniente, podemos escribir $a(\theta) = e^{-A(\theta)}$ y reescribir la ecuación (1) como

$$p_\theta(x) = a(\theta)h(x)e^{\langle \eta(\theta), T(x)\rangle}.$$

Al considerar esperanzas con respecto a una distribución de la familia exponencial, ocasionalmente podemos absorber el término $h(x)$ en $d\mu(x)$; ver, por ejemplo, el Teorema 1.

Ejemplo 2.1

Supongamos que $X \sim \text{Poisson}(\theta)$. La distribución de Poisson está dominada por la medida de conteo en $\mathbb{X} = \{0, 1, \dots\}$, con densidad

$$p_\theta(x) = \frac{e^{-\theta}\theta^x}{x!} = \frac{1}{x!}e^{x\log\theta - \theta}, \quad x = 0, 1, \dots$$

El lado derecho de esta expresión tiene la forma de (1), por lo que la distribución de Poisson pertenece a la familia exponencial.

Ejemplo 2.2

Las siguientes distribuciones son miembros de la familia exponencial: $\mathcal{N}(\theta, 1)$, $\mathcal{N}(0, \theta^2)$, $\text{Exp}(\theta)$, $\text{Gamma}(\theta_1, \theta_2)$, $\text{Beta}(\theta_1, \theta_2)$, $\text{Bin}(n, \theta)$ y $\text{Geo}(\theta)$. Algunos ejemplos comunes de distribuciones que no pertenecen a una familia exponencial incluyen $\text{Cauchy}(\theta, 1)$ y $\text{Unif}(0, \theta)$.

Existen varias propiedades estadísticas interesantes de las familias exponenciales, en particular aquellas relacionadas con la existencia de estadísticos suficientes y, posteriormente, con la existencia de estimaciones insesgadas de varianza mínima. Las siguientes propiedades matemáticas de las familias exponenciales serán útiles para demostrar estos resultados estadísticos.

Proposición 2.1

Considere una familia exponencial con densidades respecto a μ:

$$p_\theta(x) = a(\theta)h(x)e^{\langle\theta, T(x)\rangle}. \tag{2}$$

El conjunto

$$\Theta = \left\{\theta : \int h(x)e^{\langle\theta,T(x)\rangle}d\mu(x) < \infty\right\}$$

es convexo.

En la Proposición 1, el correspondiente *espacio de parámetros naturales*. Hemos expresado esto en términos de la notación θ, pero la idea básica es partir de (1) y tomar $\eta(\theta)$ como el parámetro; es decir, simplemente hacer una reparametrización. El resultado establece que el espacio de parámetros naturales es un conjunto "bien comportado".

El siguiente teorema es útil para una serie de cálculos, en particular, para calcular momentos en familias exponenciales o la información de Fisher. La demostración es una aplicación interesante del teorema de convergencia dominada.

Teorema 2.1

Sea $X \in \mathbb{X} \subseteq \mathbb{R}^d$ con densidad $p_\theta(x) = a(\theta)e^{\langle\theta,x\rangle}$ con respecto a μ. Sea $\varphi : \mathbb{X} \to \mathbb{R}$ una función μ-medible y defina el conjunto

$$\Theta_\varphi = \left\{\theta : \int |\varphi(x)|e^{\langle\theta,x\rangle}d\mu(x) < \infty\right\}.$$

Entonces, para θ en el interior de Θ_φ, la función

$$m(\theta) := \int \varphi(x)e^{\langle\theta,x\rangle}d\mu(x)$$

es continua y tiene derivadas continuas de todo orden. Además, la derivada puede tomarse dentro del signo integral; es decir, para $i = 1, \ldots, d$, se cumple

$$\frac{\partial m(\theta)}{\partial \theta_i} = \int x_i \varphi(x) e^{\langle \theta, x \rangle} d\mu(x).$$

Demostración. Consideremos primero el caso de θ unidimensional; el caso general en dimensión d es similar. Sea $\theta \in \Theta_\varphi$ un valor fijo. Para un $\varepsilon > 0$ adecuado, definimos

$$d_n(x) = \frac{e^{\varepsilon x/n} - 1}{\varepsilon/n},$$

de modo que

$$\frac{m(\theta + \varepsilon/n) - m(\theta)}{\varepsilon/n} = \int \varphi(x) e^{\theta x} d_n(x) d\mu(x).$$

Es claro que $d_n(x) \to d(x) = x$ para cada x, donde $d(x)$ es la derivada de la función $z \mapsto e^{zx}$ en $z = 0$.
Definimos $f_n(x) = \varphi(x) e^{\theta x} d_n(x)$, por lo que $f_n(x) \to x\varphi(x) e^{\theta x}$ cuando $n \to \infty$ para todo x. Resta demostrar que la μ-integral de f_n converge a la μ-integral de f. Para ello, aplicamos el teorema de convergencia dominada.
Notamos las siguientes desigualdades para la función exponencial:

$$|e^z - 1| \leqslant |z| e^{|z|} \quad \text{y} \quad |z| \leqslant e^{|z|}, \quad \forall z.$$

Con estas desigualdades, podemos escribir

$$|f_n(x)| \leqslant |\varphi(x)| e^{\theta x} \varepsilon^{-1} e^{2\varepsilon |x|} \leqslant |\varphi(x)| e^{\theta x} \varepsilon^{-1} (e^{2\varepsilon x} + e^{-2\varepsilon x}).$$

Si elegimos ε de modo que $\theta \pm 2\varepsilon$ pertenezca a Θ_φ, entonces la cota superior, digamos $g(x)$, es μ-integrable. Por lo tanto, el teorema de convergencia dominada establece que

$$\frac{dm(\theta)}{d\theta} = \lim_{n \to \infty} \int f_n(x) d\mu(x) = \int f(x) d\mu(x) = \int x\varphi(x) e^{\theta x} d\mu(x).$$

Es decir, "la derivada de la integral es la integral de la derivada", como se quería demostrar. Para demostrar que se pueden tomar más derivadas, y que estas derivadas adicionales pueden evaluarse tomando la derivada dentro de la integral, basta repetir el argumento anterior. ■

Existen un par de supuestos ocultos que vale la pena mencionar. Primero, hemos supuesto implícitamente que el soporte $\{x : p_\theta(x) > 0\}$ de la distribución no depende de θ. Esto descarta casos como $\text{Unif}(0, \theta)$. Segundo, también hemos supuesto implícitamente que Θ tiene un interior no vacío. Esto asegura que es posible encontrar un intervalo/caja abierto, dependiendo de ε, centrado en el valor dado θ y que encaje dentro de Θ. Sin esto, el enunciado podría ser falso. Estos dos supuestos forman parte de aquellas "condiciones de regularidad" mencionadas en las definiciones clásicas de familias exponenciales.

El mismo resultado es válido para familias exponenciales generales, no solo para aquellas en forma natural o canónica. El resultado clave es el siguiente: *para funciones φ suficientemente bien comportadas, el valor esperado de $\varphi(X)$ es una función bien comportada del parámetro θ*. Como aplicación del Teorema 1, tenemos lo siguiente.

Corolario 2.1

Supongamos que $X = (X_1, \ldots, X_d)$ tiene una densidad de familia exponencial de la forma

$$p_\theta(x) = a(\theta)e^{\langle \theta, x\rangle}.$$

Entonces, para $i, j = 1, \ldots, d$, se cumple que

$$\mathbb{E}_\theta(X_i) = -\frac{\partial}{\partial \theta_i}\log a(\theta) \quad \text{y} \quad \mathrm{Cov}_\theta(X_i, X_j) = -\frac{\partial^2}{\partial \theta_i \partial \theta_j}\log a(\theta).$$

Los momentos de orden superior de X pueden determinarse de manera similar.

Demostración. Partimos de la identidad

$$\int a(\theta)e^{\langle x, \theta\rangle} d\mu(x) = 1, \quad \text{para todo } \theta.$$

Ahora diferenciamos ambos lados con respecto a θ tantas veces como sea necesario, intercambiamos la derivada con la integral y resolvemos para el momento adecuado. ■

Podemos reescribir este resultado en una forma quizás más familiar reconociendo que $-\log a(\theta) = A(\theta)$. En este caso, por ejemplo, se tiene que

$$\mathbb{E}_\theta(X_i) = \frac{\partial}{\partial \theta_i}A(\theta).$$

Una derivación alternativa de este resultado se puede obtener notando que las familias exponenciales admiten una función generadora de momentos, dada por

$$M_\theta(u) = e^{A(\theta+u)-A(\theta)} = \frac{a(\theta)}{a(\theta+u)}. \tag{3}$$

2.3. Estadísticos Suficientes

2.3.1. Definición y el Teorema de Factorización

Un estadístico es simplemente una función medible de los datos; es decir, si $T : \mathbb{X} \to \mathbb{T}$ es medible, entonces $T(X)$ es un *estadístico*. Sin embargo, no todos los estadísticos serán útiles para el problema de inferencia estadística. El objetivo de esta sección es comprender qué tipo de funciones T son relevantes. La definición involucra distribuciones condicionales generales.

Definición 2.2

Suponga que $X \sim P_\theta$. Entonces, el estadístico $T = T(X)$, que mapea $(\mathbb{X}, \mathcal{A})$ a $(\mathbb{T}, \mathcal{B})$, es **suficiente** para $\{P_\theta : \theta \in \Theta\}$ si la distribución condicional de X, dado $T = t$, es independiente de θ. Más precisamente, suponga que existe una función $K : \mathcal{A} \times \mathbb{T} \to [0, 1]$, independiente de θ, tal que $K(\cdot, t)$ es una medida de probabilidad en $(\mathbb{X}, \mathcal{A})$ para

cada t, y $K(A,\cdot)$ es una función medible para cada $A \in \mathcal{A}$, con

$$P_\theta(X \in A, T \in B) = \int_B K(A,t)\, dP_\theta^T(t), \quad \forall A \in \mathcal{A}, B \in \mathcal{B}.$$

Aquí, P_θ^T denota la distribución marginal inducida de T. Entonces, T es un Estadístico Suficiente para θ.

La clave de la definición anterior es que la probabilidad condicional de X, dado $T = t$, caracterizada por el núcleo $K(\cdot, t)$, no depende de θ. Por ejemplo, si φ es alguna función integrable, entonces

$$\mathbb{E}_\theta[\varphi(X) \mid T = t] = \int \varphi(x) dK(x,t)$$

no depende de θ. Llevando este argumento un paso más allá, pues como concecuencia: la suficiencia implica que conocer el valor de T es suficiente para generar nuevos datos X' empleando métodos de muestreo basado en la distribucion condicional $X|T = t$, con las mismas propiedades probabilísticas que X.

Las dificultades de teoría de la medida que surgen con las distribuciones condicionales en el caso continuo hacen que la identificación de Estadísticos Suficientes mediante la definición sea complicada en estos casos; existe un método más práctico, el cual discutiremos en breve. Sin embargo, para problemas discretos, donde la condición es muy sencilla, la definición es adecuada.

Ejemplo 2.3

Suponga que $X_1, \ldots, X_n$ son variables iid Ber(θ). Definimos

$$T(X) = \sum_{i=1}^n X_i.$$

Entonces,

$$P_\theta(X = x \mid T(X) = t) = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1}.$$

Esto es independiente de θ, por lo que $T(X)$ es un Estadístico Suficiente para θ. Aquí, $K(\cdot, t)$ es simplemente una distribución uniforme sobre todos los n-tuplas de 0's y 1's que contienen exactamente t 1's.

Ejemplo 2.4

Supongamos que $X_1, \ldots, X_n$ son i.i.d Pois(θ). Definimos $T(X) = \sum_{i=1}^n X_i$. Entonces,

$$P_\theta(X_1 = x_1 \mid T(X) = t) = \frac{e^{-\theta}\theta^{x_1}/x_1! \cdot e^{-(n-1)\theta}[(n-1)\theta]^{t-x_1}/(t-x_1)!}{e^{-n\theta}(n\theta)^t/t!}$$

$$= \binom{t}{x_1}(1/n)^{x_1}(1-1/n)^{t-x_1}.$$

Es decir, la distribución condicional de X_1, dado $T = t$, es Bin$(t, 1/n)$. Esto se cumple

para todos los X_i, no solo para X_1. De hecho, la distribución condicional del vector X, dado $T = t$, es una distribución multinomial con tamaño t y pesos $1/n$. Como esta distribución es independiente de θ, se concluye que $T(X)$ es suficiente para θ.

Ejemplo 2.5

Sea $X_1, \ldots, X_n$ una muestra iid de una distribución P_θ. Definimos los *estadísticos de orden* $(X_{(1)}, \ldots, X_{(n)})$, que corresponden a la lista ordenada de los datos observados. Dado el orden, existen $n!$ posibles valores de X, y todos ellos tienen la misma probabilidad. Como esta probabilidad es independiente de P_θ, se concluye que los estadísticos de Orden deben ser suficientes para θ. No obstante, es importante notar que la suficiencia puede fallar si no se cumple la suposición de independencia idénticamente distribuida (iid).

En el caso en que la familia $\{P_\theta : \theta \in \Theta\}$ consista en distribuciones dominadas por una medida σ-finita común μ, existe una herramienta conveniente para identificar un estadístico suficiente.

Teorema 2.2 Teorema de Factorización de Neyman-Fisher

Sea $\{P_\theta : \theta \in \Theta\}$ dominada por una medida σ-finita común μ, con densidades $p_\theta = dP_\theta/d\mu$. Entonces, $T(X)$ es un estadístico suficiente para θ si y solo si existen funciones no negativas h y g_θ tales que

$$p_\theta(x) = g_\theta[T(x)]h(x) \quad \text{para todo } \theta, \quad \mu\text{-almost all } x. \tag{4}$$

Demostración. Para una demostración detallada, prestando especial atención a las preocupaciones de teoría de la medida sobre las condiciones, ver Keener (2010), Sec. 6.4. Sin embargo, la idea básica es bastante simple. Como T es un estadístico, una función de X, podemos ver aproximadamente que (i) la densidad conjunta de (X, T) es $g_\theta[T(x)]h(x)$, y (ii) la densidad marginal de T es $g_\theta(t)$. Entonces, la densidad condicional es el cociente de estas dos, y se observa que la dependencia en θ desaparece. Por lo tanto, la distribución condicional de X, dado T, no depende de θ, por lo que T es un estadístico suficiente. ■

Este teorema nos permite identificar fácilmente estadísticos suficientes, simplemente a través de manipulaciones algebraicas de la densidad conjunta o la función de verosimilitud.

Ejemplo 2.6

Supongamos que $X = (X_1, \ldots, X_n)$ consiste en muestras iid de Unif$(0, \theta)$. Entonces, la distribución conjunta se puede escribir como

$$p_\theta(x) = \prod_{i=1}^{n} \theta^{-1} I_{(0,\theta)}(x_i) = \theta^{-n} I_{(0,\theta)}(\max x_i).$$

Dado que $p_\theta(x)$ depende de θ solo a través de $T(X) = \max X_i$, se sigue del Teorema 2 que $T(X) = \max X_i$ es un estadístico suficiente para θ.

Ejemplo 2.7

Supongamos que $X = (X_1, \ldots, X_n)$ consiste en muestras iid de $\mathcal{N}(\mu, \sigma^2)$. La densidad conjunta es

$$\frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{n}{2\sigma^2}\mu^2 \right\}.$$

Por lo tanto, por el Teorema 2, $T(X) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ es un estadístico suficiente para (μ, σ^2). Equivalentemente, $T'(X) = (\bar{X}, s^2(X))$ también es un estadístico suficiente.

2.3.2. Estadísticos Suficientes Mínimos

Es claro que los estadísticos suficientes no son únicos; de hecho, en el Ejemplo 7 se identificaron dos estadísticos suficientes, y los estadísticos de orden también lo son, como es usual. Más generalmente, se demuestra que, si T es suficiente, entonces también lo es $\psi(T)$ para cualquier función biyectiva ψ. Dicho esto, es deseable encontrar el estadístico suficiente que sea el "más pequeño" en algún sentido. Este *estadístico suficiente mínimo* $T = T_{\min}$ es aquel para el cual, dado cualquier otro estadístico suficiente U, existe una función h tal que $T = h(U)$. Una técnica poderosa para encontrar estadísticos suficientes mínimos es descrita en el siguiente teorema.

Teorema 2.3

Supongamos que, para cada $\theta \in \Theta$, P_θ tiene una densidad $p_\theta(x) = g_\theta[T(x)]h(x)$ con respecto a μ. Si

$$p_\theta(x) = cp_\theta(y), \quad \text{para alguna } c = c(x, y),$$

implica que $T(x) = T(y)$, entonces T es un estadístico suficiente mínimo.

Demostración. Ver Keener (2010), página 47. ■

Ejemplo 2.8

Supongamos que $X = (X_1, \ldots, X_n)$ es una muestra iid con densidad común

$$p_\theta(x) = h(x)e^{\sum_{j=1}^{d} \eta_j(\theta)T_j(x) - A(\theta)}. \tag{5}$$

Entonces, $T(X) = [T_1(X), \ldots, T_d(X)]$, con $T_j(X) = \sum_{i=1}^n T_j(X_i)$, es suficiente. Para ver que T es un estadístico suficiente mínimo (bajo cierta condición), aplicamos el Teorema 2.3. Tomemos x e y tales que $p_\theta(x) = cp_\theta(y)$ para alguna función $c = c(x, y)$. Esto implica que

$$\langle \eta(\theta), T(x) \rangle = \langle \eta(\theta), T(y) \rangle + c'$$

para algún $c' = c'(x, y)$. Tomando dos puntos θ_0 y θ_1 en Θ y restando, obtenemos

$$\langle \eta(\theta_0) - \eta(\theta_1), T(x) \rangle = \langle \eta(\theta_0) - \eta(\theta_1), T(y) \rangle,$$

lo que implica

$$\langle \eta(\theta_0) - \eta(\theta_1), T(x) - T(y) \rangle = 0,$$

es decir, $T(x) - T(y)$ y $\eta(\theta_0) - \eta(\theta_1)$ son ortogonales. Como θ_0 y θ_1 son arbitrarios, esto implica que $T(x) - T(y)$ debe ser ortogonal al espacio lineal generado por

$$S = \{\eta(\theta_0) - \eta(\theta_1) : \theta_0, \theta_1 \in \Theta\}.$$

Si S genera todo $\mathbb{R}^d$ (como se explica a continuación), entonces esto implica $T(x) = T(y)$ y, por lo tanto, T es un estadístico suficiente mínimo por el Teorema 3.

El resultado de la condición en el ejemplo anterior— que el espacio S abarque todo el espacio— es suficiente para demostrar que el Estadístico Suficiente natural, $T(X)$, en la familia exponencial es un Estadístico Suficiente mínimo. Sin embargo, esta condición por sí sola no es completamente satisfactoria. Primero, no es algo obvio de verificar y, en segundo lugar, algunas propiedades deseables, como la normalidad asintótica de los estimadores de máxima verosimilitud, requieren aún más regularidad. Por esta razón, a menudo imponemos una condición más fuerte, una que implique, en particular, que el espacio S mencionado anteriormente abarque todo el espacio.

Para formalizar esto, primero definimos que una familia exponencial de la forma (5) tiene rango completo si $\eta(\Theta)$ tiene interior no vacío y $[T_1(x), \ldots, T_d(x)]$ no satisface una restricción lineal para μ-almost all x. Reconocerás estas como condiciones de regularidad adicionales impuestas clásicamente en las familias exponenciales.

Si $\eta(\Theta)$ tiene interior no vacío, entonces también lo tiene

$$\{\eta(\theta_0) - \eta(\theta_1) : \theta_0, \theta_1 \in \Theta\},$$

lo que implica que la envoltura del espacio S en el ejemplo anterior llena todo el espacio. De hecho, si Θ contiene un conjunto abierto y η es una función continua biyectiva, entonces $\eta(\Theta)$ también contiene un conjunto abierto.

Es un ejercicio útil considerar cómo una colección de vectores que contiene un conjunto abierto implicaría que su envoltura abarca todo el espacio. Para esto, consideremos un conjunto abierto en dos dimensiones. Tomemos un vector $v = (v_1, v_2)$ en este conjunto abierto. Que el conjunto sea abierto significa que existe un $\varepsilon > 0$ suficientemente pequeño tal que $\tilde{v} = (v_1 + \varepsilon, v_2 + \varepsilon)$ también pertenece al conjunto.

Se afirma que, siempre que $v_1 \neq v_2$, el par de vectores $(v, \tilde{v})$ es linealmente independiente. Desde el álgebra lineal, existe una prueba de independencia lineal basada en el determinante de la matriz formada por la superposición de estos vectores. En este caso, se tiene:

$$\det \begin{pmatrix} v_1 & v_1 + \varepsilon \\ v_2 & v_2 + \varepsilon \end{pmatrix} = v_1 v_2 + v_1 \varepsilon - v_1 v_2 - v_2 \varepsilon = \varepsilon(v_1 - v_2).$$

Por supuesto, si $v_1 \neq v_2$, entonces este determinante no puede ser cero, por lo que $(v, \tilde{v})$ son linealmente independientes. Finalmente, un par de vectores linealmente independientes en dos dimensiones constituye una base y, por lo tanto, su envoltura abarca todo el espacio.

La mayoría de las familias exponenciales que conocemos tienen rango completo, por ejemplo: Normal, Binomial, Poisson, Gamma, etc. Un ejemplo clásico de una familia exponencial que no tiene rango completo es la $N(\theta, \theta^2)$, que es una de las llamadas *familias exponenciales curvadas* (Keener 2010, Cap. 5). El término *curvado* proviene del hecho de que el espacio del parámetro natural es una curva o, más generalmente, un conjunto cuya dimensión efectiva es menor que la dimensión real. En este caso, el parámetro natural $\eta(\theta)$ está dado por

$$\eta_1(\theta) = \frac{1}{\theta} \quad \text{y} \quad \eta_2(\theta) = -\frac{1}{2\theta^2}.$$

Dado que $\eta_2 = -\eta_1^2/2$, es claro que el espacio del parámetro natural $\eta(\Theta)$ tiene la forma de una parábola invertida. Como el subconjunto unidimensional del espacio bidimensional no puede contener un conjunto abierto, concluimos que esta familia exponencial curvada no puede tener rango completo. Sin embargo, el estadístico suficiente natural $T = (T_1, T_2) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ sigue siendo mínimo suficiente. Para ver esto, necesitamos verificar que el conjunto S definido en el Ejemplo 8 abarque $\mathbb{R}^2$. Tomemos dos pares de puntos (x_1, y_1) y (x_2, y_2) y consideremos los dos vectores de diferencias:

$$v_j = \left(x_j - y_j, \frac{1}{2}(y_j^2 - x_j^2)\right)^\top, \quad j = 1, 2.$$

Coloquemos los dos vectores en una matriz, es decir,

$$\begin{pmatrix} x_1 - y_1 & x_2 - y_2 \\ \frac{1}{2}(y_1^2 - x_1^2) & \frac{1}{2}(y_2^2 - x_2^2) \end{pmatrix}.$$

Si, por ejemplo, tomamos $x_1 = 1$ y $y_1 = -1$, entonces el determinante de la matriz es igual a $y_2^2 - x_2^2$. Por lo tanto, en el caso $x_1 = 1$ y $y_1 = -1$, mientras $x_2 \neq \pm y_2$, el determinante es distinto de cero, los vectores son linealmente independientes y su espacio generado abarca todo el espacio. Por lo tanto, el estadístico suficiente natural T anterior es mínimo suficiente, incluso si la familia exponencial no es de rango completo.

La reducción de dimensión a través de la suficiencia simplifica enormemente las cosas. Por lo tanto, es interesante preguntar en qué problema es posible una reducción tan sustancial de la dimensión. Resulta que, esencialmente, **esto solo ocurre en el caso de la familia exponencial**. Como última parte de terminología, digamos que una familia de distribuciones admite un *estadístico suficiente continuo de dimensión k* si la factorización (4) es válida para todo x (no solo para casi todo x) y si $U(x) = [U_1(x), \ldots, U_k(x)]$ es continuo en x. El siguiente teorema se encuentra en Lehmann y Casella (1998, Cap. 1.6).

Teorema 2.4 Caracterización de Suficiencia

Suponga que $X_1, \ldots, X_n$ son valores reales e iid de una distribución con densidad continua $f_\theta(x)$ con respecto a la medida de Lebesgue, soportada en un intervalo $\mathbb{X}$ que no depende de θ. Denote la densidad conjunta por

$$p_\theta(x) = f_\theta(x_1) \cdots f_\theta(x_n),$$

y suponga que existe un estadístico suficiente continuo de dimensión k. Entonces:

- Si $k = 1$, entonces (5) es válido para algunas funciones h, η_1 y A.
- Si $k > 1$ y si $f_\theta(x_i)$ tiene derivadas parciales continuas con respecto a x_i, entonces (5) es válido para algún $d \leqslant k$.

Este teorema indica que, entre aquellas familias absolutamente continuas suaves con soporte fijo, esencialmente las únicas que admiten un estadístico suficiente continuo son las familias exponenciales. Nótese que el teorema no dice nada sobre aquellos problemas irregulares donde el soporte depende de θ; de hecho, la familia $\text{Unif}(0, \theta)$ admite un estadístico suficiente unidimensional para todo n.

2.3.3. Estadísticos Completos y Auxiliares

Hay Estadísticos Suficientes de diferentes dimensiones; por ejemplo, si $X_1, \ldots, X_n$ son iid $N(\theta, 1)$, entonces los estadísticos de orden y la media $\bar{X}$ son ambos suficientes. La capacidad de un Estadístico Suficiente para admitir una reducción significativa parece estar relacionada con la cantidad de información Auxiliar que contiene.

Un estadístico $U(X)$ se dice que es Auxiliar si su distribución es independiente de θ. Los estadísticos Auxiliares en sí mismos no contienen información relevante sobre θ, pero incluso los Estadísticos Suficientes mínimos pueden contener información Auxiliar. Por ejemplo, en el Ejercicio 7, el Estadístico Suficiente mínimo no es completo.

El hecho de que los estadísticos Auxiliares no contengan información sobre θ no significa que no sean útiles. Una sugerencia común, aunque no universalmente aceptada o utilizada, es realizar análisis condicionales sobre los valores de los estadísticos Auxiliares. Condicionar en algo que no contiene información sobre θ no causa dificultades lógicas, y Fisher argumentó que condicionar en estadísticos Auxiliares es una forma ingeniosa de dar un significado más relevante a la inferencia del problema en cuestión.

Intuitivamente, esto restringe el espacio muestral a un "subconjunto relevante"—el conjunto donde $U(X) = u_{\text{obs}}$—acercando la inferencia condicionada a la observación de X. Esta idea se usa a menudo cuando, por ejemplo, una estimación de máxima verosimilitud no es mínima suficiente.

Un estadístico T es completo si

$$\mathbb{E}_\theta\{f(T)\} = 0 \quad \text{para todo } \theta \text{ implica } f = 0 \text{ casi en todas partes.}$$

En otras palabras, no hay funciones no constantes de T que sean auxiliares.

Alternativamente, un Estadístico Suficiente Completo es aquel que contiene exactamente toda la información sobre θ en X; es decir, no contiene información redundante sobre θ, ya que cada característica $f(T)$ de T tiene información sobre θ. Para ver cómo esto se relaciona con la definición formal, notemos que ninguna función no nula de T es auxiliar.

Los Estadísticos Suficientes Completos son especialmente efectivos para reducir los datos; de hecho, los Estadísticos Suficientes Completos son mínimos.

Teorema 2.5

Si T es completo y suficiente, entonces T también es mínimo suficiente.

Demostración. Sea T' un Estadístico Suficiente Mínimo. Por minimalidad, tenemos $T' = f(T)$ para alguna función f. Escribimos $g(T') = \mathbb{E}_\theta(T \mid T')$, que no depende de θ por suficiencia de T'. Además, por la expectativa iterada, $\mathbb{E}_\theta g(T') = \mathbb{E}_\theta(T)$. Por lo tanto,

$$\mathbb{E}_\theta\{T - g(T')\} = 0 \quad \text{para todo } \theta$$

y, dado que $T - g(T') = T - g(f(T))$ es una función de T, la completitud implica que $T = g(T')$ casi en todas partes. Como $T = g(T')$ y $T' = f(T)$, se concluye que T y T' son equivalentes salvo transformaciones uno a uno; por lo tanto, T también es mínimo suficiente. ■

Dado el poder de un Estadístico Suficiente Completo, es útil poder identificar casos en los que existe. No es sorprendente que las familias exponenciales admitan un Estadístico Suficiente Completo.

Teorema 2.6

Si X se distribuye como una familia exponencial d-dimensional de rango completo con densidad (5), entonces $[T_1(X), \ldots, T_d(X)]$ es completo.

