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

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

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

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

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

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.