Demostración. Esto es solo un esquema en un caso simple; para la demostración detallada en el caso general, ver Brown (1986, Teorema 2.12). Consideremos el caso unidimensional con

$$p_\theta(x) = e^{\theta x - A(\theta)}$$

y medida dominante μ. Entonces $T(x) = x$. Sea $f(x)$ una función integrable con $\mathbb{E}_\theta\{f(X)\} = 0$ para todo θ. Escribiendo la forma integral de la esperanza, se obtiene

$$\int f(x)e^{\theta x} d\mu(x) = 0 \quad \forall\theta.$$

La integral es esencialmente la transformada de Laplace de f. La transformada de Laplace de la función cero es constante e igual a cero y, dado que las transformadas de Laplace son únicas (μ-a.e.), se sigue que f debe ser la función cero (μ-a.e.). Por lo tanto, X es completo. ■

Ejemplo 2.9

El Teorema 6 muestra que $T(X) = \sum_{i=1}^n X_i$ es completo cuando $X_1, \ldots, X_n$ es una muestra iid de Ber(θ), Pois(θ) y N($\theta, 1$).

Ejemplo 2.10

Sea $X_1, \ldots, X_n$ una muestra iid de N(θ, θ^2). Se mostró anteriormente que $T = (T_1, T_2) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ es minimalmente suficiente. Sin embargo, no es completo. Para ver esto, consideremos la función $f(t_1, t_2) = t_1^2 - \frac{n+1}{2} t_2$. Entonces,

$$\begin{aligned}
\mathbb{E}_\theta f(T_1, T_2) &= \mathbb{E}_\theta(T_1^2) - \frac{n+1}{2}\mathbb{E}_\theta(T_2) \\
&= n\theta^2 + (n\theta)^2 - \frac{n+1}{2} \cdot 2n\theta^2 \\
&= 0 \quad \forall\theta.
\end{aligned}$$

Dado que esta función no es exactamente cero para $T = (T_1, T_2)$ pero tiene esperanza cero, el estadístico T no es completo. Esto no contradice el Teorema 6 porque esta familia exponencial curva no tiene rango completo, es decir, el espacio de parámetros naturales es una curva unidimensional en el plano bidimensional y, por lo tanto, no contiene un conjunto abierto.

Ejemplo 2.11

Sea $X_1, \ldots, X_n$ una muestra iid de Unif$(0, \theta)$. Se afirma que $T(X) = X_{(n)}$ es completo. Un cálculo directo muestra que la densidad de T es

$$p_\theta(t) = nt^{n-1}/\theta^n, \quad 0 < t < n.$$

Supongamos que $\mathbb{E}_\theta\{f(T)\} = 0$ para todo θ. Entonces, tenemos

$$\int_0^\theta t^{n-1} f^+(t)dt = \int_0^\theta t^{n-1} f^-(t)dt \quad \forall \theta > 0.$$

Dado que esto se cumple para los intervalos de integración $[0, \theta]$ para todo θ, también debe cumplirse para todos los intervalos $[a, b]$. El conjunto de todos los intervalos genera la σ-álgebra de Borel, por lo que, en efecto,

$$\int_A t^{n-1} f^+(t)dt = \int_A t^{n-1} f^-(t)dt \quad \text{para todos los conjuntos de Borel } A.$$

Por lo tanto, f debe ser cero casi en todas partes, y así T es completo.

De acuerdo con el teorema de Basu, no tiene sentido condicionar sobre estadísticos auxiliares (como se describió brevemente al comienzo de esta sección) en los casos en los que el estadístico suficiente es completo.

Teorema 2.7 Basu

Si T es un estadístico suficiente y completo para $\{P_\theta : \theta \in \Theta\}$, entonces cualquier estadístico auxiliar U es independiente de T.

Demostración. Dado que U es auxiliar, la probabilidad $p_A = P_\theta(U \in A)$ no depende de θ para cualquier conjunto A. Definimos la distribución condicional $\pi_A(t) = P_\theta(U \in A \mid T = t)$; por expectativa iterada,

$$\mathbb{E}_\theta\{\pi_A(T)\} = p_A \quad \text{para todo } A \text{ y para todo } \theta.$$

Por lo tanto, por completitud, $\pi_A(t) = p_A$ para casi todo t. Dado que la distribución condicional $\pi_A(t)$ de U, dado $T = t$, no depende de t, las dos variables deben ser independientes. ■

Ejemplo 2.12

El teorema de Basu se puede usar para demostrar que la media y la varianza de una muestra independiente de N(μ, σ^2) son independientes. Supongamos primero que σ^2 es conocido y es igual a 1. Sabemos que la media muestral $\bar{X}$ es un estadístico suficiente y completo para μ, y también que la varianza muestral $s^2(X) = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ es auxiliar. Por lo tanto, el teorema de Basu establece que $\bar{X}$ y $s^2(X)$ son independientes. Pero esto fue para $\sigma^2 = 1$, ¿cómo extenderlo al caso de σ^2 desconocido? La clave es que el caso general de σ^2 corresponde a una transformación de escala simple de los datos, lo que claramente no puede alterar la estructura de correlación entre $\bar{X}$ y $s^2(X)$. Por lo tanto, $\bar{X}$ y $s^2(X)$ son independientes para todos los valores de (μ, σ^2).

Ejemplo 2.13

Supongamos que $X_1, \ldots, X_n$ es una muestra iid de $\mathrm{N}(0,1)$, y sea $\bar{X}$ y M la media muestral y la mediana muestral, respectivamente. El objetivo es calcular la covarianza entre $\bar{X}$ y M.
Introducimos un parámetro de media ξ; al hacerlo, encontramos que $\bar{X}$ es un estadístico suficiente y completo, mientras que $\bar{X} - M$ es auxiliar. Luego, el teorema de Basu establece que $\bar{X}$ y $\bar{X} - M$ son independientes, y por lo tanto:

$$0 = C(\bar{X}, \bar{X} - M) = V(\bar{X}) - C(\bar{X}, M) \quad \Longrightarrow \quad C(\bar{X}, M) = n^{-1}.$$

Es común en los cursos de teoría estadística y en los libros de texto dar la impresión de que el teorema de Basu es solo un truco para realizar ciertos cálculos, como en los dos ejemplos anteriores. Sin embargo, la verdadera contribución del teorema de Basu es el punto mencionado anteriormente sobre la condición de los Estadísticos Auxiliares.

2.4. Información de Fisher

2.4.1. Definición

Entendemos informalmente que un Estadístico Suficiente contiene toda la información en $X_1, \ldots, X_n$ sobre el parámetro de interés θ. El concepto de información de Fisher hará esto más preciso.

Definición 2.3

Suponga que θ es d-dimensional y que $p_\theta(x)$ es la densidad de X con respecto a μ. Entonces, asumimos las siguientes *condiciones de regularidad de la Información de Fisher*:

1. $\frac{\partial p_\theta(x)}{\partial \theta_i}$ existe μ-c.t.p. para cada i.
2. $\int p_\theta(x) d\mu(x)$ puede diferenciarse dentro del signo integral.
3. El soporte de p_θ es el mismo para todo θ.

Definición 2.4 Función Score e Información de Fisher

Suponga que se cumplen las condiciones de regularidad de la Información de Fisher. El vector score se define como $\frac{\partial \log p_\theta(X)}{\partial \theta_i}$ para $i = 1, \ldots, d$. La Información de Fisher $I_X(\theta)$ es la matriz de covarianza del vector score; es decir,

$$I_X(\theta)_{ij} = C_\theta \left(\frac{\partial \log p_\theta(X)}{\partial \theta_i}, \frac{\partial \log p_\theta(X)}{\partial \theta_j} \right). \tag{6}$$

Se tiene que el valor esperado del score es cero. En este caso, si podemos diferenciar dos veces dentro del signo integral (como en familias exponenciales; cf. Teorema 1), entonces hay una fórmula alternativa para la Información de Fisher:

$$I_X(\theta)_{ij} = -\mathbb{E}_\theta \left\{ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p_\theta(X) \right\}.$$

Si $X_1, \ldots, X_n$ son iid de una distribución que satisface las condiciones de regularidad de la Información de Fisher, entonces es fácil demostrar que $I_{X_1,\ldots,X_n}(\theta) = nI_{X_1}(\theta)$. Es decir, la información se acumula a medida que se reciben más datos, lo cual tiene sentido si se pretende medir la información en un conjunto de datos. Si los datos no son iid, entonces la información sigue aumentando, pero a una tasa no mayor que en el caso iid.

Cabe mencionar que las condiciones de regularidad de la Información de Fisher no son necesarias aquí. En particular, requerir que $\theta \mapsto p_\theta(x)$ sea diferenciable para todo x es demasiado restrictivo. La Información de Fisher puede definirse bajo la condición mucho menos estricta de *diferenciabilidad en media cuadrática*.

2.4.2. Suficiencia e información

El siguiente resultado ayuda a interpretar que los estadísticos suficientes contienen toda la información relevante sobre θ.

Teorema 2.8

Supongamos que se cumplen las condiciones de regularidad de la Información de Fisher. Supongamos que θ es d-dimensional y que P_θ está dominada por μ. Si $T = g(X)$ es un estadístico, entonces $I_X(\theta) - I_T(\theta)$ es semidefinida positiva. La matriz es nula si y solo si T es suficiente.

Demostración. Como T es una función de X, la distribución conjunta está determinada por la distribución marginal de X. En particular,

$$p_\theta^{X|T}(x \mid t) = \begin{cases} p_\theta^X(x)/p_\theta^T(t) & \text{si } T(x) = t \\ 0 & \text{en otro caso.} \end{cases}$$

Aquí utilizamos p_θ para todas las densidades, y los superíndices indican la distribución. Por lo tanto,

$$p_\theta^X(x) = p_\theta^{X,T}(x,t) = p_\theta^T(t) p_\theta^{X|T}(x \mid t), \quad \text{si } T(x) = t.$$

Tomando logaritmos, obtenemos

$$\frac{\partial \log p_\theta^X(X)}{\partial \theta_i} = \frac{\partial \log p_\theta^T(T)}{\partial \theta_i} + \frac{\partial \log p_\theta^{X|T}(X \mid T)}{\partial \theta_i}, \quad \forall \theta. \tag{7}$$

Mostraremos que los dos términos en el lado derecho son incorrelacionados y que el último término es cero si y solo si T es suficiente. Usando la expectativa iterada,

$$\begin{aligned} C_\theta\left(\frac{\partial \log p_\theta^T(T)}{\partial \theta_i}, \frac{\partial \log p_\theta^{X|T}(X|T)}{\partial \theta_i}\right) &= \mathbb{E}_\theta\left\{\frac{\partial \log p_\theta^T(T)}{\partial \theta_i}\frac{\partial \log p_\theta^{X|T}(X|T)}{\partial \theta_i}\right\} \\ &= \mathbb{E}_\theta\left\{\frac{\partial \log p_\theta^T(T)}{\partial \theta_i}\mathbb{E}_\theta\left(\frac{\partial \log p_\theta^{X|T}(X|T)}{\partial \theta_i}\Big|T\right)\right\}. \end{aligned}$$

Afirmamos que la expectativa condicional interna es cero con probabilidad P_θ^T igual a 1. Para mostrar esto, primero notemos que

$$1 = \int p_\theta^{X|T}(x \mid t)\, d\mu(x) \implies 0 = \frac{\partial}{\partial \theta_i}\int p_\theta^{X|T}(x \mid t)\, d\mu(x),$$

para todo t fuera de un conjunto nulo bajo P_θ^T. Si podemos intercambiar la última derivada y la expectativa condicional, hemos terminado. Dado que

$$\int p_\theta^{X|T}(x \mid t)\, d\mu(x) = \frac{1}{p_\theta^T(t)} \int_{\{x:T(x)=t\}} p_\theta^X(x)\, d\mu(x),$$

tomamos la derivada con respecto a θ_i y simplificamos.

$$\frac{1}{p_\theta^T(t)} \frac{\partial}{\partial \theta_i} \int_{\{x:T(x)=t\}} p_\theta^X(x)\, d\mu(x) - \frac{\partial}{\partial \theta_i} \log p_\theta^T(t). \tag{8}$$

La restricción en el rango de integración no nos impide intercambiar la derivada y la integral de p_θ^X (como en FI). Así, obtenemos

$$\begin{aligned}\frac{\partial}{\partial \theta_i} \int_{\{x:T(x)=t\}} p_\theta^X(x)\, d\mu(x) &= \int_{\{x:T(x)=t\}} \left[\frac{\partial}{\partial \theta_i} \log p_\theta^T(t) + \frac{\partial}{\partial \theta_i} \log p_\theta^{X|T}(x|t)\right] p_\theta^X(x)\, d\mu(x) \\ &= p_\theta^T(t) \frac{\partial}{\partial \theta_i} \log p_\theta^T(t) + p_\theta^T(t) \int \frac{\partial}{\partial \theta_i} \log p_\theta^{X|T}(x|t) p_\theta^{X|T}(x|t)\, d\mu(x).\end{aligned}$$

Este último cálculo muestra que (8) se simplifica a

$$\int \frac{\partial}{\partial \theta_i} \log p_\theta^{X|T}(x|t) p_\theta^{X|T}(x|t)\, d\mu(x).$$

Por lo tanto,

$$\frac{\partial}{\partial \theta_i} \int p_\theta^{X|T}(x \mid t)\, d\mu(x) = \int \frac{\partial}{\partial \theta_i} \log p_\theta^{X|T}(x|t) p_\theta^{X|T}(x|t)\, d\mu(x),$$

y dado que podemos intercambiar la derivada y la integral con respecto a la distribución condicional, obtenemos que la esperanza condicional de la función de score condicional es cero (con P_θ^T-probabilidad 1). Esto, a su vez, muestra que los dos términos en el lado derecho de (7) son incorrelacionados. Luego, la matriz de covarianza de la suma en el lado derecho de (7) es la suma de las respectivas matrices de covarianza. Por lo tanto,

$$I_X(\theta) = I_T(\theta) + \mathbb{E}_\theta\{I_{X|T}(\theta)\},$$

y es claro que $I_X(\theta) - I_T(\theta)$ es semidefinida positiva. La matriz $\mathbb{E}_\theta\{I_{X|T}(\theta)\}$ es cero si y solo si el score condicional $\frac{\partial}{\partial \theta_i} \log p_{X|T,\theta}(X|T)$ es constante (debe ser cero, ¿verdad?) en θ o, en otras palabras, T es suficiente. ■

Esto formaliza la afirmación hecha en la introducción de que los estadísticos suficientes T preservan toda la información sobre θ en los datos X. Es decir, en el caso unidimensional, se tiene $I_T(X) \leqslant I_X$ con igualdad si y solo si T es suficiente.

2.4.3. Desigualdad de Cramer–Rao

Hemos visto que la información de Fisher proporciona una justificación para la afirmación de que los estadísticos suficientes contienen toda la información relevante en una muestra. Sin embargo, la información de Fisher juega un papel aún más profundo en la inferencia estadística; en particular, está involucrada en muchos resultados de optimalidad que proporcionan una

referencia para la comparación entre estimadores, pruebas, etc. A continuación, se presenta un resultado familiar pero importante, que establece que, bajo ciertas condiciones, la varianza de un estimador no puede ser menor que un límite que involucra la información de Fisher.

Teorema 2.9 Cramer–Rao

Para simplificar, tomemos θ como un escalar y supongamos que p_θ satisface las condiciones de regularidad FI. Sea $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_\theta$ y sea $T = T(X_1, \ldots, X_n)$ un estadístico real con $\mathbb{E}_\theta(T) = g(\theta)$. Entonces,

$$V_\theta(T) \geqslant \{g'(\theta)\}^2 \{nI(\theta)\}^{-1}.$$

Demostración. La covarianza entre T y la función de score es $g'(\theta)$. Dado que la varianza del score es $nI(\theta)$, la desigualdad de Cauchy–Schwarz nos da

$$g'(\theta)^2 \leqslant V_\theta(T)\{nI(\theta)\}.$$

Despejando $V_\theta(T)$, se obtiene el resultado deseado. ■

Una aplicación de la desigualdad de Cramer–Rao se encuentra en el diseño experimental. En estos problemas, se tiene control sobre ciertas entradas y el objetivo es seleccionar dichas entradas de manera que el estimador tenga, por ejemplo, la menor varianza posible. En estos casos, la estrategia consiste en elegir esas entradas de tal forma que la información de Fisher se maximice, lo cual tiene una cierta intuición basada en Cramer–Rao, es decir, la varianza es pequeña si la información de Fisher es grande.

2.4.4. Otras medidas de información

¿Es la información de Fisher la única medida de información? Técnicamente, la respuesta es NO, existen otras medidas, pero no reciben tanta atención. La razón es que la información de Fisher es la elección "correcta" siempre que el modelo satisfaga las condiciones de regularidad de FI. Dado que la mayoría de los modelos (por ejemplo, las familias exponenciales regulares) las satisfacen, no hay mucha razón para buscar más allá de la información de Fisher. Sin embargo, existen modelos que no satisfacen las condiciones de regularidad, como la $\text{Unif}(0, \theta)$. En tales casos, la información de Fisher no está definida, por lo que obviamente no puede usarse. La pregunta es si existe algún otro tipo de información, una que se reduzca a la información de Fisher cuando esta existe, pero que sea más versátil en el sentido de que pueda definirse cuando la información de Fisher no puede.
Para extender la información de Fisher, es útil comprender de dónde proviene. Recordemos la divergencia de Kullback–Leibler del Capítulo 1, que (aproximadamente) mide la distancia entre dos modelos. Consideremos aquí dos modelos con funciones de densidad p_θ y $p_{\theta+\varepsilon}$, respecto a la misma medida dominante μ, donde el segundo representa un pequeño cambio en el parámetro. Kullback(1997) muestra que la divergencia de Kullback–Leibler de $p_{\theta+\varepsilon}$ desde p_θ es aproximadamente cuadrática en ε, en particular,

$$K(p_\theta, p_{\theta+\varepsilon}) \approx \varepsilon^\top I(\theta)\varepsilon,$$

cuando $\varepsilon \to 0$, donde $I(\theta)$ es la matriz de información de Fisher mencionada anteriormente. Entonces, la idea clave para generalizar la matriz de información de Fisher es reconocer que,

fuera de los casos regulares, la divergencia de Kullback–Leibler o, mejor aún, la divergencia de Hellinger, definida como

$$h(\theta, \theta') = \int \left(p_\theta^{1/2} - p_{\theta'}^{1/2} \right)^2 d\mu,$$

no es cuadrática en ε. Sin embargo, esta misma expansión puede realizarse y el coeficiente define una adecuada "información de Hellinger." Por ejemplo, consideremos el caso de la distribución $\text{Unif}(0, \theta)$. La divergencia de Hellinger es

$$h(\theta + \varepsilon, \theta) = \int \left[\frac{1}{\sqrt{\theta + \varepsilon}} I_{(0,\theta+\varepsilon)}(x) - \frac{1}{\sqrt{\theta}} I_{(0,\theta)}(x) \right]^2 d\mu(x) = \cdots = \frac{\varepsilon}{\theta} + o(\varepsilon). \tag{9}$$

Esto tiene una aproximación lineal en lugar de cuadrática, resultado de la no regularidad de la distribución $\text{Unif}(0, \theta)$. Sin embargo, una "información de Hellinger" para $\text{Unif}(0, \theta)$ puede definirse como θ^{-1}. También existen versiones de la cota de Cramer–Rao para la información de Hellinger, pero no se presentarán aquí.

2.5. Condicionamiento

Aquí discutimos algunos ejemplos interesantes en los cuales el enfoque frecuentista clásico da respuestas extrañas. Estos ejemplos se utilizarán para motivar el condicionamiento en la inferencia.

Ejemplo 2.14

Suponga que X_1 y X_2 son iid con distribución P_θ que satisface

$$P_\theta(X = \theta - 1) = P_\theta(X = \theta + 1) = 0{,}5, \quad \theta \in \mathbb{R}.$$

El objetivo es construir un intervalo de confianza para el desconocido θ. Considere

$$C = \begin{cases} \{\bar{X}\} & \text{si } X_1 \neq X_2, \\ \{X_1 - 1\} & \text{si } X_1 = X_2. \end{cases}$$

Para ser claros, en cualquier caso, C es un conjunto unitario. Se puede demostrar que C tiene un nivel de confianza del 75 %. Pero analicemos este procedimiento con más cuidado. A partir de la estructura del problema, si $X_1 \neq X_2$, entonces una observación es $\theta - 1$ y la otra es $\theta + 1$. En este caso, $\bar{X}$ es exactamente igual a θ, por lo que, *dado* que $X_1 \neq X_2$, C es garantizado como correcto. Por otro lado, si $X_1 = X_2$, entonces C es $\{\theta\}$ con probabilidad 0.5 y $\{\theta - 2\}$ con probabilidad 0.5, por lo que, *dado* que $X_1 = X_2$, C es correcto con probabilidad 0.5. Juntando esto, C tiene confianza del 100 % cuando $X_1 \neq X_2$ y del 50 % cuando $X_1 = X_2$. En promedio, la confianza es del 75 %, pero dado que, para un problema específico, sabemos en qué caso nos encontramos, ¿no tendría sentido reportar la *confianza condicional* de 100 % o 50 %?

Ejemplo 2.15

Suponga que los datos X pueden tomar valores en $\{1, 2, 3\}$ y que $\theta \in \{0, 1\}$. La distribución de probabilidad de X para cada θ está descrita en la siguiente tabla.

x	1	2	3
$p_0(x)$	0,0050	0,0050	0,99
$p_1(x)$	0,0051	0,9849	0,01

La prueba más poderosa de nivel $\alpha = 0{,}01$ de $H_0 : \theta = 0$ contra $H_1 : \theta = 1$ se basa en la razón de verosimilitud $p_0(x)/p_1(x)$ para un valor observado de $X = x$. Se puede demostrar que esta prueba tiene una potencia de 0.99, lo que sugiere que hay una gran confianza en la decisión basada en el valor observado de x. Pero, ¿es esto cierto? Si se observa $X = 1$, entonces la razón de verosimilitud es $0{,}005/0{,}0051 \approx 1$. En general, una razón de verosimilitud cercana a 1 no da una fuerte preferencia ni por H_0 ni por H_1, por lo que medir nuestra certeza sobre la decisión del procedimiento usando la medida "global" de potencia podría ser engañoso.

Ejemplo 2.16

Considere el siguiente experimento: lanzar una moneda justa y, si la moneda cae en cara, entonces tomar $X \sim \mathrm{N}(\theta, 1)$; de lo contrario, tomar $X \sim \mathrm{N}(\theta, 99)$. Suponga que el resultado del lanzamiento de la moneda es *conocido*. El objetivo es usar X para estimar θ. ¿Qué distribución deberíamos usar para construir un intervalo de confianza, por ejemplo? La varianza marginal de X es $(1 + 99)/2 = 50$. Sin embargo, esto parece una mala representación del error real en X como estimador de θ, ya que en realidad sabemos si X fue muestreado de $\mathrm{N}(\theta, 1)$ o de $\mathrm{N}(\theta, 99)$. Entonces la pregunta es, ¿por qué no usar la varianza "condicional", dado el resultado del lanzamiento de la moneda? Esto es algo intuitivamente natural de hacer, pero esto *no* es lo que el frecuentismo sugiere hacer.

Ejemplo 2.17

Sea (X_{1i}, X_{2i}), $i = 1, \ldots, n$ una muestra bivariada iid de una distribución con densidad $p_\theta(x_1, x_2) = e^{-\theta x_1 - x_2/\theta}$, donde x_1, x_2 y θ son todos positivos. Se puede demostrar que el estadístico suficiente mínimo es $T = (T_1, T_2)$, donde $T_j = \sum_{i=1}^{n} X_{ji}$, $j = 1, 2$. Note que el estadístico suficiente mínimo es bidimensional, mientras que el parámetro es unidimensional. Para estimar θ, una elección razonable es $\hat{\theta} = \{T_2/T_1\}^{1/2}$, el estimador de máxima verosimilitud. Sin embargo, este no es un estadístico suficiente mínimo, por lo que debemos elegir si debemos condicionar o no. Un estadístico auxiliar para condicionar es $A = \{T_1 T_2\}^{1/2}$. Como se discute en Ghosh et al. (2010), la información de Fisher incondicional en T y en $\hat{\theta}$, respectivamente, son

$$I_T(\theta) = \frac{2n}{\theta^2} \quad \text{y} \quad I_{\hat{\theta}}(\theta) = \frac{2n}{\theta^2}\frac{2n}{2n+1};$$

por supuesto, como se esperaba, $I_T(\theta) > I_{\hat{\theta}}(\theta)$. Sin embargo, la información de Fisher condicional es

$$I_{\hat{\theta}|A}(\theta) = I_T(\theta)\frac{K_1(2A)}{K_0(2A)}, \tag{10}$$

donde K_0 y K_1 son funciones de Bessel. Una gráfica de la razón—llamada $r(A)$—en el lado derecho de la ecuación anterior, como función de $A = a$, se muestra en la Figura

2.1. Cuando A es grande, $r(A)$ está cerca de 1, por lo que $I_{\hat{\theta}|A}(\theta) \approx I_T(\theta)$. Sin embargo, si A no es grande, entonces la información condicional puede ser mucho mayor y, dado que una mayor información es "mejor", podemos ver que en este caso hay una ventaja en condicionar.

El propósito de los ejemplos anteriores es destacar las desventajas del frecuentismo puro. Al menos en algunos ejemplos, hay una razón clara para considerar el condicionamiento sobre algo: a veces es evidente sobre qué condicionar (Ejemplo 16) y otras veces no lo es (Ejemplo 17). La inferencia condicional se da cuando las distribuciones muestrales se basan en distribuciones condicionales de estimadores dados los valores observados de un estadístico auxiliar; por ejemplo, en el Ejemplo 14, $|X_1 - X_2|$ es un estadístico auxiliar.

Cuando el estimador es un Estadístico Suficiente completo, el teorema de Basu establece que no hay necesidad de condicionar. Pero en problemas donde el estimador no es un Estadístico Suficiente completo (Ejemplo 17), hay necesidad de condicionar. Existen amplias discusiones en la literatura sobre la inferencia condicional, por ejemplo, Fraser (2004) y Ghosh et al. (2010); Berger (2014) proporciona una discusión más reciente. A pesar de los beneficios de la inferencia condicional, este enfoque no ha permeado realmente la estadística aplicada. Esto se debe a algunas dificultades técnicas adicionales, tanto en la identificación de un estadístico auxiliar adecuado como en la implementación efectiva del condicionamiento. Una revisión aplicada de la inferencia condicional y temas relacionados se encuentra en Brazzale et al. (2007).

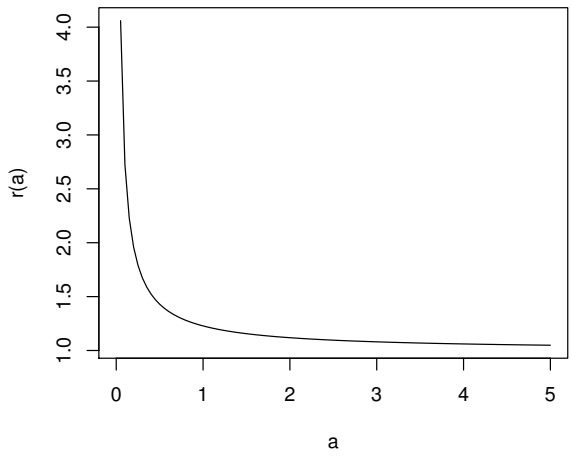


Figura 1: Gráfica de la razón $r(a)$ en el lado derecho de (10) como una función de $A = a$, el estadístico auxiliar.

2.6. Discusión

2.6.1. Modelos lineales generalizados

Una aplicación importante de las distribuciones de la familia exponencial son los llamados *modelos lineales generalizados* (GLM por sus siglas en inglés). Estos modelos generalizan los modelos lineales usuales (por ejemplo, regresión y análisis de varianza) que se presentan en

los cursos introductorios de metodología estadística. Este tema típicamente no aparece en un curso como este—no hay mención de los GLM en Keener (2010)—y creo que la razón de su omisión es que los detalles de la teoría pueden entenderse en el contexto más simple de la familia exponencial, como se discutió en la Sección 2, dejando los detalles específicos de modelado y computación para otros cursos/textos.

Sin embargo, creo que es importante que los estudiantes tengan al menos una exposición mínima a esta aplicación de los modelos de la familia exponencial en este curso teórico, aunque solo sea para que sepan de la existencia de estos temas y puedan leer más por su cuenta o tomar cursos más especializados. Aquí doy una breve explicación de los GLM con algunos ejemplos.

Consideremos un problema con dos variables: Y es llamada la *variable de respuesta* y X la *variable predictora* o *covariable*, donde X es, por ejemplo, d-dimensional. El modelo lineal usual establece que, dado $X = x$, la media de la variable respuesta Y es una función lineal de x, es decir, $\mathbb{E}(Y \mid X = x) = x^\top \beta$, donde β es un parámetro d-dimensional de coeficientes. Si tenemos muestras independientes, es decir, $\{(X_i, Y_i) : i = 1, \ldots, n\}$, entonces el modelo establece que, dado $X_i = x_i$, los Y_i son independientes con media $\mu_i = x_i^\top \beta$, $i = 1, \ldots, n$. Un punto clave es que β es el mismo para cada i; además, este es uno de los modelos independientes pero no iid más comunes que los estudiantes verán. El método de mínimos cuadrados puede utilizarse para estimar β basándose en las observaciones, y esta solución tiene muchas propiedades deseables que no profundizaremos aquí.

Una cuestión a considerar es la siguiente: ¿es este tipo de modelo lineal siempre adecuado? Es decir, ¿debería la media de la distribución de la variable de respuesta (condicionada a la variable predictora X) expresarse como una función lineal del predictor? Como ejemplo, consideremos el caso donde Y sigue una distribución de Poisson o Bernoulli. En ambos casos, la media de la distribución tiene una restricción—en $(0, \infty)$ en un caso y en $(0, 1)$ en el otro—por lo que una función lineal, que no tiene restricciones y puede tomar valores en $(-\infty, \infty)$, podría no ser adecuada. Un modelo lineal generalizado (GLM) puede abordar esto sin extenderse demasiado fuera del marco de un modelo lineal.

Supongamos que las variables de respuesta $Y_1, \ldots, Y_n$ son independientes con densidades

$$p_{\theta_i}(y_i) = h(y_i) e^{\eta(\theta_i) y_i - A(\theta_i)}, \quad i = 1, \ldots, n,$$

que tiene la forma de la familia exponencial descrita en la Sección 2, pero con un parámetro diferente θ_i para cada punto de datos. Supongamos que existe cierta estructura común, en particular, que la media $\mu_i = \mathbb{E}_\theta(Y_i)$ satisface la condición $g(\mu_i) = x_i^\top \beta$ para alguna función uno-a-uno y suave g, llamada *función de enlace*. Cuando la función de enlace es tal que $\eta(\theta_i) = x_i^\top \beta$, se denomina *enlace canónico*. El resultado de esta construcción es una forma general de introducir un modelo lineal efectivo que conecta la variable de respuesta Y_i con la variable predictora X_i, pero evitando las limitaciones de un modelo lineal real.

Como ejemplo rápido, consideremos el caso donde $Y_i \sim \text{Poisson}(\theta_i)$, con $i = 1, \ldots, n$ independientes. Es fácil comprobar que la distribución de Poisson pertenece a la familia exponencial y que $\eta(\theta_i) = \log \theta_i$. Dado que θ_i es también la media de Y_i, si queremos construir un GLM de Poisson con enlace canónico, entonces $g(u) = \log u$, por lo que

$$\theta_i = e^{x_i^\top \beta} \quad \Longleftrightarrow \quad \log \theta_i = x_i^\top \beta.$$

Esta última fórmula explica por qué este GLM de Poisson suele llamarse *modelo log-lineal*.

2.6.2. Un poco más sobre la condicionalidad

En los cursos y libros de teoría estadística, la suficiencia se trata como un aspecto críticamente importante de la inferencia estadística. Aquí quiero argumentar que no hay nada realmente especial sobre los estadísticos suficientes, siempre que se realice una condicionalidad apropiada. El mensaje aquí es que la condicionalidad es un concepto más fundamental que la suficiencia.

Voy a argumentar este punto usando un ejemplo simple. Sea X_1, X_2 una muestra iid de $N(\theta, 1)$. Un estimador razonable de θ es $\bar{X} = (X_1 + X_2)/2$, un estadístico suficiente, cuya distribución muestral es $N(\theta, 1/2)$. Por otro lado, considere el estimador $\hat{\theta} = X_1$, el cual no es un estadístico suficiente. Consideraciones clásicas sugieren que la inferencia basada en X_1 es peor que aquella basada en $\bar{X}$. Sin embargo, considere la distribución muestral condicional de X_1, dado $X_2 - X_1$. Es fácil verificar que

$$X_1 \mid (X_2 - X_1) \sim N\left(\theta + \frac{X_2 - X_1}{2}, 1/2\right),$$

y, por ejemplo, los intervalos de confianza basados en esta distribución condicional son los mismos que aquellos basados en la distribución muestral marginal del estadístico suficiente $\bar{X}$.

Por lo tanto, en este problema, se podría argumentar que no hay nada realmente especial sobre el estadístico suficiente $\bar{X}$, ya que uno puede obtener esencialmente la misma distribución muestral usando otro estadístico no suficiente, siempre que se realice una condicionalidad adecuada. El resultado aquí es más general (ver Ejercicio 20), aunque la continuidad parece ser importante.

La suficiencia, cuando proporciona algo significativo, puede ser conveniente, ya que la condicionalidad no es necesaria, lo que ahorra algo de esfuerzo. Sin embargo, hay casos en los que la suficiencia no proporciona ninguna mejora. Por ejemplo, en el problema de localización de Student-t con grados de libertad conocidos, los datos completos constituyen el estadístico suficiente mínimo. Sin embargo, se puede obtener fácilmente un estimador razonable (equivariante a la localización), como la media muestral, y condicionar en el invariante máximo, un estadístico auxiliar. El punto es que la condicionalidad funciona cuando la suficiencia no lo hace, y, aun cuando la suficiencia funcione, la condicionalidad puede ser igual de buena. Por lo tanto, argumentaría que la condicionalidad es más fundamental que la suficiencia.

2.7. Ejercicios

1. La desigualdad de Hölder es una generalización de la desigualdad de Cauchy-Schwartz.

Sea $1 \leqslant p, q \leqslant \infty$ números tales que $\frac{1}{p} + \frac{1}{q} = 1$. Sean f y g funciones tales que f^p y g^q son integrables respecto a μ. Entonces,

$$\int |fg|\, d\mu \leqslant \left(\int |f|^p \, d\mu \right)^{1/p} \left(\int |g|^q \, d\mu \right)^{1/q}.$$

La desigualdad de Cauchy-Schwartz corresponde al caso $p = q = 2$.

Utilice la desigualdad de Hölder para probar la Proposición 1.

2. Suponga que X sigue una distribución de familia exponencial con densidad

$$p_\theta(x) = h(x) e^{\eta(\theta) T(x) - A(\theta)}.$$

Derive las fórmulas de la media y la varianza:

$$\mathbb{E}_\theta[T(X)] = \frac{A'(\theta)}{\eta'(\theta)}, \quad V_\theta[T(X)] = \frac{A''(\theta)}{[\eta'(\theta)]^2} - \frac{\eta''(\theta) A'(\theta)}{[\eta'(\theta)]^3}.$$

3. Demuestre la ecuación (3), una fórmula para la función generadora de momentos de la familia exponencial.

4. Una variable aleatoria discreta con función de masa de probabilidad

$$p_\theta(x) = a(x)\theta^x / C(\theta), \quad x \in \{0, 1, \dots\}; \quad a(\theta) \geqslant 0, \quad \theta > 0$$

sigue una *distribución de serie de potencias*.

   a) Demuestre que la distribución de serie de potencias es una familia exponencial.

   b) Demuestre que las distribuciones binomial y de Poisson son casos especiales de distribuciones de serie de potencias.

5. *a*) Demuestre la identidad de Stein. Para $X \sim N(\mu, \sigma^2)$, sea φ una función diferenciable con $\mathbb{E}_\theta|\varphi'(X)| < \infty$. Entonces,

$$\mathbb{E}[\varphi(X)(X - \mu)] = \sigma^2 \mathbb{E}[\varphi'(X)].$$

   [**Pista:** Sin pérdida de generalidad, suponga $\mu = 0$ y $\sigma = 1$. Use integración por partes. Necesitará mostrar que $\varphi(x) e^{-x^2/2} \to 0$ cuando $x \to \pm\infty$. También existe un enfoque que usa el teorema de Fubini.]

   b) Sea $X \sim N(\mu, \sigma^2)$. Use la identidad de Stein para encontrar los primeros cuatro momentos, $\mathbb{E}(X^k)$, con $k = 1, 2, 3, 4$.
   [**Pista:** Para $\mathbb{E}(X^k)$, use $\varphi(x) = x^{k-1}$.]

6. Demuestre que una función uno a uno de un estadístico suficiente minimal también es un estadístico suficiente minimal.

7. Suponga que $X_1, \dots, X_n$ son iid $N(\theta, \theta^2)$.

   a) Muestre que $N(\theta, \theta^2)$ tiene la forma de una familia exponencial.

   b) Encuentre el estadístico suficiente minimal para θ.

   c) Muestre que su estadístico suficiente minimal no es completo.

8. La familia Inversa Gaussiana, denotada por $\text{IG}(\lambda, \mu)$, tiene la función de densidad

$$(\lambda/2\pi)^{1/2} \exp\{(\lambda\mu)^{1/2}\} x^{-3/2} \exp\{-(\lambda x^{-1} + \mu x)/2\}, \quad x > 0; \quad \lambda, \mu > 0.$$

   a) Mostrar que $\text{IG}(\lambda, \mu)$ es una familia exponencial.

   b) Mostrar que $\text{IG}(\lambda, \mu)$ es invariante respecto al grupo de transformaciones de escala, es decir, $\mathcal{G} = \{g_c(x) = cx : c > 0\}$.

   c) Sean

$$T_1(X) = n^{-1} \sum_{i=1}^{n} X_i, \quad T_2(X) = \sum_{i=1}^{n} (1/X_i - 1/T_1(X)).$$

   Mostrar que (T_1, T_2) es completo y suficiente.

   d) Mostrar que $T_1 \sim \text{IG}(n\lambda, n\mu)$.

9. Suponga que los pares $(X_1, Y_1), \ldots, (X_n, Y_n)$ son una muestra iid de una distribución normal bivariada, donde

$$\mathbb{E}(X_1) = \mathbb{E}(Y_1) = 0, \quad \mathbb{V}(X_1) = \mathbb{V}(Y_1) = 1, \quad \text{y} \quad \mathbb{E}(X_1 Y_1) = \theta.$$

   Aquí, $\theta \in (-1, 1)$ es la correlación entre X y Y.

   a) Encuentre un estadístico suficiente minimal (bidimensional) para θ.

   b) Demuestre que el estadístico suficiente minimal no es completo.

   c) Sea $Z_1 = \sum_{i=1}^{n} X_i^2$ y $Z_2 = \sum_{i=1}^{n} Y_i^2$. Demuestre que tanto Z_1 como Z_2 son auxiliares, pero que (Z_1, Z_2) no lo es.

10. Este ejercicio describe un enfoque alternativo para encontrar estadísticos suficientes minimales. Está relacionado con el dado en el Teorema 3.

   a) Demuestre el siguiente teorema:

   Considere una familia finita de distribuciones con densidades $p_0, p_1, \ldots, p_K$, todas con el mismo soporte. Entonces

$$T(X) = \left(\frac{p_1(X)}{p_0(X)}, \frac{p_2(X)}{p_0(X)}, \ldots, \frac{p_K(X)}{p_0(X)} \right)$$

   es suficiente minimal.

   b) Demuestre el siguiente teorema: Sea $\mathbb{P}$ una familia paramétrica de distribuciones con soporte común, y sea $\mathbb{P}_0$ un subconjunto de $\mathbb{P}$. Si T es suficiente minimal para $\mathbb{P}_0$ y suficiente para $\mathbb{P}$, entonces es suficiente minimal para $\mathbb{P}$.

   c) Use los dos resultados anteriores para demostrar que, para la familia $\text{Pois}(\theta)$, el estadístico

$$T = \sum_{i=1}^{n} X_i$$

   es suficiente minimal.

   Pista: Elija un subconjunto de dos elementos $\mathbb{P}_0 = \{p_0 = \text{Pois}(\theta_0), p_1 = \text{Pois}(\theta_1)\}$ de $\mathbb{P} = \{\text{Pois}(\theta) : \theta > 0\}$.

11. *a)* Considere una familia de localización con densidades $p_\theta(x) = p(x-\theta)$, donde $\theta \in \mathbb{R}$. Para $X \sim p_\theta$, demuestre que la información de Fisher para θ es

$$I_X(\theta) = \int_{-\infty}^{\infty} \frac{[p'(x)]^2}{p(x)} dx,$$

la cual es independiente de θ.

b) Considere una familia de escala con $p_\theta(x) = p(x/\theta)/\theta$, con $\theta > 0$. Para $X \sim p_\theta$, demuestre que la información de Fisher para θ es

$$I_X(\theta) = \frac{1}{\theta^2} \int \left[\frac{xp'(x)}{p(x)} + 1\right]^2 p(x)\, dx.$$

12. Para cada caso, encuentre la información de Fisher basada en una sola observación X.

(a) Ber(θ).

(b) Pois(θ).

(c) Cau($\theta, 1$).

(d) N($0, \theta$), donde $\theta > 0$ denota la varianza.

13. Para $X_1, \ldots, X_n$ iid, demuestre que $I_{X_1,\ldots,X_n}(\theta) = nI_{X_1}(\theta)$.

14. Sea p_θ una densidad que satisface las condiciones de regularidad FI, y sea $T = T(X_1, \ldots, X_n)$ con $\mathbb{E}_\theta(T) = g(\theta)$. Demuestre que $C_\theta(T, U_\theta) = g'(\theta)$, donde $U_\theta = \sum_{i=1}^n \frac{\partial}{\partial\theta} \log p_\theta(X_i)$ es la función de puntuación.

15. Suponga que la información de Fisher en X sobre θ es $I_X(\theta)$, donde θ es un escalar. Sea ξ una reparametrización suave y uno a uno de θ, y escriba $\tilde{I}_X(\xi)$ para la información de Fisher en X sobre ξ. Demuestre que $\tilde{I}_X(\xi) = \left(\frac{d\theta}{d\xi}\right)^2 I_X(\theta)$. Generalice al caso de vectores θ y ξ.

16. Sea $X \sim \mathcal{N}_n(\theta, \Sigma)$ una única muestra normal n-dimensional; aquí, la matriz de covarianza Σ es conocida, pero el vector θ es desconocido.

(a) Encuentre la matriz de información de Fisher $I_X(\theta)$.

(b) Suponga que $\theta = D\xi$, donde D es una matriz $n \times p$ de rango p, con $p < n$, y ξ es un vector desconocido $p \times 1$. Aquí, D es la *matriz de diseño*. Utilice el resultado en el Ejercicio 15 para encontrar la información de Fisher $\tilde{I}_X(\xi)$ en X sobre ξ.

(La matriz de información en la parte (b) depende de la matriz de diseño D, y la teoría de diseños óptimos busca elegir D para hacer $\tilde{I}_X(\xi)$ lo "más grande posible". Por supuesto, la información de Fisher aquí es una matriz, por lo que se debe definir qué significa que una matriz sea grande, pero la intuición es perfectamente clara.)

17. Sea $\{p_\theta : \theta \in \Theta\}$ una clase de μ-densidades que satisfacen las condiciones de regularidad de la información de Fisher (FI). Mediante la permutación de la diferenciación e integración, derive una aproximación de Taylor de dos términos para la función $\eta \mapsto K(p_\theta, p_\eta)$, para η cercano a θ, donde K es la divergencia de Kullback–Leibler.

18. Let $Y_i \sim \text{Ber}(\theta_i)$, $i = 1, \ldots, n$, independent.

(a) Show that the Bernoulli model is an exponential family with $\eta(\theta) = \log \frac{\theta}{1-\theta}$.

(b) Find the canonical link and write down the formula for θ_i in terms of a predictor variable x_i and a parameter β like in Section 6.1.

(c) Look up the "logistic distribution" (e.g., on **wikipedia**) to see why they call this Bernoulli GLM with canonical link *logistic regression*.

19. Sean X_1, X_2 iid Unif$(\theta - \frac{1}{2}, \theta + \frac{1}{2})$.

(a) Demuestre que $A = (X_2 - X_1)/2$ es un estadístico auxiliar.

(b) Encuentre la distribución de $\bar{X}$, dado $A = a$.

(c) Compare $\mathbb{V}(\bar{X})$ y $\mathbb{V}(\bar{X} \mid A = a)$.

(Ver el Ejemplo 2.2 en Fraser (2004) para una ilustración diferente de este ejemplo: allí se muestra que los intervalos de confianza incondicionales "óptimos"para θ son inútiles, mientras que el intervalo de confianza condicional es muy razonable.)

20. Sean X_1, X_2 iid exponenciales con media θ.

(a) Encuentre la distribución de $\bar{X} = (X_1 + X_2)/2$.

(b) Encuentre la distribución de X_1, dado X_2/X_1.

(c) Compare los intervalos de confianza obtenidos de las distribuciones en (a) y (b).

Capítulo 3: Verosimilitud y Métodos Basados en Verosimilitud

Teoría Estadística Avanzada

Sigla DES124

Prof. Jaime Lincovil

3.1. Introducción

La verosimilitud es, sin duda, uno de los conceptos más importantes en la teoría estadística. Hemos visto el papel que desempeña en la suficiencia, a través del teorema de factorización. Pero, más importante aún, la función de verosimilitud establece una preferencia entre los posibles valores del parámetro dados los datos $X = x$. Es decir, un valor de parámetro θ_1 con mayor verosimilitud es mejor que un valor de parámetro θ_2 con menor verosimilitud, en el sentido de que el modelo P_{θ_1} proporciona un mejor ajuste a los datos observados que P_{θ_2}. Esto lleva naturalmente a procedimientos de inferencia que seleccionan, como estimador puntual, el valor del parámetro que hace que la verosimilitud sea la mayor, o rechazan una hipótesis nula si el valor hipotetizado tiene una verosimilitud demasiado pequeña. La función de verosimilitud también es de considerable importancia en el análisis bayesiano, como veremos más adelante.

Los estimadores y pruebas de hipótesis basados en la función de verosimilitud tienen algunas propiedades generales deseables. En particular, existen aproximaciones ampliamente aplicables en muestras grandes de las distribuciones de muestreo relevantes. Un enfoque principal de este capítulo es la derivación rigurosa de estos importantes resultados. También se ofrece una breve introducción a algunas teorías avanzadas de verosimilitud, incluidas las aproximaciones de orden superior y el uso de pseudo-verosimilitudes, a saber, las verosimilitudes perfiladas y marginales, cuando hay parámetros molestos presentes. Se presentan algunas observaciones sobre la computación relevante en contextos basados en verosimilitud en la Sección 7. La última sección ofrece una breve discusión histórica sobre la verosimilitud y un resultado controvertido, debido a Birnbaum, conocido como el *principio de verosimilitud*.

3.2. Función de verosimilitud

Consideremos una clase de modelos de probabilidad $\{P_\theta : \theta \in \Theta\}$, definidos en el espacio medible $(\mathcal{X}, \mathcal{A})$, absolutamente continuos con respecto a una medida σ-finita dominante μ. En este caso, para cada θ, la derivada de Radon-Nikodym $\left(\frac{dP_\theta}{d\mu}\right)(x)$ es la función de densidad de probabilidad usual para la variable observable X, escrita como $p_\theta(x)$. Para un θ fijo, sabemos

que $p_\theta(x)$ caracteriza la distribución muestral de X, así como la de cualquier estadístico $T = T(X)$.

Pero, ¿cómo usamos o interpretamos $p_\theta(x)$ como una función de θ para un x fijo? Esta es una función especial con su propio nombre: la *función de verosimilitud*.

Definición 3.1

Dado que la muestra es conocida, es decir $X = x$, la función de verosimilitud se construye evaluando x en la función de probabilidad (o de densidad) conjunta $p_\theta(x)$ dejandola como una otra variable función de θ. Es denotada por $L(\theta) = p_\theta(x)$.
La intuición detrás de la elección del nombre es que un θ para el cual $L(\theta)$ es grande es "más probable" de ser el valor verdadero en comparación con un θ' para el cual $L(\theta')$ es pequeño. El término "verosimilitud" fue acuñado por Fisher (1973):

> Lo que ahora ha quedado en evidencia es que el concepto matemático de probabilidad es ... inadecuado para expresar nuestra confianza mental o indiferencia al hacer ... inferencias, y que la cantidad matemática que generalmente parece ser apropiada para medir nuestro orden de preferencia entre diferentes poblaciones posibles no obedece en realidad las leyes de la probabilidad. Para distinguirlo de la probabilidad, he utilizado el término "verosimilitud" para designar esta cantidad; dado que las palabras "verosimilitud" y "probabilidad" se usan indistintamente en el lenguaje común para referirse a ambos tipos de relación.

El punto de Fisher es que $L(\theta)$ es una medida de cuán *plausible* es θ, pero esta medida de plausibilidad es diferente de nuestra comprensión usual de la probabilidad; ver Aldrich (1997) para más información sobre Fisher y la verosimilitud. Mientras que entendemos la probabilidad (densidad) $p_\theta(x)$, para un θ fijo, como un resumen pre-experimental de nuestra incertidumbre sobre dónde caerá X, la función de verosimilitud $L(\theta) = p_\theta(x)$, para un x fijo, proporciona un resumen post-experimental de cuán probable es que el modelo P_θ haya producido el valor observado $X = x$. En otras palabras, la función de verosimilitud proporciona un ordenamiento de los posibles valores del parámetro: aquellos θ con mayor verosimilitud son mejores, en el sentido de que ajustan mejor los datos, que aquellos θ con menor verosimilitud. Por lo tanto, solo la forma de la función de verosimilitud es relevante, no su escala.
La función de verosimilitud es útil en todos los enfoques estadísticos. Ya hemos visto algunos usos de la función de verosimilitud. En particular, el teorema de factorización establece que la (forma de la) función de verosimilitud depende de los datos observados $X = x$ solo a través del estadístico suficiente. La siguiente sección discute algunos usos estándar y otros no tan estándar de la verosimilitud.

3.3. Métodos basados en verosimilitud y teoría de primer orden

3.3.1. Estimación por máxima verosimilitud

Probablemente el uso más familiar de la función de verosimilitud es la estimación por máxima verosimilitud. Dada una clase de modelos potenciales P_θ indexados por Θ, un subconjunto de $\mathbb{R}^d$, observamos $X = x$ y queremos saber cuál modelo es el más probable de haber generado este x. Esto define un problema de optimización, cuyo resultado es:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta), \tag{1}$$

el cual es el *estimador de máxima verosimilitud* (MLE) de θ. Naturalmente, $P_{\hat{\theta}}$ es entonces considerado el modelo más probable, es decir, entre la clase $\{P_\theta : \theta \in \Theta\}$, el modelo $P_{\hat{\theta}}$ proporciona el mejor ajuste a la observación $X = x$. En términos de intuición de "ordenamiento", $\hat{\theta}$ ocupa la posición más alta.

Cuando la función de verosimilitud es suave, el problema de optimización puede reformularse como un problema de búsqueda de raíces. Es decir, el MLE $\hat{\theta}$ puede verse como la solución de la ecuación:

$$\nabla \ell(\theta) = 0, \tag{2}$$

donde ∇ denota el operador gradiente, $\ell = \log L$ es la log-verosimilitud, y el lado derecho es un vector de ceros de dimensión d. La Ecuación (2) se llama la *ecuación de verosimilitud*. Algunas observaciones sobre la resolución de la ecuación de verosimilitud se presentan en la Sección 7. Nuestro enfoque aquí será el estudio de las propiedades teóricas en muestras grandes de las soluciones $\hat{\theta}$ de la ecuación (2).

Una primera propiedad deseable en muestras grandes es la **consistencia**, lo que sugiere que, si n es grande, entonces $\hat{\theta} = \hat{\theta}_n$ estará cerca de θ con alta probabilidad bajo P_θ. Más formalmente, decimos que un estimador $\hat{\theta}_n$, no necesariamente el MLE, es consistente si $\hat{\theta}_n \to \theta$ en probabilidad bajo P_θ, es decir,

$$\lim_{n\to\infty} P_\theta\left(\|\hat{\theta}_n - \theta\| > \varepsilon\right) = 0, \quad \forall \varepsilon > 0.$$

Aquí, $\|\cdot\|$ es una norma adecuada en el espacio Θ, por ejemplo, la norma euclidiana. La definición puede fortalecerse al requerir que $\hat{\theta}_n \to \theta$ con probabilidad P_θ igual a 1, aunque esto es más difícil de demostrar.

Otra propiedad útil en muestras grandes es aquella que describe la distribución límite. Esto (i) da una caracterización exacta de la tasa de convergencia y (ii) permite la construcción de procedimientos estadísticos asintóticamente exactos. Aunque es posible obtener límites no normales, en todos los problemas "estándar" que admiten una distribución límite, esta es normal. Por experiencia previa, sabemos que los estimadores de máxima verosimilitud (MLE) suelen tener una propiedad de normalidad asintótica. A continuación, presentamos una versión de un teorema de este tipo, similar al Teorema 9.14 de Keener (2010), con condiciones dadas en C1–C4. La condición C3 es la más difícil de verificar, pero se cumple en familias exponenciales regulares. Nos enfocamos en el caso unidimensional, pero el mismo teorema, con modificaciones obvias, es válido para θ en dimensión d.

C1. El soporte de P_θ no depende de θ.

C2. Para cada x en el soporte, $f_x(\theta) := \log p_\theta(x)$ es tres veces diferenciable con respecto a θ en un intervalo $(\theta^* - \delta, \theta^* + \delta)$; además, $\mathbb{E}_{\theta^*}|f'_x(\theta^*)|$ y $\mathbb{E}_{\theta^*}|f''_x(\theta^*)|$ son finitos y existe una función $M(x)$ tal que

$$\sup_{\theta \in (\theta^*-\delta, \theta^*+\delta)} |f'''_x(\theta)| \leqslant M(x) \quad \text{y} \quad \mathbb{E}_{\theta^*}[M(X)] < \infty. \tag{3}$$

C3. La esperanza con respecto a P_{θ^*} y la diferenciación en θ^* pueden intercambiarse, lo que implica que la función de puntuación tiene media cero y que la información de Fisher existe y puede evaluarse usando cualquiera de las dos fórmulas conocidas.

C4. La información de Fisher en θ^* es positiva.

Teorema 3.1

Suponga que $X_1, \ldots, X_n$ son iid bajo P_θ, donde $\theta \in \Theta \subseteq \mathbb{R}$. Suponga además que se cumplen C1–C4, y que $\hat{\theta}_n$ es una secuencia consistente de soluciones de la ecuación de verosimilitud. Entonces, para cualquier punto interior θ^*,

$$n^{1/2}(\hat{\theta}_n - \theta^*) \xrightarrow{d} N(0, I(\theta^*)^{-1}),$$

en distribución bajo P_{θ^*}.

Demostración. Definamos $\ell_n(\theta) = n^{-1} \log L_n(\theta)$ como la log-verosimilitud escalada. Dado que θ^* es un punto interior, existe un entorno abierto A de θ^* contenido en Θ. De la consistencia de $\hat{\theta}_n$, el evento $\{\hat{\theta}_n \in A\}$ tiene probabilidad bajo P_{θ^*} convergiendo a 1. Por lo tanto, basta con considerar el comportamiento de $\hat{\theta}_n$ solo cuando está en A, donde la log-verosimilitud es bien comportada; en particular, $\ell'_n(\hat{\theta}_n) = 0$. A continuación, tomemos una aproximación de Taylor de segundo orden de $\ell'_n(\hat{\theta}_n)$ alrededor de θ^*:

$$0 = \ell'_n(\theta^*) + \ell''_n(\theta^*)(\hat{\theta}_n - \theta^*) + \frac{1}{2}\ell'''_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta^*)^2, \quad \text{para } \hat{\theta}_n \text{ cerca de } \theta^*,$$

donde $\tilde{\theta}_n$ está entre $\hat{\theta}_n$ y θ^*. Después de un poco de álgebra simple, obtenemos

$$n^{1/2}(\hat{\theta}_n - \theta^*) = -\frac{n^{1/2}\ell'_n(\theta^*)}{\ell''_n(\theta^*) + \frac{1}{2}\ell'''_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta^*)}, \quad \text{para } \hat{\theta}_n \text{ cerca de } \theta^*.$$

Así, queda por demostrar que el lado derecho de la ecuación anterior tiene la distribución asintóticamente normal establecida. Analicemos el numerador y el denominador por separado.

Numerador. El numerador puede escribirse como

$$n^{1/2}\ell'_n(\theta^*) = n^{1/2} \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log p_\theta(X_i)\Big|_{\theta=\theta^*}.$$

Los sumandos son iid con media cero y varianza $I(\theta^*)$, según nuestras suposiciones sobre el intercambio de derivadas e integrales. Por lo tanto, el Teorema Central del Límite estándar establece que

$$n^{1/2}\ell'_n(\theta^*) \xrightarrow{d} N(0, I(\theta^*))$$

en distribución.

Denominador. El primer término en el denominador converge en probabilidad bajo P_{θ^*} a $-I(\theta^*)$, por la ley fuerte de los grandes números. Falta demostrar que el segundo término en el denominador es despreciable. Para ello, observemos que, por (3),

$$|\ell_n'''(\tilde{\theta}_n)| \leqslant \frac{1}{n}\sum_{i=1}^{n} M(X_i), \quad \text{para } \hat{\theta}_n \text{ cercano a } \theta^*.$$

La ley fuerte de los grandes números, nuevamente, nos dice que la cota superior converge a $\mathbb{E}_{\theta^*}[M(X_1)]$, que es finita. En consecuencia, $\ell_n'''(\tilde{\theta}_n)$ está acotada en probabilidad, y dado que $\hat{\theta}_n - \theta^* \to 0$ en probabilidad bajo P_{θ^*} por suposición, podemos concluir que

$$(\hat{\theta}_n - \theta^*)\ell_n'''(\tilde{\theta}_n) \to 0 \quad \text{en probabilidad bajo } P_{\theta^*}.$$

De aquí, por el Teorema de Slutsky [Ejercicio 7(b)], se deduce que

$$-\frac{n^{1/2}\ell_n'(\theta^*)}{\ell_n''(\theta^*) + \frac{1}{2}(\hat{\theta}_n - \theta^*)\ell_n'''(\tilde{\theta}_n)} \xrightarrow{d} \frac{N(0, I(\theta^*))}{-I(\theta^*)} = N(0, I(\theta^*)^{-1}),$$

en distribución, lo que prueba el resultado deseado. ■

El mensaje clave aquí es que, bajo ciertas condiciones, si n es grande, entonces el estimador de máxima verosimilitud (MLE) $\hat{\theta}$ tiene una distribución muestral cercana a $N(\theta, [nI(\theta)]^{-1})$ bajo P_θ. Para aplicar este resultado, por ejemplo, para construir un intervalo de confianza asintóticamente aproximado, se necesita reemplazar $I(\theta)$ por una cantidad que no dependa del parámetro desconocido. Elecciones estándar son la *información de Fisher esperada* $I(\hat{\theta}_n)$ y la *información de Fisher observada* $-\ell_n''(\hat{\theta}_n)$; ver Efron y Hinkley (1978). Esta última es a menudo preferida, ya que tiene algunas propiedades de *condicionamiento* deseables.

Con la normalidad asintótica del MLE, es posible derivar la distribución asintótica de cualquier función suave del MLE. Esto se conoce como el *teorema delta*, el cual estás invitado a demostrar. El teorema delta es en realidad más general, ya que muestra cómo crear nuevos teoremas del límite central a partir de los existentes; es decir, el teorema delta no es específico de los MLE, etc. Además, el teorema delta ofrece una alternativa—llamada *transformaciones estabilizadoras de varianza* a las reglas de sustitución discutidas anteriormente para eliminar θ de la varianza en la aproximación normal asintótica.

Es posible eliminar el requisito de que la verosimilitud sea tres veces diferenciable si se asume que la segunda derivada existe y cumple con una cierta propiedad de Lipschitz:

La función $\log p_\theta(x)$ es dos veces diferenciable en θ^*, y existe una función $g_\tau(x, \theta)$ tal que, para cada punto interior θ^*,

$$\sup_{\theta:|\theta-\theta^*|\leqslant\tau} \left| \frac{\partial^2}{\partial\theta^2}\log p_\theta(x) - \frac{\partial^2}{\partial\theta^2}\log p_{\theta^*}(x) \right| \leqslant g_\tau(x, \theta^*), \tag{4}$$

con

$$\lim_{\tau\to 0} \mathbb{E}_\theta\{g_\tau(X, \theta)\} = 0, \quad \text{para cada } \theta.$$

Bajo esta suposición, se mantiene el mismo resultado de normalidad asintótica.

Curiosamente, es posible obtener normalidad asintótica bajo una condición aún más débil, a saber, *diferenciabilidad en media cuadrática*, la cual requiere menos que la diferenciabilidad de $\theta \mapsto p_\theta(x)$, aunque los detalles son un poco más técnicos.

3.3.2. Pruebas de razón de verosimilitud

Para dos hipótesis en competencia H_0 y H_1 sobre el parámetro θ, la razón de verosimilitud se usa a menudo para hacer una comparación. Por ejemplo, para $H_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$, la razón de verosimilitud es $L(\theta_0)/L(\theta_1)$, y valores grandes (resp. pequeños) de esta razón indican que los datos x favorecen a H_0 (resp. H_1).

Un problema más difícil y algo más general es cuando la hipótesis nula es $H_0 : \theta \in \Theta_0$ frente a la alternativa $H_1 : \theta \notin \Theta_0$, donde Θ_0 es un subconjunto de Θ. En este caso, se puede definir la razón de verosimilitud como

$$T_n = T_n(X, \Theta_0) = \frac{\sup_{\theta\in\Theta_0} L(\theta)}{\sup_{\theta\in\Theta} L(\theta)}. \tag{5}$$

Teorema 3.2

Suponga que se cumplen las condiciones del Teorema 1. Bajo el esquema descrito en el párrafo anterior, se tiene que

$$W_n \xrightarrow{d} \chi^2(m) \quad \text{en distribución, bajo } P_\theta \text{ con } \theta \in \Theta_0.$$

Demostración. Nos enfocamos aquí solo en el caso $d = m = 1$. Es decir, $\Theta_0 = \{\theta_0\}$ es un conjunto unitario, y queremos conocer la distribución límite de W_n bajo P_{θ_0}. Claramente,

$$W_n = -2\ell_n(\theta_0) + 2\ell_n(\hat{\theta}_n),$$

donde $\hat{\theta}_n$ es el MLE y ℓ_n es la log-verosimilitud. Suponiendo la continuidad de la log-verosimilitud, aplicamos una aproximación de Taylor de dos términos de $\ell_n(\theta_0)$ alrededor de $\hat{\theta}_n$:

$$\ell_n(\theta_0) = \ell_n(\hat{\theta}_n) + \ell_n'(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n) + \frac{\ell_n''(\tilde{\theta}_n)}{2}(\theta_0 - \hat{\theta}_n)^2,$$

donde $\tilde{\theta}_n$ está entre θ_0 y $\hat{\theta}_n$. Como $\ell_n'(\hat{\theta}_n) = 0$, obtenemos

$$W_n = -\ell_n''(\tilde{\theta}_n)(\theta_0 - \hat{\theta}_n)^2 = -\frac{\ell_n''(\tilde{\theta}_n)}{n}\{n^{1/2}(\hat{\theta}_n - \theta_0)\}^2.$$

A partir del Teorema 1, sabemos que

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1})$$

en distribución, cuando $n \to \infty$. Además, en la demostración de ese teorema, mostramos que

$$n^{-1}\ell_n''(\tilde{\theta}_n) \to -I(\theta_0)$$

bajo P_{θ_0} para cualquier $\hat{\theta}_n$ consistente. En efecto, podemos escribir

$$\ell_n''(\tilde{\theta}_n) = \ell_n''(\theta_0) + \ell_n''(\tilde{\theta}_n) - \ell_n''(\theta_0),$$

y se tiene que

$$|\ell_n''(\tilde{\theta}_n) - \ell_n''(\theta_0)| \leqslant \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial^2}{\partial\theta^2}\log p_\theta(X_i)\Big|_{\theta=\tilde{\theta}_n} - \frac{\partial^2}{\partial\theta^2}\log p_\theta(X_i)\Big|_{\theta=\theta_0}\right|.$$

Usando la Condición C2, la cota superior está acotada por

$$n^{-1}\sum_{i=1}^{n} M(X_i)\cdot|\tilde{\theta}_n - \theta_0|,$$

lo cual converge a cero en probabilidad bajo P_{θ_0}, dado que $\hat{\theta}_n$ es consistente. Por lo tanto, $\ell_n''(\tilde{\theta}_n)$ tiene el mismo comportamiento límite que $\ell_n''(\theta_0)$. Finalmente, aplicando el Teorema de Slutsky, obtenemos

$$W_n \to I(\theta_0)N(0, I(\theta_0)^{-1})^2 \equiv N(0,1)^2 \equiv \chi^2(1).$$

■

El **teorema de Wilks** facilita la construcción de una prueba aproximada de tamaño α para H_0 cuando n es grande, es decir, rechazando H_0 si W_n es mayor que $\chi^2_{m,1-\alpha}$, el percentil $100(1-\alpha)$ de la distribución $\chi^2(m)$. La ventaja del teorema de Wilks aparece en casos donde la distribución muestral exacta de W_n es intratable, de modo que una prueba exacta (analítica) de tamaño α no está disponible. A menudo, se puede usar Monte Carlo para encontrar una prueba (ver Sección 3.7), pero el teorema de Wilks proporciona una buena solución y solo requiere el uso de una tabla de chi-cuadrado simple.

También se puede utilizar el resultado del teorema de Wilks para obtener regiones de confianza aproximadas. Definimos

$$W_n(\theta_0) = -2\log T_n(X;\theta_0),$$

donde θ_0 es un valor genérico fijo del parámetro θ en dimensión d, es decir, bajo la hipótesis nula $H_0 : \theta = \theta_0$. Entonces, una región de confianza aproximada al $100(1-\alpha)\,\%$ para θ es

$$\{\theta_0 : W_n(\theta_0) \leqslant \chi^2_{m,1-\alpha}\}.$$

Un aspecto interesante y a menudo pasado por alto del teorema de Wilks es que la distribución nula asintótica no depende de los valores verdaderos de aquellos parámetros que no están especificados bajo la hipótesis nula. Por ejemplo, en un problema de distribución gamma cuyo objetivo es probar si el parámetro de forma es igual a un valor especificado, la distribución nula de W_n no depende del valor verdadero del parámetro de escala.

3.4. Precauciones sobre la teoría de primer orden

Uno podría verse tentado a concluir que las propiedades deseables de los métodos basados en verosimilitud presentados en la sección anterior son universales, es decir, que los estimadores de máxima verosimilitud siempre *funcionan*. Además, basándose en la forma de la varianza asintótica del MLE y su similitud con la cota inferior de Cramér-Rao en el Capítulo 2, es tentador concluir que el MLE es asintóticamente eficiente. Sin embargo, ambas conclusiones son técnicamente *falsas* en general. De hecho, hay ejemplos en los que:

- El MLE no es único o incluso no existe;
- El MLE *funciona* (en el sentido de consistencia), pero las condiciones de la teoría no se cumplen, por lo que la normalidad asintótica falla y
- El MLE no es consistente.

La falta de unicidad o la inexistencia del MLE son obstáculos para su implementación práctica, pero, por alguna razón, no se consideran un problema teórico importante. El caso en el que el MLE funciona pero no es asintóticamente normal tampoco es un problema grave, siempre que se reconozca la naturaleza no regular del problema y se realicen los ajustes necesarios.

El punto más preocupante de esta lista es la inconsistencia del MLE. Dado que la consistencia es una propiedad bastante débil, la inconsistencia del MLE significa que su rendimiento es deficiente y puede producir resultados muy engañosos. El ejemplo más famoso de inconsistencia del MLE, debido a Neyman y Scott (1948), se presenta a continuación.

Ejemplo 3.1: Neyman y Scott, 1948

Sea X_{ij} una familia de variables aleatorias normales independientes,

$$X_{ij} \sim N(\mu_i, \sigma^2), \quad i = 1, \ldots, n, \quad j = 1, 2.$$

El caso de dos niveles j es el más simple, pero el resultado se mantiene para cualquier número fijo de niveles. La idea clave es que X_{i1} y X_{i2} tienen la misma media μ_i, pero hay posiblemente n medias diferentes. El parámetro completo es

$$\theta = (\mu_1, \ldots, \mu_n, \sigma^2),$$

el cual tiene dimensión $n + 1$. Es fácil comprobar que los estimadores de máxima verosimilitud (MLE) están dados por:

$$\hat{\mu}_i = \frac{1}{2}(X_{i1} + X_{i2}), \quad i = 1, \ldots, n.$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_{i1} - X_{i2})^2.$$

Un argumento rutinario (Ejercicio 17) muestra que, cuando $n \to \infty$,

$$\hat{\sigma}^2 \to \frac{1}{2}\sigma^2 \neq \sigma^2$$

en probabilidad, por lo que el MLE de σ^2 es **inconsistente**.

El problema aquí que está causando inconsistencia es que la dimensión del parámetro molesto, es decir, las medias $\mu_1, \ldots, \mu_n$, está aumentando con n. En general, cuando la dimensión del parámetro depende de n, la consistencia del MLE será una preocupación (ver Ejercicio 20), por lo que se debe tener cuidado. La modificación del enfoque básico de máxima verosimilitud puede corregir esto, ver Sección 3.6. Más generalmente, estas deficiencias del MLE estándar proporcionan motivación para modificaciones populares, como la reducción (*shrinkage*), la penalización, etc.

El hecho de que la máxima verosimilitud no sea necesariamente una estrategia confiable en general puede resultar sorprendente. Lucian Le Cam, en un artículo de 1960, escribió[5]:

> El autor está firmemente convencido de que recurrir a la máxima verosimilitud solo es justificable cuando se trabaja con familias de distribuciones que son extremadamente regulares. Los casos en los que los estimadores de máxima verosimilitud son fácilmente obtenibles y han demostrado tener buenas propiedades son extremadamente limitados.

Más tarde, en su libro de 1986, escribió:

> Los términos "verosimilitud" y "máxima verosimilitud" parecen haber sido introducidos por R. A. Fisher, quien también parece ser responsable de gran parte de la propaganda sobre los méritos del método de máxima verosimilitud... Dada la vasta influencia de Fisher, no es sorprendente que la supuesta superioridad del método siga siendo, para muchos, un artículo de fe promovido con fervor religioso. Esta situación persiste, a pesar de una larga acumulación de evidencia que indica que los estimadores de máxima verosimilitud son a menudo inútiles o gravemente engañosos.

3.5. Alternativas a la teoría de primer orden

3.5.1. Bootstrap

Bootstrap está diseñado para obtener una distribución muestral para un estimador, lo cual puede usarse para la construcción de pruebas y regiones de confianza, basado en un solo conjunto de datos. Esta es una herramienta muy popular, probablemente debido a su simplicidad. El primer artículo sobre bootstrap es Efron (1979) y el Capítulo 29 en DasGupta (2008) es un buen resumen de la literatura hasta ese punto. Desde entonces, han surgido muchas técnicas y resultados sofisticados de bootstrap, pero aquí daré solo la configuración más simple, para mantener los conceptos claros.

Recordemos que si tenemos un estimador $\hat{\theta}_n$, este se basa en un solo conjunto de datos. La distribución muestral de $\hat{\theta}_n$, que es relevante para la construcción de intervalos de confianza y pruebas, se basa en muchas muestras y muchos valores de $\hat{\theta}_n$, por lo que obviamente no está disponible para nosotros. La teoría asintótica de la sección anterior se enfoca en proporcionar una aproximación simple a esa distribución muestral desconocida. Bootstrap, en cambio, trata de producir una distribución muestral aproximada numéricamente mediante el remuestreo de los datos disponibles.

Sea P_θ la distribución de una variable observable X. Observamos copias iid $X_1, \ldots, X_n$ de P_θ, y obtenemos un estimador $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$. Para aprender la distribución muestral de $\hat{\theta}_n$, necesitaríamos muchos conjuntos de la muestra $(X_1, \ldots, X_n)$. La idea básica detrás de bootstrap es muestrear, con reemplazo, a partir del único conjunto de datos disponible. Escribimos dicha muestra remuestreada como X_i^*, $i = 1, \ldots, n$; en este caso, es posible que haya repeticiones, ya que el muestreo se realiza con reemplazo. Basado en $(X_1^*, \ldots, X_n^*)$, calculamos

$$\hat{\theta}_n^* = \hat{\theta}(X_1^*, \ldots, X_n^*).$$

Repetimos este proceso B veces, obteniendo una *muestra bootstrap* de B valores de $\hat{\theta}_n$, que escribiré como $\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,B}$. Luego, la afirmación es que la distribución de esta muestra bootstrap es una buena aproximación de la distribución muestral real de $\hat{\theta}_n$. Por ejemplo, podemos obtener un intervalo de confianza aproximado del 90 % para θ tomando los percentiles 5 y 95 de la muestra bootstrap.

Esto es muy fácil de hacer, ya que el remuestreo puede realizarse rápidamente con una computadora. La cuestión es si funciona. La base de la afirmación de que la distribución bootstrap es una buena aproximación de la distribución muestral es de naturaleza asintótica. De manera aproximada, la afirmación es que la distribución bootstrap y la distribución muestral convergen cuando $n \to \infty$. Para entender esto a un nivel intuitivo, recordemos el teorema fundamental de la estadística, que establece que la función de distribución empírica converge uniformemente, casi seguramente, a la función de distribución verdadera. Entonces, el paso de remuestreo es simplemente un muestreo iid de la distribución empírica.

Así, si la distribución empírica está cerca de la distribución verdadera, entonces muestrear de la primera debería ser equivalente a muestrear de la segunda, justificando así la afirmación. Curiosamente, algunos métodos de bootstrap, quizás más sofisticados que lo presentado aquí, tienen propiedades de precisión de orden superior .ªutomáticas"(por ejemplo, DiCiccio y Romano 1995).

A pesar de la simplicidad de bootstrap y de la amplia gama de aplicaciones en las que funciona, no es una herramienta que funcione universalmente. Es decir, existen casos conocidos en los que bootstrap falla, por lo que no se puede usar a ciegas. Hay herramientas disponibles para corregir bootstrap cuando falla, pero estas modificaciones no son intuitivas. Véase la discusión en DasGupta (2008, Cap. 29) y las referencias allí mencionadas para más detalles sobre bootstrap.

3.5.2. Monte Carlo y funciones de plausibilidad

El enfoque en la teoría asintótica se debe, en gran medida, a la tradición—cuando la estadística estaba en desarrollo, no había computadoras disponibles, por lo que solo eran posibles las aproximaciones analíticas asintóticas. La tecnología ha cambiado drásticamente desde entonces, por lo que una pregunta interesante es si todavía necesitamos aproximaciones asintóticas. Es decir, ¿por qué no realizar todos los cálculos exactamente utilizando la potencia computacional disponible? No estoy sugiriendo que las aproximaciones asintóticas no sean útiles, sino que es importante recordar para qué sirven realmente, es decir, para ayudar a simplificar cálculos que son demasiado difíciles de realizar exactamente.

Martin (2015b) propuso un enfoque basado en obtener una estimación de Monte Carlo (véase la Sección 7.2) de la función de distribución del estadístico de razón de verosimilitud; el desafío es que Monte Carlo generalmente necesita ejecutarse para muchos valores del parámetro. Se puede demostrar que este enfoque, que define una *función de plausibilidad*, proporciona inferencia exacta sin asintótica. Es decir, los *intervalos de plausibilidad* del 95 % tienen una probabilidad de cobertura exactamente igual a 0.95. Entonces, la pregunta es, si se tienen los recursos para realizar estos cálculos (que no son costosos en problemas pequeños), ¿por qué no hacerlo y evitar cualquier aproximación asintótica? Ese artículo muestra algunos ejemplos para demostrar la eficiencia del método, pero hay problemas teóricos y computacionales que deben resolverse.

También debo mencionar que, aunque el artículo antes mencionado no lo dice explícitamente, este enfoque tiene algunas conexiones con el marco del *modelo inferencial* (IM) (por ejemplo, Martin y Liu 2013, 2015a,c). La idea es que este enfoque define un IM sin especificación completa del modelo de muestreo, lo que simplifica mucho las cosas, pero aparentemente sin sacrificar demasiada eficiencia; véase Martin (2015a).

3.6. Sobre la teoría avanzada de verosimilitud

3.6.1. Visión general

Uno de los objetivos de la teoría avanzada de verosimilitud es manejar problemas con parámetros molestos. Por ejemplo, supongamos que θ se divide en un par de subvectores $\theta = (\psi, \lambda)$, donde ψ es el parámetro de interés y λ es un parámetro molesto, es decir, un parámetro desconocido pero que no es de interés. Nos gustaría hacer inferencia sobre ψ en presencia de un λ desconocido.

Este es un problema desafiante porque la función de verosimilitud depende tanto de ψ como de λ, y no es inmediatamente claro cómo eliminar λ. Por ejemplo, si el modelo es suficientemente regular, entonces $\hat{\theta}$ es asintóticamente normal. Se podría proceder a hacer inferencia sobre ψ extrayendo $\hat{\psi}$ de $\hat{\theta}$ y el bloque correspondiente de la matriz de covarianza asintótica. Sin embargo, dicha matriz de covarianza generalmente dependerá de todo θ, por lo que la pregunta es si sustituir $\hat{\lambda}$ en esa matriz de covarianza es una forma suficiente de eliminar λ—no lo creo.

Un segundo objetivo de la teoría avanzada de verosimilitud es obtener aproximaciones más precisas que las obtenidas mediante la normalidad asintótica del estimador de máxima verosimilitud (MLE) o el teorema de Wilks. La herramienta básica que impulsa las demostraciones de estos dos resultados es una aproximación de Taylor de dos términos del log-verosimilitud. Si tomamos una aproximación de orden superior, manejando cuidadosamente los términos residuales, a menudo podemos obtener aproximaciones asintóticas más precisas. Esto conducirá a pruebas más precisas y/o regiones de confianza más ajustadas.

Los detalles de estas aproximaciones de orden superior están fuera de nuestro alcance, por lo que solo daré un ejemplo de esto. Algunas referencias buenas y accesibles incluyen Young y Smith (2005, Cap. 9) y Brazzale et al. (2007, Cap. 2); una buena visión general de este tipo de asintótica avanzada se encuentra en Reid (2003).

3.6.2. Verosimilitud Modificada

Escribimos $\theta = (\psi, \lambda)$, donde ψ es el parámetro de interés y λ es un parámetro molesto desconocido. ¿Cómo construir una función de verosimilitud solo para ψ? Básicamente, hay tres técnicas posibles; la primera de ellas probablemente la hayas visto en un curso introductorio de teoría estadística, en el contexto de pruebas de razón de verosimilitud.

Verosimilitud perfilada

Supongamos, por el momento, que ψ es conocido y solo λ es desconocido. En este caso, podríamos encontrar el MLE de λ, dado este valor conocido de ψ, el cual denotamos por $\hat{\lambda}_\psi$. Esto se puede hacer para cualquier valor de ψ, por lo que podemos escribir

$$L_p(\psi) = L(\psi, \hat{\lambda}_\psi),$$

lo que se conoce como la *verosimilitud perfilada* de ψ. Esta función, en general, puede tratarse como una función de verosimilitud genuina. Por ejemplo, el teorema de Wilks establece que -2 veces el logaritmo de la razón de verosimilitud perfilada es asintóticamente chi-cuadrado bajo un H_0 dado. Algunos usos no asintóticos de la razón de verosimilitud perfilada se discuten en Martin (2015b). Sin embargo, se debe tener precaución en cuanto al orden: en el ejemplo de Neyman-Scott mencionado anteriormente, el estimador de máxima verosimilitud perfilado $\hat{\sigma}^2$ es *inconsistente*!

Verosimilitud marginal y condicional

Gran parte del material en esta sección está tomado de la Sección 2.4 de Boos y Stefanski (2013). Para los datos X, sea (S, T) una función uno a uno de X; por supuesto, la función no debe depender del parámetro $\theta = (\psi, \lambda)$. Entonces, en términos de verosimilitudes, con un ligero abuso de notación, podemos escribir $p_\theta(X) = p_\theta(S, T)$. Por supuesto, el lado derecho, que es una densidad conjunta de T y S, puede factorizarse en un producto de una densidad marginal y una densidad condicional. Supongamos que se cumple alguna de las siguientes factorizaciones:

$$p_\theta(T, S) = p_\theta(T \mid S) p_\psi(S) \tag{6}$$

$$p_\theta(T, S) = p_\psi(T \mid S) p_\theta(S). \tag{7}$$

Consideramos ambos casos por separado.

- En el primer caso, la Ecuación (6), la distribución marginal de S depende solo del parámetro de interés, por lo que podemos tomarla como una verosimilitud "modificada" o "pseudo":

  $$L_m(\psi) = p_\psi(S).$$

  Esto se denomina *verosimilitud marginal* ya que se basa en la distribución marginal de una función $S = S(X)$. Sin embargo, hay que notar que esta no es una verosimilitud real porque se ha descartado información relevante para ψ, contenida en la parte condicional

$p_\theta(T \mid S)$. Sin embargo, la esperanza es que la eliminación del parámetro molesto λ mediante marginalización conduzca a ventajas que superen la pérdida de información.

- En el segundo caso, la Ecuación (7), la distribución condicional de T dado S no depende de λ, por lo que podemos tomarla como una verosimilitud modificada o pseudo:

$$L_c(\psi) = p_\psi(T \mid S).$$

De nuevo, esto no es una verosimilitud real, ya que se ha descartado algo de información, pero la eliminación del parámetro molesto tiene valor.

En el ejemplo de Neyman–Scott, consideremos la transformación $X = (X_{ij})$ a

$$S_i = 2^{-1/2}(X_{i1} - X_{i2}) \quad \text{y} \quad T_i = 2^{-1/2}(X_{i1} + X_{i2}), \quad i = 1, \ldots, n.$$

Entonces, las siguientes propiedades son fáciles de verificar:

- La distribución marginal de $S = (S_1, \ldots, S_n)$ no depende de $(\mu_1, \ldots, \mu_2)$;
- $S_1, \ldots, S_n$ son iid $N(0, \sigma^2)$;
- S y T son independientes.

Por lo tanto, condicionar no tiene efecto, por lo que la verosimilitud marginal/condicional para el parámetro de interés σ^2, basada solo en S, es

$$L_m(\sigma^2) \propto (\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n S_i^2}.$$

Es sencillo verificar que el estimador de máxima verosimilitud marginal/condicional de σ^2 es $\hat{\sigma}^2 = (2/n)\sum_{i=1}^n S_i^2 = 2\sigma^2$. Dado que $\hat{\sigma}^2 \to \frac{1}{2}\sigma^2$, es claro que $\hat{\sigma}^2$ es consistente.

El desafío principal en la implementación de un enfoque de verosimilitud marginal o condicional es encontrar los estadísticos adecuados (S, T). Desafortunadamente, no existen estrategias generales para encontrarlos; la experiencia es la única guía.

Probablemente, la única estrategia semi-general que se puede aplicar para obtener una verosimilitud condicional es la siguiente, la cual aplica a ciertos modelos de familia exponencial. Supongamos que la función de densidad de X tiene la forma:

$$p_\theta(x) \propto \exp\{\langle \psi, T(x) \rangle + \langle \lambda, S(x) \rangle - A(\psi, \lambda)\}.$$

Entonces, es bastante fácil ver que la distribución condicional de $T = T(X)$, dado $S = S(X)$, será de forma de familia exponencial y no dependerá del parámetro molesto λ. Boos y Stefanski (2013), Sección 2.4.6, presentan una muy buena aplicación de esta estrategia en el contexto de la regresión logística; véase también la Sección 5 de Bølviken y Skovlund (1996).

3.6.3. Expansiones asintóticas

Las dos herramientas técnicas principales son las expansiones de *Edgeworth* y *saddlepoint*. Ambas se basan en funciones generadoras de cumulantes—el logaritmo de la función generadora de momentos. La idea es aproximar la densidad $p_n(s)$ de la suma normalizada

$$S_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}},$$

donde $X_1, \ldots, X_n$ son iid con media μ y varianza σ^2. El teorema del límite central establece que $p_n(s) \to \varphi(s)$, la densidad normal estándar, cuando $n \to \infty$. Estas expansiones están diseñadas para obtener una aproximación más precisa de $p_n(s)$ para n finito. No discutiremos los detalles generales aquí—véase, por ejemplo, DasGupta (2008)—solo una aplicación.

Para una muestra dada, supongamos que el estadístico suficiente mínimo S para θ se puede expresar como (T, A), donde T es el MLE y A es un estadístico auxiliar. Dado que A es auxiliar, es natural basar la inferencia en la distribución muestral de T, dado $A = a$, donde a es el valor observado de A. A pesar de que esto es un procedimiento natural, generalmente no es fácil calcular esta distribución condicional. La fórmula p^*, basada en las expansiones asintóticas anteriores, proporciona una muy buena aproximación a esta distribución condicional.

Escribamos la función de log-verosimilitud como $\ell(\theta; t, a)$ y la matriz de información de Fisher observada $J(\theta; t, a)$. La fórmula p^* es entonces

$$p^*_\theta(t \mid a) = c(\theta, a)\, |\det J(t; t, a)|^{1/2}\, e^{\ell(\theta_{t,a}) - \ell(t,a)},$$

donde $c(\theta, a)$ es una constante de normalización que no depende de t. Entonces, el resultado de aproximación afirmado es que, para cualquier t, cuando $n \to \infty$,

$$p_\theta(t \mid a) = p^*_\theta(t \mid a)\{1 + O(n^{-1})\};$$

es decir, la densidad condicional exacta $p_\theta(t \mid a)$ de T, dado $A = a$, es igual a $p^*_\theta(t \mid a)$ módulo un error que desaparece a razón de n^{-1}. Esto proviene de una expansión *saddlepoint* que incluye cotas de aproximación generales. Para algunos problemas, incluyendo problemas de transformación de grupos, la fórmula p^* es exacta.

También existen muy buenas aproximaciones de la función de distribución de un estadístico, por ejemplo, la aproximación r^* explicada en Reid (2003). Desafortunadamente, esto es demasiado técnico para considerarlo aquí. Sin embargo, existen otros trucos para mejorar las aproximaciones asintóticas, como la corrección de Bartlett, que es relativamente fácil de usar.

3.7. Un poco sobre computación

3.7.1. Optimización

El método de Newton es una herramienta simple y poderosa para realizar optimización o, más precisamente, búsqueda de raíces. Deberías estar familiarizado con este método de un curso de cálculo. La idea se basa en el hecho de que, localmente, cualquier función diferenciable puede ser adecuadamente aproximada por una función lineal. Esta función lineal se usa luego para definir un procedimiento recursivo que, bajo condiciones adecuadas, eventualmente encuentra la solución deseada.

Recordemos la ecuación de verosimilitud (3.2). Entonces, el MLE es una solución de esta ecuación, es decir, una raíz del gradiente de la función de log-verosimilitud. Supongamos que el gradiente $\nabla\ell(\theta)$ también es diferenciable, y sea $D(\theta)$ la matriz de derivadas, es decir,

$$D(\theta)_{ij} = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta).$$

Supongamos que $D(\theta)$ es no singular para todo θ. La idea detrás del método de Newton es la siguiente. Escogemos una estimación inicial, digamos $\theta^{(0)}$ del MLE $\hat{\theta}$. Ahora aproximamos $\nabla\ell(\theta)$ mediante una función lineal:

$$\nabla\ell(\theta) = \nabla\ell(\theta^{(0)}) + D(\theta^{(0)})(\theta - \theta^{(0)}) + \text{error}.$$

Ignoramos el error, resolvemos para θ y llamamos a la solución $\theta^{(1)}$:

$$\theta^{(1)} = \theta^{(0)} - D(\theta^{(0)})^{-1}\nabla\ell(\theta^{(0)}).$$

Si $\theta^{(0)}$ está cerca de la solución de la ecuación de verosimilitud, entonces también lo estará $\theta^{(1)}$ (¡haz un dibujo!). La idea es iterar este proceso hasta que las soluciones converjan. Por lo tanto, el método consiste en elegir un valor inicial "razonable" $\theta^{(0)}$ y, en la iteración $t \geqslant 0$, definir:

$$\theta^{(t+1)} = \theta^{(t)} - D(\theta^{(t)})^{-1}\nabla\ell(\theta^{(t)}).$$

Luego, detén el algoritmo cuando t sea grande y/o $\|\theta^{(t+1)} - \theta^{(t)}\|$ sea pequeño.

Existen muchas herramientas disponibles para la optimización. El método de Newton descrito anteriormente es solo un enfoque simple. Afortunadamente, hay buenas implementaciones de estos métodos disponibles en software estándar. Por ejemplo, la rutina `optim` en R es una herramienta muy poderosa y fácil de usar para la optimización genérica. Para problemas que tienen una forma específica, en particular, problemas que pueden escribirse en una forma de "variable latente", existe una herramienta muy ingeniosa llamada el algoritmo EM (por ejemplo, Dempster et al. 1977) para maximizar la verosimilitud. La Sección 9.6 de Keener (2010) ofrece alguna descripción de este método.

Un resultado interesante e inesperado es que, a veces, la optimización puede usarse para realizar integración. El resultado técnico al que me refiero es la *Aproximación de Laplace*, y algunos comentarios adicionales sobre esto se harán en el Capítulo 4 sobre métodos bayesianos.

3.7.2. Integración de Monte Carlo

En pruebas de hipótesis, supongamos que $H_0 : \theta = \theta_0$, es decir, la hipótesis nula proporciona una especificación completa del parámetro. En este caso, es sencillo derivar pruebas exactas utilizando Monte Carlo.

Es decir, generamos muchas muestras de datos $(X_1, \ldots, X_n) \overset{\text{iid}}{\sim} P_{\theta_0}$ y, para cada conjunto de datos, calculamos el correspondiente W_n, o cualquier otro estadístico de prueba. Una prueba de tamaño-α establece que

$$\text{Rechazar } H_0 \text{ si } W_n > k_\alpha,$$

donde k_α depende de la distribución nula de W_n. Ahora, elegimos k_α como el percentil $100(1-\alpha)$ de la muestra Monte Carlo de los W_n. Esto es fácil de hacer.

El desafío es intentar hacer algo similar cuando la hipótesis nula deja algunos componentes de θ sin especificar. En este caso, no está claro de qué distribución se debe muestrear en el paso de Monte Carlo. Por ejemplo, si $\theta = (\theta_1, \theta_2)$ y la hipótesis nula especifica el valor $\theta_{1,0}$ de θ_1, entonces ¿qué valor de θ_2 se debe usar para simular? En general, la distribución nula de W_n en este caso, al menos para n finito, dependerá del valor particular de θ_2, por lo que esta es una cuestión importante. En algunos casos, se puede demostrar que la distribución de W_n no depende del parámetro no especificado, pero esto puede ser difícil de lograr. Martin (2015b) tiene algunos comentarios relevantes sobre esto.

3.8. Discusión

La verosimilitud, los métodos basados en la verosimilitud y los resultados deseables en muestras grandes presentados anteriormente han sido importantes para el desarrollo de la práctica estadística. Sin embargo, surge una pregunta interesante: ¿es la función de verosimilitud fundamental para la estadística? Entendemos que, en cierto sentido, la función de verosimilitud contiene toda la información relevante en los datos sobre el parámetro desconocido. Pero los métodos estadísticos descritos anteriormente (por ejemplo, la prueba de razón de verosimilitud) utilizan información más allá de lo contenido en la función de verosimilitud. En particular, la distribución muestral del estadístico relevante es necesaria para elegir el punto de corte de la prueba. Que los resultados de un procedimiento estadístico dependan de aspectos distintos a los datos observados es algo controvertido.

Aceptar completamente la afirmación: la verosimilitud contiene toda la información relevante referente al parámetro contenida en los datos, uno debe estar dispuesto a aceptar

- **Principio de verosimilitud**: dos conjuntos de datos que producen la misma función de verosimilitud (hasta una constante de proporcionalidad) deberían producir los mismos resultados respecto al parámetro de interés.

Los métodos clásicos que dependen de distribuciones muestrales, incluidas las pruebas de razón de verosimilitud discutidas anteriormente, violan el principio de verosimilitud. Además del principio de verosimilitud, hay otros dos principios estadísticos que han ganado relevancia:

- **Principio de suficiencia**: Cualquier par de conjuntos de datos que admitan los mismos estadísticos suficientes (mínimos) deberían conducir a las mismas conclusiones sobre θ.
- **Principio de condicionalidad**: Si se consideran dos experimentos diferentes y la elección entre ellos es aleatoria, y dicha aleatorización no depende de θ, entonces las conclusiones sobre θ deberían basarse solo en el experimento realmente realizado.

Estos dos principios son difíciles de refutar y, históricamente, han sido aceptados como principios razonables. (Por supuesto, estoy omitiendo algunos detalles no triviales para resumir la idea principal).

Existe un famoso resultado de Birnbaum (1962), posiblemente el más controversial en toda la estadística, que establece que la inferencia estadística que sigue los principios de suficiencia y condicionalidad también debe seguir el principio de verosimilitud. Esto sugiere que aquellos métodos estadísticos basados en distribuciones muestrales (por ejemplo, las pruebas de razón de verosimilitud), que violan el principio de verosimilitud, también deben violar el principio de suficiencia o el principio de condicionalidad. En resumen: el resultado de Birnbaum implica que los métodos frecuentistas son "ilógicos" en este sentido específico.

El resultado de Birnbaum ha tenido un efecto significativo, en particular, en el desarrollo y aceptación de los métodos bayesianos. El hecho es que el único enfoque estadístico conocido que satisface el principio de verosimilitud es el enfoque bayesiano (con una priori subjetiva). Por ejemplo, Jimmie Savage, en su discusión sobre el artículo de Birnbaum, escribe:

> Yo mismo comencé a tomar ... la estadística bayesiana ... en serio solo a través del reconocimiento del principio de verosimilitud.

Es decir, si no fuera por el resultado de Birnbaum sobre el principio de verosimilitud, Savage nunca habría tomado en serio el enfoque bayesiano. De este modo, el resultado de Birnbaum es, indirectamente, responsable de gran parte del desarrollo de la estadística bayesiana.

Mientras lees esto, deberías sentirte un poco incómodo: *¡¿esos métodos clásicos enseñados en los cursos introductorios de estadística violan algunos principios lógicos?!* Ha habido siempre dudas sobre la validez del resultado de Birnbaum y, recientemente, se ha demostrado que es, de hecho, **falso**! Véase Evans (2013) y Mayo (2014). Además de liberar a la estadística de las restricciones impuestas por la afirmación de Birnbaum, estos desarrollos abren la puerta a nuevas ideas sobre los fundamentos de la estadística; véase Martin y Liu (2014).

3.9. Ejercicios

1. Sea $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$. Encuentra el MLE de (μ, σ^2).
2. Sea $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Gamma}(\theta, 1)$, donde el parámetro de forma $\theta > 0$ es desconocido.
   a) No existe una expresión en forma cerrada para el MLE $\hat{\theta}$. Escribe código en R para encontrar el MLE numéricamente; intenta hacerlo sin usar rutinas de optimización incorporadas.

b) Simula 1000 conjuntos de datos (con $\theta = 7$ y $n = 10$) y, para cada uno, calcula el MLE. Resume tu simulación con un histograma que describa la distribución muestral del MLE.

c) ¿Parece que la distribución muestral es aproximadamente normal?

3. Los estimadores de máxima verosimilitud tienen una propiedad deseable de *invarianza*. Es decir, si $\hat{\theta}$ es el MLE de θ, y $\eta = g(\theta)$ es alguna transformación, entonces el MLE de η es $\hat{\eta} = g(\hat{\theta})$. Explica la intuición detrás de este hecho.

4. Para datos iid $X_1, \ldots, X_n$, sea $\ell_n(\theta)$ la log-verosimilitud.

   - Usa la desigualdad de Jensen y la ley de los grandes números para argumentar que, para cualquier a distinto de cero,

   $$\ell_n(\theta^* + a) - \ell_n(\theta^*) < 0$$

   para todo n suficientemente grande con probabilidad 1 bajo P_{θ^*}.
   - Usa esto para argumentar la consistencia de una secuencia de soluciones de la ecuación de verosimilitud. **[Pista:** Fija $\varepsilon > 0$ y aplica (a) con $a = \varepsilon$ y $a = -\varepsilon$.]

5. Supón que la densidad p_θ de X satisface las condiciones del teorema de factorización del Capítulo 2. Usa el Teorema 3 para demostrar que si el MLE $\hat{\theta} = \hat{\theta}(X)$ es único y suficiente, entonces también es mínimamente suficiente.

6. Demuestra la siguiente versión del *Teorema del Mapeo Continuo*:

   > Sea $X, \{X_n : n \geqslant 1\}$ una secuencia de variables aleatorias que toman valores en un espacio métrico $(\mathbb{X}, |\cdot|)$, como $\mathbb{R}^d$. Sea g una función continua en todo su dominio que mapea $\mathbb{X}$ a otro espacio métrico $\mathbb{X}'$. Demuestra que si $X_n \to X$ en probabilidad, entonces $g(X_n) \to g(X)$ en probabilidad.

   Observaciones: (i) g no necesita ser continua en todo su dominio, es suficiente que X esté en el conjunto de continuidad de g con probabilidad 1, y (ii) el mismo resultado se cumple si se reemplaza la convergencia en probabilidad por convergencia en distribución o convergencia con probabilidad 1.

7. *a*) Considera dos secuencias de variables aleatorias, X_n y Y_n, tales que $X_n \to X$ y $Y_n \to c$, ambas en distribución, donde X es una variable aleatoria y c es una constante. Demuestra que $(X_n, Y_n) \to (X, c)$ en distribución.

   [Pista: Por definición, una secuencia de variables aleatorias X_n converge en distribución a X si $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ para todas las funciones f acotadas y continuas.**]**

   b) Usa (a) y el Teorema del Mapeo Continuo para demostrar el *Teorema de Slutsky*:

   > Si $A_n \to a$ y $B_n \to b$, ambos en probabilidad, y $X_n \to X$ en distribución, entonces $A_n + B_n X_n \to a + bX$ en distribución.

   [Pistas: (i) La convergencia en distribución y la convergencia en probabilidad son equivalentes cuando el límite es una constante, y (ii) para aplicar la parte (a), piensa en (A_n, B_n) como una única secuencia con límite constante (a, b).**]**

   c) Como caso especial del Teorema de Slutsky, demuestra que si $Y_n \to Y$ en distribución y B_n es un evento con $P(B_n) \to 1$, entonces $Y_n I_{B_n} + Z_n I_{B_n^c} \to Y$ para cualquier secuencia Z_n de variables aleatorias.

8. Demuestra el *Teorema Delta*:

   Para variables aleatorias T_n, supongamos que

   $$n^{1/2}(T_n - \theta) \to N(0, v(\theta))$$

   en distribución, donde $v(\theta)$ es la varianza asintótica. Sea $g(\cdot)$ una función diferenciable en θ, con $g'(\theta) \neq 0$. Entonces,

   $$n^{1/2}\{g(T_n) - g(\theta)\} \to N(0, v_g(\theta))$$

   en distribución, donde $v_g(\theta) = [g'(\theta)]^2 v(\theta)$.

9. El teorema delta en el ejercicio anterior asume que g' existe y es distinto de cero en θ. ¿Qué sucede cuando $g'(\theta) = 0$?

   Teorema. Supón que $g'(\theta) = 0$, g'' es continua y $g''(\theta) \neq 0$. Entonces, existe una secuencia de constantes c_n y una función $h(\cdot)$ tales que

   $$c_n h(\theta)[g(\hat{\theta}_n) - g(\theta)] \to \text{ChiSq}(1)$$

   en distribución, cuando $n \to \infty$.

   a) Demuestra el teorema e identifica las constantes particulares c_n y $h(\theta^*)$.
   [Pista: Usa la aproximación de Taylor cuadrática de $g(\hat{\theta}_n)$ en θ y el teorema del mapeo continuo.**]**

   b) Da un ejemplo de un modelo iid, un estimador $\hat{\theta}_n$, un valor verdadero θ y una función g tal que la aproximación chi-cuadrado anterior sea exacta.
   [Pista: $N(0,1)^2 = \text{ChiSq}(1)$.**]**

10. Sea $\hat{\theta}_n$ una secuencia de estimadores tal que

    $$\sqrt{n}(\hat{\theta}_n - \theta) \to N(0, v(\theta))$$

    en distribución, donde $v(\theta) > 0$ es la función de varianza (asintótica).

    a) Una *transformación estabilizadora de varianza* es una función g tal que la varianza asintótica de $g(\hat{\theta}_n)$ no depende de θ. Usa el teorema delta para encontrar la condición que debe cumplir una función g para ser estabilizadora de varianza.

    b) Sea λ un número real fijo y supón que $v(\theta) = \theta^{2(1-\lambda)}$, $\theta > 0$. Usa la condición suficiente derivada en la Parte (a) para mostrar que

    $$g(\theta) = \begin{cases} \frac{\theta^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0, \\ \log\theta & \text{si } \lambda = 0, \end{cases}$$

    es una transformación estabilizadora de varianza. (Esta es la *transformación de Box–Cox.*)

    c) Supón que $X_1, \ldots, X_n$ son iid con densidad

    $$p_\theta(x) = \frac{1}{\theta} e^{-x/\theta}, \quad x > 0, \theta > 0.$$

    El estimador de máxima verosimilitud $\hat{\theta}_n$ es asintóticamente normal, con una función de varianza $v(\theta)$ de la forma dada en la Parte (b). Encuentra los valores correspondientes de λ y g.

d) En el contexto de la Parte (c), usa la distribución asintótica de $g(\hat{\theta}_n)$ para encontrar un intervalo de confianza asintóticamente correcto de $100(1-\alpha)\,\%$ para θ.

11. Considera una densidad de familia exponencial

$$p_\theta(x) = h(x)e^{\eta(\theta)T(x)-A(\theta)}.$$

¿Bajo qué condiciones se puede encontrar una función $M(x)$ que satisfaga (3)?

12. Muestra que (4) implica la existencia de una función $M(x)$ que satisfaga (3).

13. Sea $X_1, \ldots, X_n$ una muestra iid de una distribución exponencial con media desconocida θ.

a) Encuentra la prueba exacta de razón de verosimilitud de tamaño-α para $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

b) Encuentra la prueba aproximada de tamaño-α basada en el teorema de Wilks.

c) Grafica las funciones de potencia de las dos pruebas anteriores y compáralas.

14. *No unicidad del MLE.* Sea $X_1, \ldots, X_n$ una muestra iid con densidad

$$p_\theta(x) = 2e^{-|x-\theta|}$$

para $x \in \mathbb{R}, \theta \in \mathbb{R}$. Esta distribución se conoce como la distribución de Laplace desplazada (o doble exponencial).

a) Argumenta que el MLE de θ no es único.

b) Para verificar esto, toma $\theta = 0$, simula $n = 10$ observaciones de la distribución de Laplace, grafica la verosimilitud e identifica el pico plano.
[Pista: Para simular de la distribución de Laplace estándar, simula una exponencial estándar y luego lanza una moneda justa para decidir si el signo debe ser positivo o negativo.**]**

15. *No existencia del MLE.* Considera una mezcla de dos distribuciones normales, es decir,

$$\pi N(\mu_1, \sigma_1^2) + (1-\pi)N(\mu_2, \sigma_2^2).$$

Supón que $X_1, \ldots, X_n$ son iid de la mezcla anterior. Argumenta que el MLE de $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \pi)$ no existe.
[Pista: ¿Qué sucede con la verosimilitud como función de σ_1 si $\mu_1 = X_1$, por ejemplo?**]**

16. Sea $X_1, \ldots, X_n$ una muestra iid de Unif$(0, \theta)$.

a) Muestra que el MLE es $\hat{\theta} = \max X_i$.

b) Explica por qué el MLE no puede ser asintóticamente normal y por qué esto no es un contraejemplo a la teoría presentada en la Sección 3.

c) Muestra que $n(\theta - \hat{\theta})$ converge en distribución a Exp(θ).

17. Refiérete al Ejemplo 1, el problema de Neyman–Scott.

a) Deriva el MLE indicado $\hat{\sigma}^2$ para σ^2.

b) Muestra que $\hat{\sigma}^2$ es inconsistente.

18. Supón que $X_1, \ldots, X_n$ son independientes, con $X_i \sim N(\theta_i, 1)$, $i = 1, \ldots, n$. En notación vectorial, podemos escribir $X \sim N_n(\theta, I_n)$, donde $X = (X_1, \ldots, X_n)^\top$ es el vector observable y $\theta = (\theta_1, \ldots, \theta_n)^\top$ es el vector de medias desconocido.

    a) Usa el Ejercicio 3 para encontrar el MLE de $\psi = \|\theta\|^2$, la norma cuadrada de θ.

    b) Muestra que el MLE de ψ está sesgado.

    c) ¿Desaparece el sesgo anterior cuando $n \to \infty$? Explica qué está ocurriendo.

19. Para datos iid, la varianza asintótica del MLE $\hat{\theta}$ es $[nI(\theta)]^{-1}$ y, para construir intervalos de confianza, se necesita una estimación de $nI(\theta)$. Una elección razonable es $nI(\hat{\theta})$, pero hay otra opción que a menudo es más fácil de obtener y funciona al menos igual de bien. Para un θ escalar y una función de log-verosimilitud $\ell(\theta)$, definimos la *información de Fisher observada* $J(\theta)$ como:

$$J(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta^2}.$$

Cuadro 3.1: Los datos de Rubin (1981) sobre experimentos de entrenamiento para el SAT.

Escuela (i)	**Efecto del Tratamiento (X_i)**	**Error Estándar (σ_i)**
1	28.39	14.9
2	7.94	10.2
3	-2.75	16.3
4	6.82	11.0
5	-0.64	9.4
6	0.63	11.4
7	18.01	10.4
8	12.16	17.6

existe también una versión para múltiples parámetros, donde $J(\theta)$ es la matriz de segundas derivadas parciales negativas de la log-verosimilitud. La afirmación es que $J(\hat{\theta})$ es una buena estimación de $nI(\theta)$ en comparación con $nI(\hat{\theta})$.

Supongamos que tenemos muestras $(X_1, Y_1), \ldots, (X_n, Y_n)$ iid de una distribución bivariada con densidad

$$p_\theta(x, y) = e^{-x/\theta - y}, \quad x, y > 0, \theta > 0.$$

En este caso, $\hat{\theta}$, $nI(\hat{\theta})$ y $J(\hat{\theta})$ pueden encontrarse analíticamente. Usa simulaciones para comparar las distribuciones muestrales de

$$[nI(\hat{\theta})]^{1/2}(\hat{\theta} - \theta) \quad \text{y} \quad [J(\hat{\theta})]^{1/2}(\hat{\theta} - \theta).$$

20. Supón que $X_1, \ldots, X_n$ son independientes, con $X_i \sim N(\lambda, \sigma_i^2 + \psi)$, donde $\sigma_1, \ldots, \sigma_n$ son conocidos, pero $\theta = (\psi, \lambda)$ es desconocido. Aquí $\psi \geqslant 0$ es el parámetro de interés y λ es un parámetro molesto.

    Para los datos en el Cuadro 3.1, tomados del estudio de entrenamiento SAT en Rubin (1981), encuentra y grafica la función de verosimilitud perfilada para ψ.

21. Sea $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ y consideremos la prueba de hipótesis $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

    a) Muestra que el estadístico de razón de verosimilitud $W = W(\mu_0)$ se puede expresar como
    $$W = n \log\{1 + T^2/(n-1)\},$$
    donde $T = n^{1/2}(\bar{X} - \mu_0)/S$ es el estadístico t-usual.

    b) Muestra que $\mathbb{E}(W) = 1 + bn^{-1} + O(n^{-2})$, donde $b = 3/2$.
    [Pista: Busca (por ejemplo, en *Wikipedia*) fórmulas para los momentos pares de variables aleatorias t de Student.**]**

    c) Define $W_b = W/(1 + bn^{-1})$; esto se conoce como el *estadístico de razón de verosimilitud corregido de Bartlett*. Compara, usando simulaciones, la precisión de las aproximaciones ChiSq(1) para W y W_b cuando n es relativamente pequeño; puedes tomar $\mu = \mu_0 = 0$ y $\sigma = 1$.

22. *One-step estimation* es un método mediante el cual un estimador consistente se actualiza, a través de una única iteración del método de Newton en la Sección 7, para obtener un estimador asintóticamente eficiente. Es decir, sea $\hat{\theta}_0$ un estimador consistente de θ, y definimos la versión de un paso como:
    $$\hat{\theta}_1 = \hat{\theta}_0 - D(\hat{\theta}_0)^{-1} \nabla \ell(\hat{\theta}_0),$$
    donde $D(\theta)$ es la matriz de segundas derivadas de $\ell(\theta)$. Se puede demostrar que $\hat{\theta}_1$ es asintóticamente eficiente, como el MLE.

    Como un ejemplo simple de esto, supongamos que $X_1, \ldots, X_n$ son iid $N(\theta, 1)$. Tomamos $\hat{\theta}_0$ como la mediana muestral, que es consistente pero no asintóticamente eficiente. Encuentra la versión de un paso $\hat{\theta}_1$ y argumenta su eficiencia.

23. En el contexto del Ejercicio 2, considera la prueba de hipótesis $H_0 : \theta = 1$ versus $H_1 : \theta \neq 1$. Usa Monte Carlo para encontrar el punto de corte k_α para la prueba de razón de verosimilitud de tamaño-α.

    Toma $n = 10$, $\alpha = 0{,}05$ y genera una muestra de Monte Carlo de tamaño $M = 5000$. Compara tu punto de corte k_α con aquel basado en la aproximación ji-cuadrado para muestras grandes.

24. Considera un modelo general de ubicación-escala, es decir, donde X tiene densidad
    $$\sigma^{-1} p(\sigma^{-1}(x - \mu)),$$
    con $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+$, y p una densidad en $\mathbb{R}$. Dado que $X_1, \ldots, X_n$ son iid de este modelo, supongamos que el objetivo es probar $H_0 : \sigma = \sigma_0$ versus $H_1 : \sigma \neq \sigma_0$.

    Demuestra o argumenta que la distribución nula (exacta, no asintótica) del estadístico de razón de verosimilitud no depende de μ.

25. La formulación matemática del teorema de Birnbaum sobre el principio de verosimilitud y la clarificación de Evans involucra *relaciones de equivalencia*.

    Una relación de equivalencia $\sim$ en $\mathbb{X}$ es una relación binaria que satisface:

    - **Reflexividad**: $x \sim x$ para todo $x \in \mathbb{X}$.

- **Simetría**: Si $x \sim y$, entonces $y \sim x$.
- **Transitividad**: Si $x \sim y$ y $y \sim z$, entonces $x \sim z$.

a) Uno de los ejemplos más comunes de relaciones de equivalencia en estadística es la igualdad de funciones μ-medibles hasta conjuntos de medida-μ cero. Es decir, escribimos $f \sim g$ si $f = g$ μ-casi en todas partes. Demuestra que $\sim$ es una relación de equivalencia en el conjunto de todas las funciones μ-medibles.

b) Otro ejemplo de relaciones de equivalencia aparece en teoría de grupos. Considera un grupo G de transformaciones $g : \mathbb{X} \to \mathbb{X}$. Escribimos $x \sim y$ si existe $g \in G$ tal que $y = gx$. Demuestra que $\sim$ es una relación de equivalencia en $\mathbb{X}$.

c) Sea $\mathbb{X}$ un conjunto equipado con una relación de equivalencia $\sim$. Dado $x \in \mathbb{X}$, define la *clase de equivalencia*

$$E_x = \{y \in \mathbb{X} : y \sim x\},$$

el conjunto de todos los elementos equivalentes a x. Para cualesquiera x, y, demuestra que E_x y E_y son disjuntos o exactamente iguales. Es decir, $\sim$ induce una partición $\{E_x : x \in \mathbb{X}\}$ de $\mathbb{X}$ en clases de equivalencia.

Capítulo 4: Inferencia Bayesiana

Teoría Estadística Avanzada

Sigla DES124

Prof. Jaime Lincovil

4.1. Introducción

El enfoque clásico frecuentista de la estadística es uno de los cuales la mayoria de los estudiantes están familiarizados. Es decir, para un procedimiento dado—estimador, prueba, intervalo de confianza, etc.—el frecuentista está interesado en el desempeño de ese procedimiento en términos de muestreo repetido. Por ejemplo, la calidad de una prueba se mide por su función de potencia, que no es más que la proporción límite de veces que la prueba rechaza la hipótesis nula al muestrear desde una distribución contenida en la hipótesis alternativa. Esto está bien, pero es importante entender las limitaciones de tales consideraciones. En particular, la función de potencia de una prueba no proporciona tranquilidad cuando se dispone de un conjunto fijo de datos y se quiere medir la incertidumbre sobre la veracidad de la hipótesis nula. Por lo tanto, hay razones para buscar un enfoque diferente, uno que pueda permitirte informar una suerte de *probabilidad* de que la hipótesis nula sea cierta, dados los datos observados. Un enfoque bayesiano hace esto posible, pero necesitamos ver el problema desde una perspectiva muy diferente.

Cuando los estudiantes llegan a un curso de postgrado, seguramente ya saben algo sobre el enfoque bayesiano. Por ejemplo, estoy seguro de que todos saben que, en el contexto bayesiano, el parámetro desconocido se trata como una variable aleatoria, con una distribución previa, y el teorema de Bayes se usa para producir una distribución posterior. Pero es natural preguntarse por qué el parámetro, que es una cantidad fija pero desconocida, debería ser tratado como aleatorio. Por ejemplo, parece absurdo asumir que el ingreso medio en el condado de Cook se selecciona al azar, ¿verdad? Este es un punto sutil pero importante. La justificación del enfoque bayesiano se basa en el siguiente tipo de "axioma:"

Las incertidumbres solo pueden describirse con probabilidad.

Esto significa que, para cualquier cosa que no conozcamos—por ejemplo, el parámetro θ en un problema estadístico—la única manera lógica de describir nuestras creencias es con probabilidad. Esto es lo que distingue el enfoque bayesiano del enfoque clásico. En este último, θ se asume fijo pero desconocido.

Pero, ¿qué significa que θ sea "desconocido"? ¿Realmente no sabemos nada al respecto? ¿No sabemos cómo resumir el conocimiento que tenemos o nos sentimos incómodos al usar este

conocimiento?

Parece poco realista asumir que realmente no sabemos *nada* sobre θ. Por ejemplo, si θ es el ingreso medio en el condado de Cook, sabemos que θ es positivo y menor a 1 billón de dólares; también creeríamos que $\theta \in (40K, 60K)$ es más probable que $\theta \in (200K, 220K)$.

Si, para cada evento relacionado con θ, asignamos algún puntaje numérico que represente nuestra incertidumbre y dichos puntajes satisfacen ciertas propiedades de consistencia[6], entonces hemos asignado efectivamente una distribución de probabilidad en el espacio de parámetros. Esto es lo que se llama la *distribución previa*.

Lo particularmente interesante de este argumento es que no hay noción de muestreo repetido, etc., como estamos acostumbrados a ver en un curso básico de probabilidad. Es decir, esta distribución previa es simplemente una descripción de la propia incertidumbre y no necesita estar relacionada con el *azar* per se. Esto no es tan ajeno como podría parecer inicialmente. Por ejemplo, supongamos que tú y varios amigos han sido invitados a una fiesta el próximo sábado. Cuando tu amigo te pregunta si asistirás, podrías responder con algo como: "hay un 50-50 de probabilidad de que vaya". Aunque esto no está en la escala de probabilidades, tiene una interpretación similar. Lo mismo ocurre con los pronósticos del clima, por ejemplo: "hay un 30 % de probabilidad de lluvia mañana".

Observa que estos eventos son diferentes del tipo de experimentos que pueden repetirse una y otra vez, como lanzar un dado, y sin embargo, se pueden definir probabilidades. Afortunadamente, estas probabilidades subjetivas pueden manipularse de la misma manera que la probabilidad basada en frecuencias ordinarias.

En el problema estadístico, tenemos incertidumbre sobre el parámetro θ. Luego describimos nuestra incertidumbre con probabilidades (subjetivas). Es decir, asignamos probabilidades a eventos como $\{\theta > 7\}$, $\{-0{,}33 \leqslant \theta < 0{,}98\}$, etc., lo que describe la distribución previa Π para θ. Esto es, en efecto, lo mismo que asumir que el parámetro desconocido en sí mismo es una variable aleatoria con una distribución especificada. Es un error común pensar que el análisis bayesiano *asume* que el parámetro es una variable aleatoria. Por el contrario, un bayesiano comienza asignando probabilidades a todas las cosas que son inciertas; que esto sea equivalente a considerar θ como una variable aleatoria es solo una consecuencia.

Teniendo una comprensión básica de la lógica detrás del enfoque bayesiano, en el resto de este capítulo investigaremos algunos aspectos específicos del análisis bayesiano. Aquí no entraremos en discusiones filosóficas sobre enfoques bayesianos frente a no bayesianos, aunque tales debates han existido por muchos años[7]. Primero, describiré el modelo de Bayes y cómo la distribución previa se actualiza a una distribución posterior mediante el teorema de Bayes. Luego, discutiremos cómo se usa esta distribución posterior para la inferencia y daremos algunos ejemplos.

Después, intentaré describir varias motivaciones para un análisis bayesiano. Si uno elige utilizar un análisis bayesiano, entonces quizás la pregunta más importante es cómo elegir la distribución previa. Existen varios métodos ahora bastante estándar, que describiré breve-

[6] Las propiedades de consistencia son bastante razonables, por ejemplo, si un evento es un subconjunto de otro, entonces el primero no puede tener un puntaje mayor que el segundo. Sin embargo, construir un sistema de puntuación desde cero no es tan fácil; usualmente se hace asumiendo un modelo de probabilidad particular.

[7] Hoy en día, la mayoría de las personas entienden que tanto los enfoques bayesianos como los no bayesianos tienen sus ventajas y desventajas, es decir, ninguno es claramente mejor que el otro. Por lo tanto, la discusión se centra más en hacer el mejor uso de las herramientas disponibles en un problema dado.

mente.

4.2. Análisis Bayesiano

4.2.1. Configuración básica de un problema de inferencia bayesiana

Al igual que antes, comenzamos con un espacio muestral (medible) $(\mathbb{X}, \mathcal{A})$ que está equipado con una familia de distribuciones de probabilidad $P = \{P_\theta : \theta \in \Theta\}$. Supongamos también que existe una medida σ-finita μ tal que $P_\theta \ll \mu$ para todo θ, de modo que tenemos derivadas de Radon-Nikodym (densidades) $p_\theta(x) = (dP_\theta/d\mu)(x)$ con respecto a μ.

La diferencia es que también está disponible alguna distribución de probabilidad Π en Θ, a la que llamamos la *distribución previa*. Para ayudar a diferenciar lo que es aleatorio de lo que es fijo, usaré la notación Θ para una variable aleatoria distribuida según Π, y θ para los valores observados. No debería haber confusión al usar la notación Θ tanto para el espacio de parámetros como para la versión aleatoria del parámetro.

El enfoque bayesiano asume el siguiente modelo jerárquico:

$$\Theta \sim \Pi \quad \text{y} \quad X \mid (\Theta = \theta) \sim p_\theta(x). \tag{1}$$

El objetivo es tomar la información del valor observado $X = x$ y actualizar la información previa sobre el "parámetro" Θ. Esto se logra, en términos generales, mediante el teorema de Bayes. Pero antes de entrar en los detalles técnicos, es útil comprender el razonamiento detrás de esta elección en particular.

Si la incertidumbre sobre θ está descrita por la distribución de probabilidad (subjetiva) Π, entonces la incertidumbre sobre θ *después* de observar los datos x debería estar descrita por la distribución condicional Π_x, la *distribución posterior* de Θ dado $X = x$. Pronto discutiremos cómo se utiliza esta distribución posterior para la inferencia.

4.2.2. Teorema de Bayes

Todos estamos familiarizados con el teorema de Bayes de un curso introductorio de probabilidad. En presentaciones simples, el teorema proporciona una fórmula para la probabilidad $P(A \mid B)$ en términos de la probabilidad condicional opuesta $P(B \mid A)$ y las probabilidades marginales $P(A)$ y $P(B)$. Aquí presentamos una versión muy general de esta resultado en términos de teoría de la medida.

Teorema 4.1 Teorema de Bayes

Bajo la configuración descrita anteriormente, sea Π_x la distribución condicional de Θ dado $X = x$. Entonces, $\Pi_x \ll \Pi$ para Π-casi todo x, donde

$$P_\Pi = \int p_\theta d\Pi(\theta)$$

es la distribución marginal de X en el modelo (4.1). Además, la derivada de Radon-Nikodym de Π_x con respecto a Π es

$$\frac{d\Pi_x}{d\Pi}(\theta) = \frac{p_\theta(x)}{p_\Pi(x)},$$

para aquellos x tales que la densidad marginal $p_\Pi(x) = (dP_\Pi/d\mu)(x)$ no es ni 0 ni ∞. Dado que el conjunto de todos los x tales que $p_\Pi(x) \in \{0, \infty\}$ es un conjunto nulo con respecto a P_Π, la derivada de Radon-Nikodym puede definirse arbitrariamente para dichos x.

Demostración. Esta demostración proviene de Schervish (1995, p. 16–17). Definimos

$$C_0 = \{x : p_\Pi(x) = 0\} \quad \text{y} \quad C_\infty = \{x : p_\Pi(x) = \infty\}.$$

Dado que $P_\Pi(A) = \int_A p_\Pi(x)d\mu(x)$, se sigue que

$$P_\Pi(C_0) = \int_{C_0} p_\Pi(x)d\mu(x) = 0,$$

$$P_\Pi(C_\infty) = \int_{C_\infty} p_\Pi(x)d\mu(x) = \int_{C_\infty} \infty \, d\mu(x).$$

La última integral será igual a ∞ si $\mu(C_\infty) > 0$; pero dado que esta cantidad no puede ser igual a ∞ (ya que es una probabilidad), debe ser que $\mu(C_\infty) = 0$ y, por lo tanto, $P_\Pi(C_\infty) = 0$. Esto prueba la última afirmación en el teorema sobre el denominador.

Para probar la afirmación principal, recordemos que la distribución posterior Π_x debe satisfacer

$$P(\Theta \in B, X \in A) = \int_A \Pi_x(B)\, dP_\Pi(x), \quad \text{para todos los conjuntos medibles } A, B. \quad (2)$$

Las distribuciones conjuntas son simétricas (es decir, "condicional por marginal" puede ir en ambas direcciones), por lo que el lado izquierdo (LHS) de la ecuación (2) también se puede escribir como

$$\text{LHS} = \int_B \int_A p_\theta(x)d\mu(x)d\Pi(\theta) = \int_A \left[\int_B p_\theta(x)d\Pi(\theta)\right] d\mu(x),$$

donde la segunda igualdad se sigue del teorema de Fubini. Dado que estamos forzando la igualdad entre los lados izquierdo y derecho, es decir, LHS = RHS, debemos tener que

$$\text{RHS} = \int_A \left[\int_B p_\theta(x)d\Pi(\theta)\right] d\mu(x).$$

Pero, RHS también se puede escribir como

$$\text{RHS} = \int_A \left[\Pi_x(B) \int_\Theta p_\theta(x)d\Pi(\theta)\right] d\mu(x).$$

Dado que ambas expresiones para RHS deben ser iguales para todos los conjuntos A y B, debemos tener

$$\Pi_x(B) \int_\Theta p_\theta(x)d\Pi(\theta) = \int_B p_\theta(x)d\Pi(\theta), \quad \text{para } P_\Pi\text{-casi todo } x.$$

Resolviendo para $\Pi_x(B)$, vemos que $\Pi_x \ll \Pi$ y que la fórmula para la densidad posterior (derivada de Radon-Nikodym) también se sigue. ■

En el caso en que la distribución previa Π tenga una densidad con respecto a alguna medida ν, obtenemos la forma más familiar de la actualización bayesiana del posterior.

Corolario 4.1

Supongamos que $\Pi \ll \nu$ con derivada de Radon-Nikodym π. Entonces, la distribución posterior Π_x también es absolutamente continua con respecto a ν, y su densidad, denotada como π_x, está dada por

$$\pi_x(\theta) = \frac{p_\theta(x)\pi(\theta)}{p_\Pi(x)} \propto p_\theta(x)\pi(\theta).$$

Demostración. Se sigue directamente de la propiedad de la regla de la cadena para las derivadas de Radon-Nikodym; ver Ejercicio 1. ■

El mensaje clave es que, dada una distribución previa Π y una verosimilitud $p_\theta(x)$, se puede construir una distribución posterior Π_x y, en el caso en que Π tenga una densidad, el posterior también tendrá una densidad y será proporcional a la densidad previa multiplicada por la verosimilitud.

4.2.3. Inferencia

La distribución posterior es todo lo que se necesita para hacer inferencia sobre θ. Es decir, una vez que la distribución posterior está disponible, podemos usarla para calcular diversos estimadores. Por ejemplo, una estimación puntual típica para θ es la media o la moda posterior. La media posterior se define como

$$\hat{\theta}_{\text{mean}} = \mathbb{E}(\Theta \mid X = x) = \int \theta \, d\Pi_x(\theta),$$

y, en el caso en que Π_x tenga una densidad π_x, la moda posterior[8] se define como

$$\hat{\theta}_{\text{mode}} = \arg\max_\theta \pi_x(\theta),$$

lo cual es similar al estimador de máxima verosimilitud. Existen nociones más formales de estimadores de Bayes (o, más generalmente, reglas de Bayes) que encontraremos más adelante.

Para la estimación de conjuntos, un bayesiano usa lo que se conoce como *conjunto creíble*. Un conjunto creíble del $100(1-\alpha)$

$$C = \{\theta : \theta \text{ está entre los cuantiles } \alpha/2 \text{ y } 1-\alpha/2 \text{ de } \Pi_x\}.$$

De manera alternativa, y más generalmente, si la distribución posterior Π_x tiene una densidad π_x, entonces se puede usar una *región de densidad posterior más alta*. Es decir,

$$C = \{\theta : \pi_x(\theta) \geqslant c_\alpha\},$$

[8] También llamado el estimador *MAP*, por *maximum a posteriori*.

donde c_α se elige de modo que $\Pi_x(C) = 1 - \alpha$.

El punto clave aquí es que, a diferencia de un intervalo de confianza frecuentista, un intervalo creíble *no* necesariamente tiene la propiedad de que la probabilidad de cobertura de C sea igual a $1 - \alpha$.

La prueba de hipótesis es similar. Una hipótesis sobre θ define un subconjunto H del espacio de parámetros, y su probabilidad previa es $\Pi(H)$. De acuerdo con el teorema de Bayes, la probabilidad posterior de H es

$$\Pi_x(H) = \frac{\int_H p_\theta(x) d\Pi(\theta)}{p_\Pi(x)}.$$

Entonces, el bayesiano *rechazará* H si esta probabilidad posterior es demasiado pequeña.

Un punto a notar es que, en este esquema, si la probabilidad previa de H es cero, entonces la probabilidad posterior también lo será. Por lo tanto, un bayesiano debe adoptar un enfoque diferente cuando $\Pi(H) = 0$. Estas diferencias están relacionadas con los factores de Bayes y la selección de modelos, pero no discutiremos más sobre esto aquí.

4.2.4. Marginalización

En nuestra discusión sobre la verosimilitud, consideramos el problema en el que θ era un vector, pero solo una característica o un componente de θ era de interés. Para concretar, supongamos que $\theta = (\psi, \lambda)$, donde tanto ψ como λ son desconocidos, pero solo ψ es de interés. En ese caso, se requiere alguna modificación a la función de verosimilitud usual, por ejemplo, la *verosimilitud perfilada* obtenida al maximizar la verosimilitud sobre λ, punto a punto en ψ. Aunque esta modificación es conceptualmente simple, la verosimilitud perfilada tiene algunas limitaciones, por ejemplo, no incorpora un ajuste natural para la incertidumbre en λ.

En el contexto bayesiano, la marginalización es sencilla. Desde la teoría de la probabilidad, sabemos que, dada una densidad conjunta $p(x, y)$ para un vector aleatorio (X, Y), la densidad marginal para X se puede obtener mediante integración, es decir,

$$p(x) = \int p(x, y)\, dy.$$

En nuestro contexto actual, la distribución posterior para $\Theta = (\Psi, \Lambda)$ es una densidad conjunta, y la distribución posterior marginal para Ψ se puede obtener integrando la densidad conjunta posterior $\pi_x(\psi, \lambda)$ sobre λ:

$$\pi_x(\psi) = \int \pi_x(\psi, \lambda)\, d\lambda.$$

Este enfoque es, en muchos casos, más sencillo que en el contexto no bayesiano, donde se debe idear e implementar alguna modificación a la verosimilitud. Para el bayesiano, las reglas de la probabilidad indican cómo manejar la marginalización.

En muchas situaciones prácticas, no es posible realizar la integración requerida a mano, por lo que se necesita algún método numérico. En algunos casos, se puede utilizar integración numérica (por ejemplo, mediante sumas de Riemann, la regla del trapecio/de Simpson, etc., o a través de la función `integrate` en R) para este propósito.

Generalmente, la distribución conjunta posterior es intratable, por lo que se requieren algunos tipos de simulaciones para la inferencia posterior. En este caso, la marginalización es particularmente simple. Supongamos que se tiene una muestra $\{(\Psi^{(m)}, \Lambda^{(m)}), m = 1, \ldots, M\}$ de la distribución posterior $\pi_x(\psi, \lambda)$. Luego, una muestra de la distribución posterior marginal de Ψ se obtiene ignorando la parte de Λ en la muestra posterior, es decir, construyendo un intervalo creíble para ψ basado en $\Psi^{(1)}, \ldots, \Psi^{(M)}$.

Una versión extrema de la marginalización es la de la predicción, es decir, cuando la cantidad de interés es X_{n+1}, la siguiente observación. En el problema de predicción, el bayesiano integrará la densidad condicional $p_\theta(x)$ con respecto a la distribución posterior de Θ, dado $(X_1, \ldots, X_n)$, para obtener la llamada *distribución predictiva* de X_{n+1}, dado $(X_1, \ldots, X_n)$.

Los Ejercicios 6 y 7 te invitan a explorar más a fondo el problema de la predicción.

4.3. Algunos Ejemplos

Ejemplo 4.1

Supongamos que $X_1, \ldots, X_n$ son iid $N(\theta, \sigma^2)$, donde σ es conocido. Además, supongamos que θ tiene una distribución previa $N(\omega, \tau^2)$ para valores fijos de ω y τ. En particular, las creencias previas sugieren que θ estaría ubicado en algún lugar cerca de ω, pero ser mayor o menor que ω es igualmente probable.
Dado que $\bar{X}$ es un estadístico suficiente, la distribución posterior dependerá de $(X_1, \ldots, X_n)$ solo a través de la media $\bar{X}$ (¿por qué?). En este caso, la distribución posterior $\theta \mid (\bar{X} = \bar{x})$ es normal, con

$$\text{media} = \frac{n\tau^2}{\sigma^2 + n\tau^2}\bar{x} + \frac{\sigma^2}{\sigma^2 + n\tau^2}\omega$$

y

$$\text{varianza} = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}.$$

Dado que la distribución posterior tiene una expresión en forma cerrada, es fácil calcular momentos posteriores, intervalos creíbles o probar hipótesis. Por ejemplo, una estimación puntual de θ es $\hat{\theta} = \text{media}$, donde la media se muestra arriba.

Ejemplo 4.2

Supongamos que $X_1, \ldots, X_n$ es una muestra independiente de una población Pois(θ). Consideremos una distribución previa Gamma(a, b) para θ. Multiplicando la verosimilitud y la distribución previa, obtenemos

$$p_\theta(x)\pi(\theta) = \text{const} \times e^{-n\theta}\theta^{n\bar{x}}\theta^{a-1}e^{-\theta/b} = \text{const} \times \theta^{n\bar{x}+a-1}e^{-(n+1/b)\theta}.$$

Es claro que, después de la normalización, la densidad posterior debe seguir una distribución Gamma$(n\bar{x}+a, [n+1/b]^{-1})$. De nuevo, el hecho de que la distribución posterior tenga una expresión en forma cerrada hace que los cálculos sean sencillos.
Por ejemplo, supongamos que $a = 5$ y $b = 2$; además, en una muestra de tamaño $n = 10$, la media observada es $\bar{x} = 7$. Entonces, la distribución posterior para θ es Gamma$(75, 0{,}095)$ y un intervalo creíble del 95 % para θ es

$$\texttt{qgamma(c(0.025, 0.975), shape=75, scale=0.095)} = (5{,}60, 8{,}23).$$

Una observación importante es que, en los dos ejemplos anteriores, la distribución posterior pertenece a la misma familia que la distribución previa. Estos son casos especiales de un concepto más general de *distribuciones previas conjugadas*. Específicamente, una clase de distribuciones forma una clase conjugada si, para cualquier distribución previa Π en la clase, la distribución posterior Π_x también pertenece a la misma clase. En todos estos problemas, el análisis de la distribución posterior es sencillo, tal como en los dos ejemplos anteriores.
Sin embargo, uno puede cuestionar cuán realista es una distribución previa conjugada cuando su única justificación es que permite cálculos simples.
A continuación, se presenta un ejemplo con un modelo no estándar y una distribución previa no conjugada. Este ejemplo ilustra un método numérico—*Markov Chain Monte Carlo* (MCMC)—que es utilizado frecuentemente por los bayesianos para calcular la distribución posterior cuando no existe una expresión en forma cerrada.

Ejemplo 4.3

Supongamos que $X_1, \ldots, X_n$ son muestras iid con densidad

$$p_\theta(x) = \frac{1 - \cos(x - \theta)}{2\pi}, \quad 0 \leqslant x \leqslant 2\pi,$$

donde θ es un parámetro desconocido en $[-\pi, \pi]$. Tomamos una distribución previa Unif$(-\pi, \pi)$ para Θ. En este caso, la distribución posterior no tiene una forma cerrada, aunque sus características pueden determinarse mediante integración numérica. En este caso, utilizamos *Markov Chain Monte Carlo* (MCMC) para simular desde la distribución posterior y estimar diversas cantidades. Emplearemos el algoritmo de *Metropolis-Hastings*, que usa una distribución de propuesta basada en un paseo aleatorio uniforme, con ancho de ventana $a = 0{,}5$. El código en R se muestra en la Figura 1.
Un histograma de las 10,000 muestras de Π_x se muestra en la Figura 2(a), junto con un gráfico de la densidad posterior verdadera, y es claro que el procedimiento de muestreo está funcionando correctamente; el panel (b) muestra un gráfico de trazas que ayuda a evaluar si la cadena de Markov a "convergido".

Ejemplo 4.4

Sea $X_1, \ldots, X_n$ una muestra iid de $N(\mu, \sigma^2)$, donde tanto μ como σ^2 son desconocidos. Como distribución previa, consideremos una "distribución" con densidad $\pi(\mu, \sigma^2) \propto 1/\sigma^2$. Este tipo de distribución previa se denomina *impropia* porque la integral sobre (μ, σ^2) no es finita. La validez de una distribución previa impropia es cuestionable[a], pero aún se puede aplicar formalmente el teorema de Bayes para obtener una distribución posterior. Aunque esto pueda parecer extraño, el uso de distribuciones previas impropias es bastante común; ver Sección 4.5.4.

Aquí el objetivo es simplemente determinar la distribución marginal posterior de μ. Primero, la verosimilitud normal puede escribirse como

$$L(\mu, \sigma^2) = (1/\sigma^2)^{n/2} e^{-D/2\sigma^2},$$

donde

$$D = \frac{1}{2}\{(n-1)s^2 + n(\mu - \bar{x})^2\},$$

con

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Con una distribución previa $\pi(\mu, \sigma^2) \propto 1/\sigma^2$, la densidad posterior satisface

$$\pi_x(\mu, \sigma^2) \propto (1/\sigma^2)^{(n/2)+1} e^{-D/2\sigma^2}.$$

El lado derecho de la ecuación anterior es proporcional a la densidad de una distribución conocida, a saber, la distribución normal-inversa gamma[b], aunque esto no es particularmente relevante.
Para la distribución marginal posterior de μ, necesitamos integrar σ^2 en la distribución posterior $\pi_x(\mu, \sigma^2)$. La clave aquí es que el lado derecho de la ecuación anterior, como función de σ^2, es proporcional a la densidad de una distribución gamma inversa, que tiene la forma

$$\frac{b^a}{\Gamma(a)}\left(\frac{1}{x}\right)^{a+1} e^{-b/x}.$$

Por lo tanto, si integramos sobre σ^2 en la ecuación anterior, obtenemos

$$\int_0^\infty (1/\sigma^2)^{n/2+1} e^{-D/2\sigma^2} d\sigma^2 = \frac{\Gamma(n/2)}{D^{n/2}},$$

y la densidad marginal posterior debe satisfacer

$$\pi_x(\mu) \propto \frac{\Gamma(n/2)}{D^{n/2}} \propto \left(\frac{1}{(n-1)s^2 + n(\mu - \bar{x})^2}\right)^{n/2}.$$

La expresión en el lado derecho anterior, como función de μ, es proporcional a una transformación de ubicación y escala de la densidad t de Student con $n-1$ grados de libertad; es decir, dado x, la distribución de

$$n^{1/2}(\mu - \bar{x})/s$$

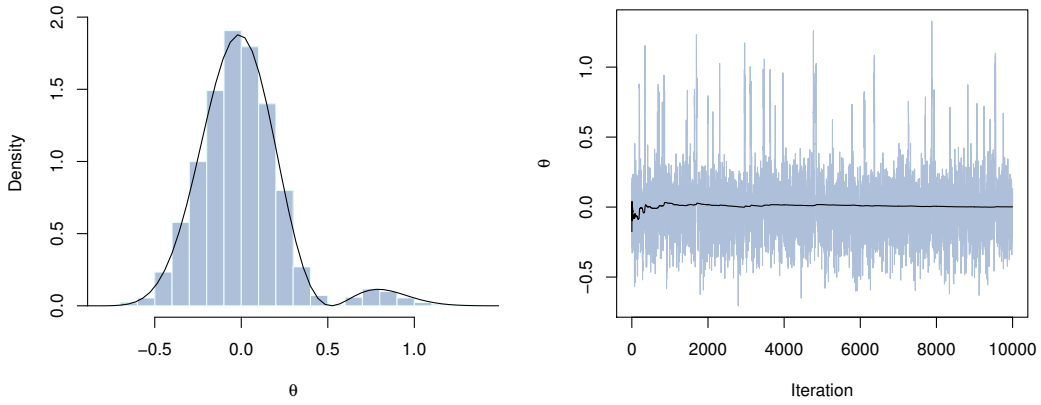
sigue una distribución $t(n-1)$.

[a] Probablemente, la mejor manera de interpretar una distribución previa impropia es como una especie de peso adjunto a cada punto del parámetro, en este caso, (μ, σ^2). Para la distribución previa en este ejemplo, aquellos pares (μ, σ^2) con σ^2 pequeño reciben más peso previo.
[b] http://en.wikipedia.org/wiki/Normal-inverse-gamma_distribution

Por último, también es interesante considerar los beneficios de un enfoque bayesiano basado en los criterios clásicos. En términos generales, si la distribución previa es razonable, entonces un procedimiento bayesiano que utiliza esta información generalmente superará a un procedi-

```
mh <- function(x0, f, dprop, rprop, N, B) {

  x <- matrix(NA, N + B, length(x0))
  fx <- rep(NA, N + B)
  x[1,] <- x0
  fx[1] <- f(x0)
  ct <- 0
  for(i in 2:(N + B)) {

    u <- rprop(x[i-1,])
    fu <- f(u)
    r <- log(fu) + log(dprop(x[i-1,], u)) - log(fx[i-1]) - log(dprop(u, x[i-1,]))
    R <- min(exp(r), 1)
    if(runif(1) <= R) {

      ct <- ct + 1
      x[i,] <- u
      fx[i] <- fu

    } else {

      x[i,] <- x[i-1,]
      fx[i] <- fx[i-1]

    }

  }
  out <- list(x=x[-(1:B),], fx=fx[-(1:B)], rate=ct / (N + B))
  return(out)

}

X <- c(3.91, 4.85, 2.28, 4.06, 3.70, 4.04, 5.46, 3.53, 2.28, 1.96, 2.53,
       3.88, 2.22, 3.47, 4.82, 2.46, 2.99, 2.54, 0.52, 2.50)
lik <- function(theta) {

  o <- drop(exp(apply(log(1 - cos(outer(X, theta, "-"))), 2, sum)))
  ind <- (theta <= pi) & (theta >= -pi)
  o <- o * ind
  return(o)

}
a <- 0.5
dprop <- function(theta, theta0) dunif(theta, theta0 - a, theta0 + a)
rprop <- function(theta0) runif(1, theta0 - a, theta0 + a)
den <- integrate(lik, -pi, pi)$value
dpost <- function(theta) lik(theta) / den
x <- seq(-pi, pi, len=150); dpost.x <- dpost(x)
ylim <- c(0, 1.05 * max(dpost.x))
N <- 10000
B <- 5000
theta.mcmc <- mh(runif(1, -pi, pi), lik, dprop, rprop, N, B)
hist(theta.mcmc$x, freq=FALSE, col="gray", border="white", ylim=ylim, xlab=expression(theta), main="")
lines(x, dpost.x)
plot(theta.mcmc$x, type="l", col="gray", xlab="Iteration", ylab=expression(theta))
lines(1:N, cumsum(theta.mcmc$x) / (1:N))
print(quantile(theta.mcmc$x, c(0.05, 0.95)))
```

Figura 2: Codigo en R para el algoritmo de Metropolis-Hastings del ejemplo 3



(a) Posterior distribution (b) Trace plot

Figura 3: Panel (a): Histograma de la muestra de Monte Carlo de la distribución posterior Π_x en el Ejemplo 4.3, con la densidad posterior verdadera superpuesta; Panel (b): gráfico de trazas de la muestra de Monte Carlo, con la media en ejecución, sugiriendo que la cadena de Markov se "mezcló bien".

miento no bayesiano que la ignora. Esto es esencialmente un problema de teoría de decisiones, pero a continuación se muestra una ilustración sencilla de la idea principal.

Ejemplo 4.5

Supongamos que $\omega = 0$, $\tau = 1$ y $\sigma = 1$. En este caso, la media posterior de θ es

$$\hat{\theta} = \frac{n\bar{X}}{n+1}.$$

Veamos cómo se compara esto con la estimación habitual $\bar{X}$, que es el estimador de máxima verosimilitud (MLE). Un criterio por el cual se pueden comparar es el error cuadrático medio:

$$\text{mse}(\theta; \hat{\theta}) := \mathbb{E}_\theta(\hat{\theta} - \theta)^2,$$

como función del verdadero θ. Es fácil verificar que

$$\text{mse}(\theta; \bar{X}) = \frac{1}{n}, \quad \text{y} \quad \text{mse}(\theta; \hat{\theta}) = \frac{\theta^2 + n}{(n+1)^2}.$$

Es fácil ver que si el verdadero θ está cerca de cero (la media de la distribución previa), entonces la estimación bayesiana es mejor; sin embargo, si la distribución previa está muy alejada y el verdadero θ es muy diferente de cero, la regla de Bayes puede ser superada por el MLE.

El mensaje clave es que, para distribuciones previas elegidas adecuadamente, los procedimientos bayesianos generalmente superan a los procedimientos no bayesianos. La dificultad es que la(s) distribución(es) previa(s) necesaria(s) dependen del valor real del parámetro desconocido.

4.4. Motivaciones para el enfoque bayesiano

4.4.1. Algunas motivaciones

- Existen conjuntos de axiomas de racionalidad y se ha demostrado que, si uno desea ser "racional", entonces debe ser bayesiano. Algunas descripciones de estas ideas se pueden encontrar en Ghosh et al. (2006). La teoría de la utilidad se discute en Keener (2010, Cap. 7.3).

- Existe un tipo específico de axiomas de "racionalidad", llamados *coherencia*, que es relativamente fácil de entender. Proviene de una perspectiva de apuestas. La idea es que un resumen razonable de la incertidumbre debe ser tal que uno esté dispuesto a hacer apuestas basadas en él. Por razones de espacio, no entraré en los detalles aquí, pero la idea es que, bajo algunos supuestos básicos[9], a menos que las incertidumbres que especifiques satisfagan las reglas de la probabilidad, es decir, los axiomas de Kolmogorov[10], yo podría diseñar una estrategia de apuestas que te garantice perder siempre. Por supuesto, esto indica que tu sistema de incertidumbres es defectuoso en algún sentido, o *incoherente*. El mensaje aquí es que, desde un punto de vista estadístico, si no eres bayesiano y resumes tus incertidumbres sobre θ de alguna manera que no sea una probabilidad genuina, entonces hay algo incorrecto en tu evaluación de la incertidumbre. Por lo tanto, solo el enfoque bayesiano es coherente. El Capítulo 1 de Kadane (2011) ofrece una excelente descripción de esta idea.

- En un enfoque bayesiano, es fácil incorporar cualquier información conocida sobre el parámetro en el análisis. Por ejemplo, supongamos que se sabe que la media θ de una población normal satisface $a \leqslant \theta \leqslant b$. Entonces, el enfoque bayesiano puede manejar esto fácilmente eligiendo una distribución previa con soporte en $[a, b]$ que refleje esta información. El enfoque clásico (frecuentista) no puede manejar dicha información con la misma facilidad.

- Existen teoremas en teoría de decisiones (llamados *teoremas de clase completa*) que establecen que, para un problema de inferencia dado, para cualquier regla de decisión, existe una regla de Bayes (aproximada) que es igual o mejor[11]. En otras palabras, no hay una razón real para buscar fuera de la clase de procedimientos bayesianos (aproximados), ya que, para cualquier procedimiento no bayesiano, siempre hay un procedimiento bayesiano que es igual de bueno.

4.4.2. Intercambiabilidad y el teorema de de Finetti

En cursos introductorios estamos acostumbrados a ver suposiciones de datos "independientes e idénticamente distribuidos". Sin embargo, también sabemos que existen otros tipos de estructuras de dependencia que suelen ser más realistas, pero no tan fáciles de manejar. En

[9] La suposición clave aquí es que estás igualmente dispuesto a comprar boletos de mí o a venderme boletos similares. Esto puede o no ser razonable.

[10] En realidad, no se necesita la aditividad numerable de Kolmogórov; los teoremas de coherencia están disponibles para medidas que son solo finitamente aditivas.

[11] Por ejemplo, se sabe que la media muestral $\bar{X}$ es una buena estimación de la media de una distribución normal, y esto corresponde a la media bayesiana posterior bajo una distribución previa uniforme "impropia" en $(-\infty, \infty)$, la cual puede verse como el límite de una secuencia de distribuciones previas "propias".

esta sección, discutiremos la noción de variables aleatorias *intercambiables*, que incluye el caso iid como un caso especial, y una consecuencia notable debido a de Finetti y posteriormente desarrollada por otros. Este material proviene de Schervish, Capítulo 1.

Definición 4.1

Un conjunto finito de variables aleatorias $X_1, \ldots, X_n$ es *intercambiable* si todas las permutaciones de $(X_1, \ldots, X_n)$ tienen la misma distribución conjunta. Una colección infinita de variables aleatorias es intercambiable si cada subconjunto finito es intercambiable.

Por ejemplo, supongamos que $X_1, \ldots, X_{50}$ son intercambiables. Entonces, todos los X_i tienen la misma distribución marginal. Además, (X_1, X_2) y (X_{33}, X_{44}) tienen la misma distribución conjunta, al igual que (X_2, X_7, X_5) y (X_{47}, X_{21}, X_{15}). De hecho, en el Ejercicio 10 se le pide demostrar que un conjunto $X_1, \ldots, X_n$ es intercambiable si y solo si todos los subconjuntos finitos tienen la misma distribución conjunta. También es fácil ver que las variables aleatorias iid son intercambiables[12].

Es importante comprender, de manera intuitiva, qué significa la intercambiabilidad. La intercambiabilidad no implica más que simetría en la distribución. Es decir, si el orden de las observaciones es irrelevante, entonces los datos son intercambiables. Esta es, obviamente, una suposición muy débil. La suposición de datos iid es bastante fuerte y, además, existen algunas dificultades filosóficas al asumir la existencia de un parámetro fijo pero desconocido en tal contexto[13]. Resulta que la intercambiabilidad es casi tan simple como iid y conduce a una motivación muy interesante para el análisis bayesiano. El siguiente ejemplo nos acerca más al teorema principal.

Ejemplo 4.6

Consideremos variables aleatorias $X_1, \ldots, X_n$ que modelamos como "iid condicionalmente". Es decir, existe una variable aleatoria Θ tal que $X_1, \ldots, X_n$ son iid dado el valor θ de Θ. Más formalmente,

$$\Theta \sim \Pi \quad \text{y} \quad (X_1, \ldots, X_n) \mid (\Theta = \theta) \overset{\text{iid}}{\sim} p_\theta. \tag{3}$$

Entonces $(X_1, \ldots, X_n)$ son intercambiables; ver el Ejercicio 11.
Esta estructura de "iid condicionalmente" es exactamente la misma que encontramos en los ejemplos de la Sección 4.3. Es decir, el modelo bayesiano implica un modelo intercambiable para los datos (marginalmente). El hecho sorprendente es que la relación también funciona en sentido contrario—*una colección infinita de variables aleatorias intercambiables es iid condicionalmente* (con respecto a algún prior Π y densidad p_θ). Esta es una versión del teorema de de Finetti para observables binarios.

[12] Las distribuciones conjuntas son productos y la multiplicación es conmutativa.

[13] Por ejemplo, en un experimento de lanzamiento de moneda, el parámetro θ que representa la probabilidad de que la moneda caiga en cara se suele considerar como la proporción límite de caras en una secuencia infinita de lanzamientos. ¿Cómo puede existir tal parámetro antes de que la secuencia de lanzamientos haya sido realizada?

Teorema 4.2

Una secuencia X_n de variables aleatorias binarias es intercambiable si y solo si existe una variable aleatoria Θ, que toma valores en $[0,1]$, tal que, dado $\Theta = \theta$, los X_n son iid Ber(θ). *Además, si la secuencia es intercambiable, entonces la distribución de* Θ *es única y* $n^{-1}\sum_{i=1}^{n} X_i$ *converge casi seguramente a* Θ.

En otras palabras, la intercambiabilidad implica que, para alguna medida de probabilidad Π en $[0,1]$, la distribución conjunta de $(X_1,\ldots,X_n)$ se puede escribir como

$$P(X_1 = x_1,\ldots,X_n = x_n) = \int \theta^{t(x)}(1-\theta)^{n-t(x)} d\Pi(\theta),$$

donde $t(x) = \sum_{i=1}^{n} x_i$. Se puede interpretar Π como un prior en el sentido bayesiano, junto con la verosimilitud Bernoulli. Sin embargo, Π está determinada por el límite de la secuencia X, lo cual se alinea con nuestra intuición en el problema del lanzamiento de moneda: el parámetro θ representa la proporción límite de caras en una cantidad infinita de lanzamientos. Así que el punto clave es que una simple suposición de intercambiabilidad es suficiente para implicar que el modelo bayesiano/jerárquico está en juego.

Existen versiones más generales del teorema de de Finetti. Aunque estas son más complicadas matemáticamente, la intuición es la misma que en el Teorema 2. Aquí se presenta uno de esos resultados.

Teorema 4.3 Hewitt-Savage

Una secuencia de variables aleatorias X_n en $(\mathbb{X},\mathcal{A})$ es intercambiable si y solo si existe una medida de probabilidad aleatoria $\mathcal{P}$ tal que, dado $\mathcal{P} = P$, las variables $X_1, X_2, \ldots$ son iid con distribución P. *Además, si el modelo es intercambiable, entonces* $\mathcal{P}$ *es única y está determinada por el límite*

$$\mathcal{P}_n(A) := n^{-1}\sum_{i=1}^{n} \mathbb{I}_A(X_i) \to \mathcal{P}(A) \quad \text{casi seguramente para cada } A \in \mathcal{A}.$$

Otra forma de interpretar el concepto de "medida de probabilidad aleatoria" es a través de una mezcla. El Teorema 3 establece que la secuencia X_n es intercambiable si y solo si existe una medida de probabilidad Π en el conjunto de todas las distribuciones sobre $(\mathbb{X},\mathcal{A})$, y la distribución marginal de $(X_1,\ldots,X_n)$ está dada por

$$\int \prod_{i=1}^{n} P(X_i \in A_i) d\Pi(P),$$

una mezcla de modelos iid. Esta formulación es más general, pero establece una conexión con la noción de variables aleatorias iid condicionales. Esta ilustración también arroja luz sobre las implicaciones de este teorema para los bayesianos. De hecho, el teorema establece que una simple suposición de intercambiabilidad implica la existencia de un modelo jerárquico como el de (1), que puede interpretarse como un prior y una verosimilitud a ser actualizados mediante el teorema de Bayes. Sin embargo, una advertencia respecto a la interpretación del teorema de de Finetti en un contexto bayesiano es que la intercambiabilidad no especifica cuál debe ser el prior y la verosimilitud, solo que existe tal pareja.

Una visión alternativa del teorema de de Finetti como motivación para un enfoque bayesiano, comunicada por Stephen Walker y centrada en la predicción, es la siguiente. Sea $X_1, X_2, \ldots$ una secuencia de observables independientes, y supongamos que el objetivo es predecir la siguiente observación con base en las ya observadas. El analista comienza el proceso con dos elementos:

- Una (estimación de) la distribución predictiva para X_1.
- Una regla para actualizar dicha estimación basada en una nueva observación.

Es importante mencionar que, al menos por ahora, la regla de actualización no tiene que ser la actualización predictiva bayesiana mencionada brevemente en la Sección 4.2.4. Cuando se observa $X_1 = x_1$, el analista actualiza la estimación inicial basada en la regla mencionada anteriormente para obtener una distribución predictiva para X_2. Luego, se observa $X_2 = x_2$ y se obtiene una distribución predictiva para X_3. El proceso puede continuar indefinidamente, pero nos detendremos aquí para reflexionar sobre la estructura. En particular, ¿depende la distribución predictiva de X_3 del orden de las observaciones (x_1, x_2)? En otras palabras, ¿sería la distribución predictiva de X_3 la misma si hubiéramos observado (x_2, x_1) en su lugar?

Esta pregunta solo puede responderse conociendo la regla de actualización del analista. Sin embargo, si ocurre que el orden de las observaciones previas no importa, lo cual es bastante intuitivo dado que la fuente de datos es independiente, entonces se deduce del teorema de de Finetti que la regla de actualización del analista debe ser la regla bayesiana discutida en la Sección 4.2.4.

4.5. Eleccion de priors

Hemos mencionado anteriormente varias razones para adoptar un enfoque bayesiano. Sin embargo, ninguna de estas justificaciones indica qué prior elegir para un problema dado; en el mejor de los casos, los resultados simplemente establecen que existe un prior "razonable". ¿Qué se puede hacer si uno quiere adoptar un enfoque bayesiano pero no sabe qué prior elegir? Aquí hay algunas ideas.

4.5.1. Elicitación de priors

La elicitación de priors implica mantener discusiones con expertos para codificar su conocimiento previo sobre el problema en cuestión en una distribución de probabilidad. Este es un desafío por varias razones. Primero, puede ser un proceso que consume mucho tiempo. Segundo, incluso los expertos (que a menudo tienen poco o ningún conocimiento de probabilidad y estadística) pueden tener dificultades para comunicar sus creencias sobre el parámetro desconocido de una manera precisa (y consistente) que permita a un estadístico convertirlas en una distribución a priori. En resumen, este paso de elicitación es difícil de llevar a cabo y rara vez se realiza en su máxima extensión.

4.5.2. Priors convenientes

Como vimos en la Sección 4.3, hay algunos priors que son particularmente convenientes para el modelo en cuestión. Los priors conjugados son un conjunto de priors convenientes. Para ampliar el conjunto de priors conjugados, se pueden considerar mezclas de priors conjugados. El problema es que puede ser difícil confiar en los resultados de un análisis bayesiano basado en una suposición poco realista desde el principio. Durante años, este fue el único tipo de análisis bayesiano que se podía realizar, ya que, de lo contrario, los cálculos eran demasiado complejos. Hoy en día, con computadoras rápidas y algoritmos avanzados, realmente no hay necesidad de limitarse a un conjunto de priors "convenientes". Así que los priors conjugados, entre otros, son en cierto modo una cosa del pasado[14].

4.5.3. Muchos priors candidatos y Bayes robusto

Una alternativa a elegir un prior conveniente es considerar una clase de priors razonables y relativamente convenientes, analizar cada posterior candidato de manera individual y decidir si los resultados son sensibles a la elección del prior. Aquí hay un ejemplo tomado de Ghosh et al. (2006, Sec. 3.6).

Ejemplo 4.7

Supongamos que X sigue una distribución Pois(θ). Supongamos además que se cree que el prior para θ es continuo con percentiles 50 y 75 iguales a 2 y 4, respectivamente. Si estos son los únicos insumos previos, entonces los siguientes son tres candidatos para Π:

1. $\Pi_1 : \Theta \sim \text{Exp}(a)$ con $a = \log(2)/2$.
2. $\Pi_2 : \log \Theta \sim \mathcal{N}(\log(2), (\log(2)/0{,}67)^2)$.
3. $\Pi_3 : \log \Theta \sim \text{Cauchy}(\log 2, \log 2)$.

x	0	1	2	3	4	5	10	15	20	50
Π_1	0.75	1.49	2.23	2.97	3.71	4.46	8.17	11.88	15.60	37.87
Π_2	0.95	1.48	2.11	2.81	3.56	4.35	8.66	13.24	17.95	47.02
Π_3	0.76	1.56	2.09	2.63	3.25	3.98	8.87	14.07	19.18	49.40

Cuadro 4.1: Valores esperados a posteriori $\mathbb{E}(\Theta \mid x)$ para distintos priors y valores de x.

Bajo estas elecciones de prior, se puede calcular la media a posteriori. La Tabla 4.1 lista estos valores para distintos x. Aquí vemos que cuando x es relativamente pequeño (es decir, $x \leqslant 10$), la elección del prior no afecta mucho. Sin embargo, cuando x es algo grande, las medias a posteriori parecen variar significativamente.

Existen otros enfoques relacionados que definen una gran clase Γ de priors que son razonables en cierto sentido y buscan obtener cotas superiores e inferiores para ciertas cantidades a posteriori de interés. Veremos un resultado de este tipo que acota la media a posteriori $\psi(\theta)$ sobre una clase de priors unimodales simétricos. Para más detalles, ver Ghosh et al. (2006, Teorema 3.6).

[14] Ocasionalmente aparecen en niveles superiores de priors jerárquicos...

Teorema 4.4

Supongamos que los datos X tienen una densidad p_θ, con $\theta \in \mathbb{R}$, y consideremos una clase Γ de priors unimodales simétricos alrededor de θ_0, es decir,

$$\Gamma = \{\pi : \pi \text{ es simétrica y unimodal alrededor de } \theta_0\}.$$

Para una función real ψ dada, tenemos las siguientes cotas

$$\sup_{\pi\in\Gamma} \mathbb{E}_\pi(\psi(\Theta) \mid x) = \sup_{r>0} \frac{\int_{\theta_0-r}^{\theta_0+r} \psi(\theta)p_\theta(x)\,d\theta}{\int_{\theta_0-r}^{\theta_0+r} p_\theta(x)\,d\theta}.$$

$$\inf_{\pi\in\Gamma} \mathbb{E}_\pi(\psi(\Theta) \mid x) = \inf_{r>0} \frac{\int_{\theta_0-r}^{\theta_0+r} \psi(\theta)p_\theta(x)\,d\theta}{\int_{\theta_0-r}^{\theta_0+r} p_\theta(x)\,d\theta}.$$

4.5.4. Priors objetivos o no informativos

La idea principal detrás de los priors objetivos es elegir un prior que tenga un impacto mínimo en el posterior; en otras palabras, los priors objetivos permiten que los datos guíen el análisis. Existen básicamente tres enfoques en Bayes objetivo:

- Definir una "distribución uniforme" con respecto a la geometría del espacio de parámetros.
- Minimizar una medida adecuada de información en el prior.
- Elegir un prior de manera que las inferencias posteriores resultantes (por ejemplo, los intervalos creíbles) posean algunas propiedades frecuentistas deseables.

Sorprendentemente, en problemas con un solo parámetro, un único prior cumple con los tres criterios.

Definición 4.2

El prior de Jeffreys para θ tiene densidad $\pi(\theta) \propto (\det\{I_X(\theta)\})^{1/2}$, donde $I_X(\cdot)$ denota la matriz de información de Fisher.
Es interesante notar que el prior de Jeffreys es una distribución uniforme en Θ si, en lugar de la geometría euclidiana usual, se considera la geometría inducida por la métrica riemanniana, la cual está determinada por la información de Fisher; para más detalles, ver Ghosh y Ramamoorthi (2003). Cuando θ es un parámetro de localización, la información de Fisher es constante y la geometría inducida por ella es exactamente la geometría usual; por lo tanto, el prior de Jeffreys para θ es, en este caso, una distribución uniforme ordinaria, aunque suele ser impropia. Además, se puede demostrar que el prior de Jeffreys minimiza la divergencia de Kullback-Leibler asintótica entre el prior y el posterior. También hay resultados que muestran cómo el prior de Jeffreys produce conjuntos creíbles posteriores con una cobertura frecuentista aproximadamente nominal. Ghosh et al. (2006) proporciona una descripción clara de estos hechos.
Existen otras nociones de priors objetivos (por ejemplo, los priors invariantes), pero algo que tienen en común es que no integran a uno; es decir, son impropios. Por ejemplo,

en un problema de localización, tanto el prior de Jeffreys como los priors invariantes corresponden a la medida de Lebesgue en $(-\infty, \infty)$, la cual no es una medida finita. Más generalmente, los priors invariantes de Haar (izquierda y derecha) en problemas de transformación de grupos suelen ser impropios. Esto plantea una pregunta natural: ¿puede la teoría de la probabilidad en general, y el teorema de Bayes en particular, extenderse al caso impropio? Existen esencialmente dos maneras de abordar este problema: permitir que las probabilidades sean infinitas o eliminar la suposición de aditividad numerable. En ambos casos, muchos de los resultados conocidos en probabilidad deben ser descartados o demostrados nuevamente. Sin embargo, existen versiones del teorema de Bayes que se cumplen para probabilidades impropias o finitamente aditivas. No obstante, estos temas son demasiado técnicos para abordarlos aquí.

4.6. Teoría Bayesiana en grandes muestras

4.6.1. Configuración

La teoría de grandes muestras en el contexto clásico es útil para derivar procedimientos estadísticos en casos donde las distribuciones de muestreo exactas no están disponibles. Antes de que el cálculo del poder estadístico fuera tan accesible, este tipo de teoría asintótica era la única esperanza para manejar problemas no triviales. En el contexto bayesiano, existe una teoría asintótica correspondiente.

Además de la obvia aproximación asintótica de las probabilidades posteriores, que puede simplificar los cálculos de varias maneras, una consecuencia importante del Teorema 6 es que, bajo condiciones mínimas, la elección del prior es irrelevante cuando el tamaño de la muestra es grande. Dado que la elección del prior es el único obstáculo real en el análisis bayesiano, este es un resultado fundamental. Existen otros resultados de convergencia bayesiana más débiles (por ejemplo, la consistencia del posterior, Ejercicio 19) que pueden ser discutidos más adelante.

Otra aplicación de la teoría asintótica bayesiana es como una herramienta para identificar "malos priors" que no deberían ser utilizados. Es decir, si un prior en particular no admite un comportamiento deseable del posterior cuando $n \to \infty$, entonces hay algo incorrecto con ese prior. Esto es especialmente útil en problemas bayesianos no paramétricos, donde no es tan fácil definir un "buen prior" basado en intuición o experiencia. De hecho, la principal motivación del reciente auge de trabajos en teoría asintótica bayesiana es ayudar a identificar buenos priors para su uso en problemas no paramétricos desafiantes.

4.6.2. Aproximación de Laplace

La *aproximación de Laplace* es una técnica maravillosamente simple pero muy poderosa para aproximar ciertos tipos de integrales; ver Ghosh et al. (2006, Sec. 4.3). La aproximación en sí no está directamente relacionada con el análisis bayesiano, pero los problemas de integración en los que es útil aparecen frecuentemente en estadística bayesiana. También es la base del popular Criterio de Información Bayesiano (BIC, por sus siglas en inglés) en la selección de

modelos.

Consideremos una integral de la forma

$$\text{integral} = \int q(\theta) e^{nh(\theta)}\, d\theta,$$

donde tanto q como h son funciones suaves de una cantidad θ de dimensión p; es decir, la integral anterior se toma sobre $\mathbb{R}^p$. Aquí se asume que n es grande o tiende a ∞. Sea $\hat{\theta}$ el único máximo de h. Entonces, la aproximación de Laplace proporciona una forma de calcular la integral sin necesidad de integración, ¡solo mediante optimización!

Teorema 4.5

Sean h' y h'' las derivadas de h y sea $\det(\cdot)$ el determinante de una matriz. Entonces, cuando $n \to \infty$,

$$\int q(\theta) e^{nh(\theta)}\, d\theta = q(\hat{\theta}) e^{nh(\hat{\theta})} (2\pi)^{p/2} n^{-p/2} \det\{-h''(\hat{\theta})\}^{-1/2} \{1 + O(n^{-1})\}.$$

Nótese que $h''(\hat{\theta})$ es definida negativa, por hipótesis, por lo que $-h''(\hat{\theta})$ es definida positiva.

Demostración. Aquí esbozamos el caso $p = 1$. La primera observación es que, si h tiene un máximo único en $\hat{\theta}$ y n es muy grande, entonces la contribución principal a la integral proviene de un pequeño intervalo alrededor de $\hat{\theta}$, digamos $\hat{\theta} \pm a$.
En segundo lugar, dado que este intervalo es pequeño y $q(\theta)$ es una función suave, es razonable aproximar $q(\theta)$ por la función constante $q(\hat{\theta})$ para $\theta \in (\hat{\theta} - a, \hat{\theta} + a)$. Ahora, la idea es usar una aproximación de Taylor de $h(\theta)$ hasta orden dos alrededor de $\theta = \hat{\theta}$:

$$h(\theta) = h(\hat{\theta}) + h'(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2} h''(\hat{\theta})(\theta - \hat{\theta})^2 + \text{error}.$$

Dado que $h'(\hat{\theta}) = 0$ por definición de $\hat{\theta}$, al sustituir esto en el término exponencial de la integral (e ignorando los términos de error), obtenemos:

$$\begin{aligned}
\text{integral} &\approx \int_{\hat{\theta}-a}^{\hat{\theta}+a} q(\theta) \exp\{n[h(\hat{\theta}) + (1/2)h''(\hat{\theta})(\theta - \hat{\theta})^2]\}\, d\theta \\
&\approx \int_{\hat{\theta}-a}^{\hat{\theta}+a} q(\hat{\theta}) \exp\{nh(\hat{\theta}) - (\theta - \hat{\theta})^2/2\sigma^2\}\, d\theta \\
&= q(\hat{\theta}) e^{nh(\hat{\theta})} \int_{\hat{\theta}-a}^{\hat{\theta}+a} e^{-(\theta-\hat{\theta})^2/2\sigma^2}\, d\theta.
\end{aligned}$$

donde $\sigma^2 = [-nh''(\hat{\theta})]^{-1}$, que es pequeño. El último integrando se asemeja a una función de densidad normal, excepto que le falta el factor $(2\pi\sigma^2)^{-1/2}$. Multiplicamos y dividimos por esta cantidad para obtener

$$\text{integral} \approx q(\hat{\theta}) e^{nh(\hat{\theta})} (2\pi)^{1/2} n^{-1/2} [-h''(\hat{\theta})]^{-1/2},$$

lo cual es exactamente lo que buscábamos. ■

Una aplicación simple pero interesante de la aproximación de Laplace es la aproximación de

Stirling para $n!$ (Ejercicio 16). La conexión con la distribución normal en el bosquejo de la demostración anterior es la clave para el teorema principal sobre la normalidad del posterior, que se discutirá a continuación.

4.6.3. Teorema de Bernstein–von Mises

Sea $\hat{\theta}_n$ una secuencia consistente de soluciones a la ecuación de verosimilitud, y sea $I(\theta)$ la información de Fisher. El teorema de Bernstein–von Mises establece que la distribución a posteriori de $n^{1/2}(\Theta - \hat{\theta}_n)$ es aproximadamente normal con media cero y varianza $I(\theta^*)^{-1}$ en probabilidad P_{θ^*} cuando $n \to \infty$.

Las condiciones requeridas para este resultado son esencialmente las mismas que se utilizan para demostrar la normalidad a posteriori del estimador de máxima verosimilitud (MLE). En particular, además de las condiciones C1–C4 del Capítulo 3, asumimos lo siguiente:

C5. Para cualquier $\delta > 0$, con probabilidad P_{θ^*} igual a 1, existe $\varepsilon > 0$ tal que

$$\sup_{\theta:|\theta-\theta^*|>\delta} n^{-1}\{\ell_n(\theta) - \ell_n(\theta^*)\} \leqslant -\varepsilon$$

para todo n suficientemente grande, donde $\ell_n = \log L_n$.

C6. La densidad a priori $\pi(\theta)$ es continua y positiva en θ^*.

La condición C6 es fácil de verificar y se cumple para cualquier prior razonable. La condición C5, por otro lado, es un poco más desafiante, aunque se cumple en la mayoría de los ejemplos; ver Ejercicio 17. Nos enfocaremos aquí en el caso unidimensional de Θ, aunque un resultado similar se cumple para Θ de dimensión d con modificaciones obvias.

Teorema 4.6

Supongamos que se cumplen las condiciones C1–C6, y sea $\hat{\theta}_n$ una secuencia consistente de soluciones a la ecuación de verosimilitud. Definimos $Z_n = n^{1/2}(\Theta - \hat{\theta}_n)$ y sea $\tilde{\pi}_n(z)$ la densidad a posteriori de Z_n dado $X_1, \ldots, X_n$. Entonces, se cumple que

$$\int \left|\tilde{\pi}_n(z) - N(z \mid 0, I(\theta^*)^{-1})\right| dz \to 0 \quad \text{con probabilidad 1 bajo } P_{\theta^*}.$$

El mensaje aquí es que, bajo condiciones adecuadas, si el tamaño de la muestra es grande, entonces la distribución a posteriori de Θ se aproximará a una normal con media $\hat{\theta}_n$ y varianza $[nI(\theta^*)]^{-1}$. Bajo las mismas condiciones, $I(\theta^*)$ puede ser reemplazada por n^{-1} veces la información de Fisher observada; ver el argumento de la aproximación de Laplace más adelante.

El tipo de convergencia que se discute aquí es la convergencia en L_1 de las densidades, que es más fuerte que la convergencia "en distribución" usual. Es decir, según el Teorema 6, la esperanza de cualquier función de Θ puede ser aproximada por la misma esperanza bajo la distribución normal límite.

Además de ser una herramienta para la inferencia aproximada a posteriori, estos resultados pueden ser útiles para el desarrollo de métodos computacionales que permitan simular desde el posterior exacto.

Los detalles de la demostración se encuentran en Ghosh et al. (2006, Sec. 4.1.2). Aquí se presenta un bosquejo de la demostración basado en la aproximación de Laplace. La idea es aproximar la log-verosimilitud mediante una función cuadrática a través de una aproximación de Taylor; recordemos que hicimos algo similar en la demostración de la normalidad asintótica del estimador de máxima verosimilitud (MLE).

Sea $Z = n^{1/2}(\Theta - \hat{\theta})$ el parámetro reescalado. Entonces,

$$\Pi_n(-a < Z < a) = \Pi_n(\hat{\theta} - an^{-1/2} < \Theta < \hat{\theta} + an^{-1/2}) = \frac{\text{num}}{\text{den}}.$$

Definiendo

$$q(\theta) = \pi(\theta) \quad \text{y} \quad h(\theta) = \frac{1}{n}\log L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log p_\theta(X_i),$$

entonces el denominador anterior puede ser aproximado (vía Laplace) por

$$\text{den} = \int L_n(\theta)\pi(\theta)\,d\theta = \int \pi(\theta)e^{nh(\theta)}\,d\theta \approx L_n(\hat{\theta})\pi(\hat{\theta})(2\pi/nv)^{1/2},$$

donde $v = -h''(\hat{\theta}) = -n^{-1}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log p_\theta(X_i)\big|_{\theta=\hat{\theta}}$ es la información de Fisher observada.

El numerador puede aproximarse de manera similar:

$$\text{num} = \int_{\hat{\theta}-an^{-1/2}}^{\hat{\theta}+an^{-1/2}} L_n(\theta)\pi(\theta)\,d\theta \approx L_n(\hat{\theta})\pi(\hat{\theta})n^{-1/2}\int_{-a}^{a} e^{-vu^2/2}\,du.$$

Tomando el cociente entre el numerador y el denominador, obtenemos

$$\Pi_n(-a < Z < a) = \frac{\text{num}}{\text{den}} \approx \frac{L_n(\hat{\theta})\pi(\hat{\theta})n^{-1/2}\int_{-a}^{a} e^{-vu^2/2}\,du}{L_n(\hat{\theta})\pi(\hat{\theta})(2\pi/nv)^{1/2}} = \int_{-a}^{a}\frac{\sqrt{v}}{\sqrt{2\pi}}e^{-vu^2/2}\,du,$$

y esta última expresión es la probabilidad de que una variable aleatoria normal con media cero y varianza v^{-1} esté en el intervalo $(-a, a)$, que es lo que queríamos demostrar.

Es intuitivamente claro, a partir del teorema de Bernstein–von Mises, que la media a posteriori debería comportarse como la secuencia consistente de soluciones $\hat{\theta}_n$. El siguiente resultado formaliza esta intuición.

Teorema 4.7

$$n^{1/2}(\tilde{\theta}_n - \hat{\theta}_n) \to 0 \quad \text{y} \quad n^{1/2}(\tilde{\theta}_n - \theta^*) \to N(0, I(\theta^*)^{-1}).$$

Demostración. Ver Ghosh y Ramamoorthi (2003, p. 39). El Ejercicio 18 describe la demostración de un resultado más débil utilizando la aproximación de Laplace. □

4.7. Conclusiones

4.7.1. Más detalles sobre la inferencia bayesiana

Aquí no intentamos profundizar demasiado en la metodología y filosofía bayesiana. Un curso formal de inferencia bayesiana tomaría estos puntos con mayor seriedad. El libro de Berger (1985) proporciona una justificación "filosófica" interesante del análisis bayesiano, junto con muchas otras cosas. Lo mismo puede decirse de Ghosh et al. (2006).

El modelado y la metodología bayesiana son desafiantes. Un buen punto de partida para aprender sobre estos temas es el libro de Gelman et al. (2004), que ofrece una perspectiva moderna del análisis bayesiano desde un enfoque aplicado. Existen desafíos metodológicos más allá de la justificación filosófica del enfoque bayesiano. Un ejemplo importante es el caso de pruebas de hipótesis cuando la hipótesis nula es un singleton (o algún otro conjunto con medida de Lebesgue cero). En este caso, la formulación simple de la prueba de hipótesis bayesiana deja de ser válida y se requieren enfoques alternativos.

Los *factores de Bayes* suelen ser la opción típica en estos casos; ver Kass y Raftery (1995) y Ghosh et al. (2006). Sin embargo, curiosamente, dado que estos no son funciones de la distribución a posteriori, en realidad no son completamente bayesianos.

El punto es que, aunque las ideas presentadas aquí son bastante simples, en general constituyen una simplificación excesiva. Finalmente, cabe mencionar que la computación es crucial para los bayesianos, ya que casi nada admite expresiones en forma cerrada. Se emplean diversas formas de Monte Carlo, que es un método para la integración numérica, y el mejor recurso para aprender sobre estos temas es el libro de Robert y Casella (2004).

En estas notas, el objetivo es introducir algunas ideas clave dentro del marco bayesiano. Algunos de estos puntos serán útiles en nuestra cobertura de teoría de decisión estadística más adelante. En cualquier caso, todos los estadísticos deben estar familiarizados con los fundamentos de todas las principales ideas en estadística; limitarse a una sola perspectiva no es más que eso, una limitación.

4.7.2. Sobre Bayes y el principio de verosimilitud

Recordemos los breves comentarios sobre el *principio de verosimilitud* al final del Capítulo 3. Este principio es una postura filosófica que sostiene que la inferencia final debe depender únicamente de los datos y el modelo a través de la función de verosimilitud observada. Aunque esta es una posición algo extrema, Birnbaum (1962) demostró que el principio de verosimilitud se derivaba de dos principios bastante razonables.

Aunque este teorema ha sido refutado recientemente (e.g., Evans 2013; Martin y Liu 2014; Mayo 2014), el hecho es que el resultado de Birnbaum ayudó a convencer a muchos estadísticos de considerar seriamente el enfoque bayesiano. La conclusión rápida del teorema de Birnbaum es que el principio de verosimilitud es deseable (porque, según el teorema, es equivalente a otras propiedades deseables), y dado que el único enfoque bayesiano clásico (con un "prior subjetivo", véase más adelante) satisface el principio de verosimilitud, el único enfoque lógico es el bayesiano.

Esta atención dada a los métodos bayesianos tras el descubrimiento de Birnbaum puede considerarse como un catalizador de los desarrollos teóricos, metodológicos y computacionales en los últimos 20-30 años.

Aquí quiero hacer dos observaciones sobre el principio de verosimilitud y el análisis bayesiano. Primero, la afirmación de que "el enfoque bayesiano obedece el principio de verosimilitud" es incompleta.

Esta afirmación es cierta si el prior se basa en algunas consideraciones subjetivas. Sin embargo, el enfoque estándar actual es utilizar priors predeterminados "no informativos", como el prior de Jeffreys, que depende del propio modelo. El uso de inferencia bayesiana con un prior predeterminado de este tipo *no* satisface el principio de verosimilitud.

En segundo lugar, en algunos casos, el hecho de que el enfoque bayesiano satisfaga el principio de verosimilitud podría considerarse una desventaja, o al menos no ser intuitivamente atractivo. Un ejemplo es el caso del muestreo e inferencia en poblaciones finitas. Buenos libros sobre enfoques bayesianos y no bayesianos a este problema son Ghosh y Meeden (1997) y Hedayat y Sinha (1991), respectivamente.

Para tales problemas, se dedica un esfuerzo considerable al diseño de un buen esquema de muestreo, de modo que la muestra obtenida sea "representativa" en cierto sentido. Sin embargo, el hecho de que el posterior bayesiano dependa únicamente de la función de verosimilitud observada implica que *el diseño del muestreo es irrelevante para la inferencia*. Esto es algo contraintuitivo e incluso puede ser controvertido.

Por lo tanto, en este caso, algunos podrían decir que obedecer el principio de verosimilitud es una desventaja para el enfoque bayesiano, aunque no creo que sea tan simple.

4.7.3. Sobre la etiqueta "Bayesiano"

Los estadísticos suelen usar las etiquetas "bayesiano' y "no bayesiano". Personalmente, no me gusta esto porque creo que (a) es algo divisivo y (b) puede dar una falsa impresión de que algunos problemas pueden resolverse con análisis bayesiano mientras que otros no. Todos trabajamos en los mismos problemas, y deberíamos estar abiertos a considerar diferentes perspectivas.

Por esta razón, aunque me considero conocedor de la perspectiva bayesiana, no me clasificaría a mí mismo como un bayesiano, en sí mismo. De hecho, podría ser considerado muy "no bayesiano" porque tengo dudas sobre el significado de las probabilidades a posteriori bayesianas, etc., en general, y he trabajado en el desarrollo de un nuevo marco teórico diferente de, pero no ortogonal a, las ideas bayesianas existentes.

4.7.4. Sobre la "objetividad"

Recuerden que, al comienzo del curso, mencioné que el problema estadístico está mal planteado y que no hay una "respuesta correcta". Cada enfoque hace suposiciones para poder desarrollar una teoría.

En el contexto bayesiano, se asume la existencia de una distribución a priori. En el enfoque

frecuentista, se asume que los datos observados son uno de esos resultados típicos. Por lo tanto, en cualquier caso, no podemos evitar hacer algún tipo de suposición.

Sean escépticos ante argumentos, estadísticos o de otro tipo, que afirman ser "objetivos". La siguiente subsección describe brevemente algunas preocupaciones sobre el uso de priors objetivos y, en general, sobre el uso de la probabilidad para la inferencia.

4.7.5. Sobre el papel de la probabilidad en la inferencia estadística

Recordemos que la motivación básica del enfoque bayesiano es que la probabilidad es la herramienta correcta para resumir la incertidumbre. ¿Existe alguna justificación para que la probabilidad sea la herramienta adecuada? Para mí, la justificación es clara cuando hay una distribución a priori significativa para θ. En este caso, el problema de inferencia estadística se reduce a un cálculo de probabilidad.

Sin embargo, el problema científico típico es aquel en el que se conoce poco o nada sobre θ, lo que implica que no hay una distribución a priori significativa disponible, o bien que se tiene reticencia a utilizar la poca información disponible por miedo a influir en los resultados.

En tales casos, se puede introducir un prior predeterminado no informativo para θ y llevar a cabo un análisis bayesiano. Este tipo de enfoque bayesiano se ha aplicado con éxito en muchos problemas, pero eso no significa que no podamos cuestionar su uso. La principal preocupación es la siguiente:

La distribución a priori establece, de manera efectiva, la escala en la que se interpretan las probabilidades a posteriori. Esto se vuelve claro si se piensa en el posterior como un prior actualizado basado en los datos observados. Otra forma extrema de entender esto es que el prior y el posterior tienen los mismos conjuntos nulos. Así, al menos en muestras finitas, el prior juega un papel en la escala de los valores numéricos del posterior.

Cuando el prior en sí mismo no tiene una escala significativa, por ejemplo, si es impropio, ¿cómo se puede asignar un significado a las probabilidades a posteriori correspondientes? Solo se puede asignar significado a las probabilidades a posteriori (de ciertos subconjuntos de Θ) asintóticamente, que es esencialmente lo que establece el teorema de Bernstein–von Mises.

Por lo tanto, cuando el prior es no informativo, las probabilidades a posteriori carecen de una interpretación significativa, al menos en muestras finitas.

Otro problema con la probabilidad es la regla de la complementariedad, es decir, $P(A^c) = 1 - P(A)$. Esta propiedad tiene sentido cuando el objetivo es predecir el resultado de algún experimento a realizar—X estará exactamente en uno de A o A^c. Este es el contexto en el que se desarrolló originalmente la probabilidad, es decir, cuando el objetivo es predecir el resultado de un experimento en el que se dispone de toda la información sobre dicho experimento.

Sin embargo, argumentaría que esto es muy diferente del problema de la inferencia estadística. La cantidad de interés no es una realización de algún experimento, sino más bien un valor fijo sobre el cual tenemos información limitada, en la forma de un modelo y datos observados. ¿Por qué, entonces, la probabilidad debería ser la herramienta adecuada para resumir nuestra incertidumbre?

Por ejemplo, no creo que la regla de la complementariedad sea lógica en el problema de

inferencia. En realidad, θ está exactamente en uno de A y A^c, pero con la información limitada de los datos, puede no ser razonable llegar a una conclusión tajante de que A está fuertemente apoyado y A^c está débilmente apoyado, o viceversa. De hecho, parece bastante razonable que los datos no puedan apoyar fuertemente ni A ni A^c, en cuyo caso, la "probabilidad" de estos eventos debería sumar un número menor que 1.

Ninguna probabilidad puede satisfacer esta propiedad, por lo que quizás la probabilidad no sea la herramienta correcta. Podrías preguntarte si existe algo más que pueda acomodar esto, y la respuesta es SÍ, una *función de creencia* (*belief function*). Esto es parte de la motivación detrás del marco de *modelos inferenciales* (IM, por sus siglas en inglés); ver Martin y Liu (2013, 2015a,b,c) y Liu y Martin (2015).

4.8. Ejercicios

1. *a*) Consideremos algunas medidas σ-finitas genéricas μ, ν y λ, todas definidas en el espacio medible $(\mathbb{X}, \mathcal{A})$. Supongamos que $\mu \ll \nu$ y $\nu \ll \lambda$. Demuestre que $\mu \ll \lambda$ y que la derivada de Radon-Nikodym de μ con respecto a λ satisface una regla de la cadena, es decir,

$$\frac{d\mu}{d\lambda} = \frac{d\mu}{d\nu} \cdot \frac{d\nu}{d\lambda} \quad (\lambda\text{-almost everywhere}).$$

   b) Use la parte (a) para demostrar el Corolario 1.

2. Dado $\theta \in (0, 1)$, supongamos que $X \sim \text{Bin}(n, \theta)$.

   a) Demuestre que la familia $\text{Beta}(a, b)$ es conjugada.

   b) Bajo un prior $\text{Beta}(a, b)$, encuentre la media y la varianza a posteriori. Proporcione una interpretación de lo que ocurre cuando $n \to \infty$.

   c) Considere el prior $\pi(\theta) = [\theta(1-\theta)]^{-1}$ para $\theta \in (0, 1)$. Demuestre que el prior es impropio pero que, siempre que $0 < x < n$, el posterior resulta ser propio.

3. Suponga que $X = (X_1, \ldots, X_n)$ son iid $\text{Unif}(0, \theta)$ y que θ tiene un prior $\text{Unif}(0, 1)$.

   a) Encuentre la distribución a posteriori de θ.

   b) Encuentre la media a posteriori.

   c) Encuentre la mediana a posteriori.

   d) Encuentre la moda a posteriori.

4. Considere el escenario del Ejemplo 2. Diseñe un estudio de simulación para evaluar la probabilidad de cobertura del intervalo creíble del 95 % para θ en varias elecciones de θ, tamaño de muestra n y los hiperparámetros gamma (a, b).

5. Encuentre el intervalo creíble marginal del 95 % para μ basado en los cálculos del Ejemplo 4. Demuestre que la probabilidad de cobertura frecuentista del intervalo creíble del 95 % es 0.95.

6. Problema 7.11 en Keener (2010, p. 126).

7. Problema 7.12 en Keener (2010, p. 126).

8. Considere el caso de Poisson en el Ejemplo 2.

   a) Describa un enfoque para simular a partir de la distribución predictiva. ¿Cómo utiliza esta muestra a posteriori para producir un intervalo de predicción del 95 % para la siguiente observación X_{n+1}?

   b) Proponga un intervalo de predicción no bayesiano del 95 % para X_{n+1}.

   c) Diseñe un estudio de simulación para evaluar el desempeño (cobertura y longitud) de sus intervalos de predicción bayesiano y no bayesiano. Pruebe varios valores de θ, tamaño de muestra n e hiperparámetros previos (a, b).

9. Considere un modelo con densidad $p_\theta(x) = h(x)\exp\{\theta x - A(\theta)\}$, es decir, una familia exponencial de un solo parámetro. Sea $X_1, \ldots, X_n$ una muestra iid de p_θ.

   a) Considere un prior para θ, con densidad $\pi(\theta) = g(\theta)e^{\eta\theta - B(\eta)}$. Demuestre que este es conjugado y escriba la densidad a posteriori correspondiente.

   b) Considere el caso especial donde $\pi(\theta) \propto e^{\eta\theta - mA(\theta)}$ para valores fijos de (η, m), y suponga que $\pi(\theta)$ se anula en la frontera del espacio paramétrico.

      1) Demuestre que la media a priori de $A'(\Theta)$ es η/m.
      2) Demuestre que la media a posteriori de $A'(\Theta)$ es un promedio ponderado entre la media a priori y la media muestral.

10. Demuestre que $X_1, \ldots, X_n$ son intercambiables si y solo si, para todo $k \leqslant n$, todos los k-tuplas $(X_{i_1}, \ldots, X_{i_k})$ tienen la misma distribución conjunta.

11. Demuestre que variables aleatorias condicionalmente iid, que satisfacen (3), son intercambiables. Puede asumir la existencia de densidades si esto facilita la demostración.

12. Problema 7.14(b) en Keener (2010, p. 127). **[Sugerencia: Use "covarianza iterada."]**

13. Suponga que $X|\theta \sim N(\theta, 1)$ y el objetivo es probar $H_0 : \theta \leqslant \theta_0$. Considere la clase Γ de priors simétricos y unimodales sobre θ_0. Utilice el Teorema 4.4 para obtener cotas superior e inferior para $\Pi_x(H_0)$, la probabilidad a posteriori de H_0. **[Sugerencia: Se involucrará el valor-p.]**

14. Sea h una función diferenciable uno a uno y considere la reparametrización $\xi = h(\theta)$. Sea π_θ y π_ξ los priors de Jeffreys para θ y ξ, respectivamente. Demuestre que

$$\pi_\theta(u) = \pi_\xi(h(u))\,|h'(u)|\,.$$

    Esta propiedad establece que el prior de Jeffreys es invariante ante reparametrizaciones suaves.

15. La integración de Monte Carlo es una parte importante del análisis bayesiano moderno. La idea clave es reemplazar integrales difíciles con simulación y promedios simples. En otras palabras, si el objetivo es evaluar $\mathbb{E}[h(X)] = \int h(x)\,dP(x)$ para una medida de probabilidad P, entonces una estrategia de Monte Carlo consiste en simular $\{X_t : t = 1, \ldots, T\}$ independientes[15] de P y aproximar $\mathbb{E}[h(X)]$ mediante $T^{-1}\sum_{t=1}^{T} h(X_t)$.

    Suponga que $X \sim \text{Unif}(0, 1)$. El objetivo es utilizar integración de Monte Carlo para aproximar la función generadora de momentos $M_X(u)$ de X en el intervalo $u \in (-2, 2)$.

a) Describa su algoritmo y utilice la desigualdad de Hoeffding y el lema de Borel–Cantelli (Capítulo 1) para demostrar la consistencia de su estimador de Monte Carlo. Es decir, si $\hat{M}_X(u)$ es su estimador de Monte Carlo, entonces pruebe que $\hat{M}_X(u) \to M_X(u)$ con probabilidad 1 para cada u fijo.

b) Implemente su método, trace un gráfico de su aproximación junto con la verdadera función generadora de momentos superpuesta y comente sobre la calidad de la aproximación.
[Sugerencias: (i) 1000 muestras de Monte Carlo deberían ser suficientes; (ii) puede usar las mismas muestras de Monte Carlo para cada punto u en la cuadrícula.**]**

16. Utilice la aproximación de Laplace para derivar la *fórmula de Stirling*:
$$n! \approx n^{n+(1/2)} e^{-n} \sqrt{2\pi}, \quad \text{para } n \text{ grande.}$$
[Sugerencia: Use la función gamma: $n! = \Gamma(n+1) = \int_0^\infty e^{-u} u^n \, du$.**]**

17. Verifique que $N(\theta, 1)$ satisface la Condición C5 en el marco de Bernstein–von Mises.

18. *a*) Escriba $\pi_n(\theta) \propto L_n(\theta)\pi(\theta)$, la densidad a posteriori de un parámetro real. Para una función $g(\theta)$, la media a posteriori se define como
$$\mathbb{E}\{g(\Theta) \mid X\} = \frac{\int_{-\infty}^{\infty} g(\theta) L_n(\theta)\pi(\theta)\, d\theta}{\int_{-\infty}^{\infty} L_n(\theta)\pi(\theta)\, d\theta}.$$
Si $g(\theta)$ es suficientemente suave, utilice la aproximación de Laplace tanto en el numerador como en el denominador para obtener una fórmula para $\mathbb{E}\{g(\Theta) \mid X\}$ en términos del estimador de máxima verosimilitud $\hat{\theta}_n$.

b) Sea $\tilde{\theta}_n = \int \theta \pi_n(\theta)\, d\theta$ la media a posteriori basada en datos iid $X_1, \ldots, X_n$. Utilice la aproximación de Laplace para probar que $\tilde{\theta}_n$ es un estimador consistente de θ.
[Sugerencia: Demuestre que $(\tilde{\theta}_n - \theta^*)^2 \to 0$ en probabilidad bajo P_{θ^*}. Primero use la desigualdad de Jensen, luego tome $g(\theta) = (\theta - \theta^*)^2$ para la aproximación de Laplace.**]**

19. Escriba Π_n para la distribución a posteriori basada en una muestra iid de tamaño n; para simplificar, suponga que el parámetro es un escalar, aunque las ideas son mucho más generales. Se dice que la distribución a posteriori es *consistente* en θ^* si
$$\Pi_n(\{\theta : |\theta - \theta^*| > \varepsilon\}) \to 0 \quad \text{en probabilidad bajo } P_{\theta^*}, \text{ para todo } \varepsilon > 0, \text{ cuando } n \to \infty.$$
Como aplicación, sea $X_1, \ldots, X_n$ una muestra iid $N(\theta, 1)$ y considere un prior $\pi(\theta) \propto 1$, un prior constante. Utilice la desigualdad de Markov y propiedades básicas de las muestras normales para demostrar que la distribución a posteriori correspondiente es consistente en todo $\theta^* \in \mathbb{R}$.

20. Para la consistencia a posteriori, etc., el prior Π debe asignar suficiente masa cerca del valor verdadero del parámetro, digamos, θ^*. Una forma de garantizar esto es asumir que
$$\Pi(\{\theta : K(p_{\theta^*}, p_\theta) < \varepsilon\}) > 0, \quad \forall \varepsilon > 0,$$
donde K denota la divergencia de Kullback–Leibler introducida en las Notas I.
La condición anterior se interpreta como: "Π satisface la propiedad de Kullback–Leibler en θ^*."

a) Sea p_θ una familia exponencial. Encuentre $K(p_{\theta^*}, p_\theta)$.

b) Suponiendo regularidad de p_θ, argumente que un prior Π satisface la propiedad de Kullback–Leibler en θ^* si tiene una densidad positiva π en una vecindad de θ^*.

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,

$$\begin{aligned}\mathbb{E}[\delta_0(X) \mid T = t] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[I_{(-\infty,c]}(X_i) \mid T = t] \\ &= P(X_1 \leqslant c \mid T = t) = \Phi\left(\frac{c - t}{\sqrt{(n-1)/n}}\right), \qquad (2)\end{aligned}$$

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.

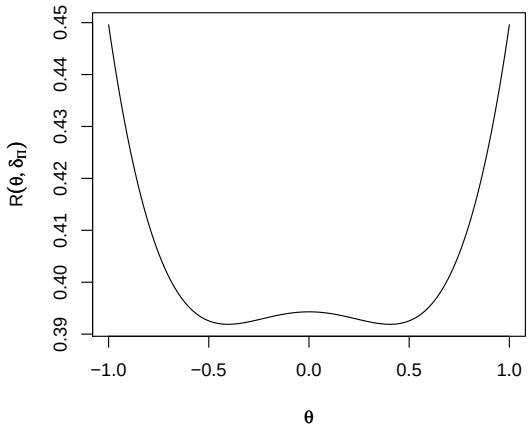


Figura 4: Grafica de la Funcion Riesgo del ejemplo 10.

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).

Capítulo 6: Teoría Asintótica

Teoría Estadística Avanzada

SIGLA DES124

PROF. JAIME LINCOVIL

6.1. Teoría Asintótica Estadística

- Cuando el modelo estadístico del estimador puntual o del estadístico tiene forma analitica conocida, no es necesario usar resultados asintóticos.
- Cuando el modelo estadístico de los anteriores no tiene una forma de encontrar por medio de métodos analiticos, este se puede aproximar por el modelo que tendría cuando el tamaño de la muestra es grande ($n \to \infty$).
- Es decir

$$P_{\text{asintótico}}(X \in A) \approx P_\theta(X \in A).$$

Sea $\{X_n\}$ una secuencia de vectores aleatorios definidos sobre el mismo espacio de probabilidad.

Definición 6.1 Tipos de convergencia de V.A

(i) Escribimos $X_n \to_{c.c.} X$ (casi ciertamente) ssi

$$P\Big[\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\Big] = 1.$$

(ii) Escribimos $X_n \to_p X$ (convergencia en probabilidad) ssi

$$\lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0. \quad \forall \epsilon > 0.$$

(iii) Escribimos $X_n \to {}_{L_r} X$ (converge a X en L_r) sii

$$\lim_{n\to\infty} E|X_n - X|^r = 0, \text{en que } r \in \mathcal{N}_+.$$

iv) Sea $F, F_n, n = 1, 2, \ldots$, una secuencia de funciones de distribución acumulada definidas sobre $\mathcal{R}^k$ y $P, P_n, n = 1, \ldots$, son las respectivas medidas de probabilidad. Decimos que $\{F_n\}$ convege a F debilmente (o $\{P_n\}$ converge a P debilmente) y escribimos $F_n \to_w F$ (o $P_n \to_w P$) si y solo si, para cada punto de continuidad x de F,

$$\lim_{n\to\infty} F_n(x) = F(x).$$

Nosotros decimos que $\{X_n\}$ converge a X en distribución y escribimos $X_n \to_d X$ si y solo si $F_{X_n} \to_w F_X$.

6.2. Interpretación de las convergencias

(i) El conjunto de puntos $\omega \in \Omega$ para los cuales X_n converge puntalmente a X tiene probabilidad 1.

(ii) Para todo nivel de proximidad ϵ, la probabilidad de que la diferencia entre X_n y X sea menor que ϵ converge a 1.

(iii) El valor esperado de orden r entre la diferencia entre X_n y X converge a 0.

(iv) La función de probabilidad acumulada F_n converge puntualmente a F en cada punto de continuidad de F.

Teorema 6.1 Implicancias entre convergencias

Sea $X, X_1, X_2, \ldots$ una secuencia de variables aleatorias definidos sobre el mismo espacio de probabilidad.
(i) Si $X_n \to_{c.c.} X$, entonces $X_n \to_p X$.
(ii) Si $X_n \to_{L_r} X$ para $r > 0$, entonces $X_n \to_p X$.
(iii) Si $X_n \to_p X$, entonces $X_n \to_d X$.
(iv) (Teorema de Skorohod). Si $X_n \to_d X$, entonces existen vectores aleatorios $Y, Y_1, Y_2, \ldots$ definimos sobre un mismo espacio de probabilidades tal que $P_Y = P_X$, $P_{Y_n} = P_{X_n}, n = 1, 2, \ldots$, and $Y_n \to_{c.c.} Y$.

6.3. Ordenes $O(\cdot), o(\cdot)$, $Op(\cdot)$ y $op(\cdot)$

Las secuencias reales $\{a_n\}$ y $\{b_n\}$,
- Satisface $a_n = O(b_n)$ ssi $|a_n| \leqslant c\,|b_n|$ para todo n y una constante c.
- Denotamos por $a_n = o(b_n)$ ssi $a_n/b_n \to 0$ cuando $n \to \infty$.

Definición 6.2 Orden de convergencia

Sea $X_1, X_2, \ldots$ y $Y_1, Y_2, \ldots$ dos secuencia de variables aleatorias definidas sobre un mismo espacio de probabilidad.
(i) $X_n = O(Y_n)$ c.c. si y solo si $P(|X_n| = O(|Y_n|)) = 1$.
(ii) $X_n = o(Y_n)$ c.c. si y solo si $X_n/Y_n \to_{c.c.} 0$.
(iii) $X_n = Op(Y_n)$ si y solo si, para cualquier $\epsilon > 0$, existe una constante $C_\epsilon > 0$ tal que $\sup_n P(|X_n| \geqslant C_\epsilon |Y_n|) < \epsilon$.
(iv) $X_n = op(Y_n)$ si y solo si $X_n/Y_n \to_p 0$.

6.3.1. Interpretación de las ordenes

(i) El evento $|X_n| \leqslant |Y_n|$ tiene probabilidad 1.

(ii) El conjunto de puntos $\omega \in \Omega$ para los cuales X_n/Y_n converge puntalmente a 0 tiene probabilidad 1.

(iii) Para todo nivel de proximidad ϵ, la probabilidad máxima para todo n de que la diferencia entre X_n y Y_n sea menor que ϵ esta acotada y depende de ϵ.

(iv) X_n/Y_n converge en probabilidad a 0.

6.4. Convergencia de funciones aleatorias

Para variables aleatorias X_n que convergen a X en algun sentido, nosotros a menudo queremos conocer si $g(X_n)$ converge a $g(X)$ en algun sentido. El siguiente resultado provee una respuestaa este tipo de preguntas.

Teorema 6.2 Convergencia de funciones de variables

Sea $X, X_1, X_2, \ldots$ una secuencia de variables aleatorias definidas sobre un mismo espacio de probabilidad y sea g una función medible de $\left(\mathcal{R}^k, \mathcal{B}^k\right)$ to $\left(\mathcal{R}^l, \mathcal{B}^l\right)$. Supongamos que g es continua a.s. P_X. Entonces,
(i) $X_n \to_{c.c.} X$ implica $g(X_n) \to_{c.c.} g(X)$;
(ii) $X_n \to_p X$ implica $g(X_n) \to_p g(X)$;
(iii) $X_n \to_d X$ implica $g(X_n) \to_d g(X)$.

6.5. Teorema del Límite Central

El Teorema del Límite Central (CLT), que desempeña un papel fundamental en la teoría asintótica estadística para aproximar la medida de probabilidad de sumas de variables aleatorias.

Teorema 6.3 T. del Límite Central (TLC) de Lindeberg

Sea $\{X_n\}$ variables aleatorias independientes con $0 < s_n^2 = \text{Var}\left(\sum_{j=1}^n X_j\right) < \infty$ y $E(X_i) < \infty$. Si

$$\textbf{[Condición de Lindebeg]} \sum_{j=1}^{n} E\left[(X_j - EX_j)^2 I_{\{|X_j - EX_j| > \epsilon s_n\}}\right] = o\left(s_n^2\right)$$

para cualquier $\epsilon > 0$, entonces

$$\frac{1}{s_n} \sum_{j=1}^{n} (X_j - EX_j) \to_d N(0,1).$$