Conceptos preliminares de inferencia estadística paramétrica

18 de marzo de 2025

1. Introducción

Ahora enfocaremos nuestra atención en los modelos paramétricos, es decir, aquellos que suponen que la distribución que describe el fenómeno aleatorio de interés pertenece a una familia específica de distribuciones caracterizadas por un conjunto finito de parámetros. En términos generales, un modelo paramétrico se expresa como $M = \{f(x \mid \theta) : \theta \in \Theta\}$, donde $\Theta \subset \mathbb{R}^k$ es el espacio paramétrico y $\theta = (\theta_1, \ldots, \theta_k)$ representa el conjunto de parámetros desconocidos.

En este contexto, asumimos que los datos $X_1, X_2, \ldots, X_n$ son una muestra aleatoria de la variable $X \sim f(x \mid \theta)$. Así, el problema de la inferencia estadística consiste en estimar el parámetro θ a partir de la información contenida en los datos observados.

1.1. Métodos paramétricos Vs no paramétricos

Una pregunta natural que surge al introducir modelos paramétricos es:

> *¿Cómo podemos estar seguros de que la distribución que generó los datos pertenece a una familia paramétrica específica?*

Esta es una inquietud válida y fundamental. En la práctica, rara vez tenemos certeza de que los datos siguen exactamente un modelo paramétrico, lo que justifica el desarrollo de métodos no paramétricos que permiten mayor flexibilidad.

A pesar de esta limitación, el estudio de la inferencia paramétrica es relevante por varias razones.

1. En algunos casos, el conocimiento previo sugiere que un modelo paramétrico proporciona una aproximación razonable a la realidad. Por ejemplo, el número de accidentes de tráfico en un periodo de tiempo fijo suele modelarse mediante una distribución de Poisson, siempre que los eventos ocurran de forma independiente y a una tasa constante. De manera similar, muchas características biológicas, como la altura y el peso en una población homogénea, pueden aproximarse por una distribución normal, lo que se justifica por el Teorema del Límite Central y observaciones empíricas. Asimismo, los errores de medición en experimentos físicos suelen modelarse con una distribución normal debido a la acumulación de múltiples fuentes de variabilidad pequeñas e independientes. Aunque estas aproximaciones no siempre son exactas, en la práctica pueden facilitar el análisis y la inferencia estadística con resultados robustos.

2. Los métodos paramétricos suelen ser preferibles en situaciones donde se dispone de pocos datos, ya que al imponer una estructura específica sobre la distribución pueden producir estimaciones más estables y eficientes con menos observaciones. También pueden ser útiles cuando se cuenta con una gran cantidad de datos y el modelo paramétrico es una buena aproximación, ya que en tales casos los estimadores paramétricos tienden a ser consistentes y computacionalmente más eficientes que sus contrapartes no paramétricas.

Por otro lado, los métodos no paramétricos son preferibles cuando no se tiene certeza de la forma de la distribución subyacente y se desea evitar imponer una estructura restrictiva. Si se dispone de suficientes datos, estos métodos permiten estimar la distribución con gran flexibilidad sin asumir una familia paramétrica específica.

En resumen, aunque los modelos paramétricos requieren una suposición estructural sobre los datos, su estudio es fundamental no solo por su aplicabilidad en casos específicos, sino también porque proporcionan herramientas conceptuales clave para comprender métodos más generales, incluidos algunos enfoques no paramétricos.

2. Función de Verosimilitud

Una vez establecido el marco de los modelos paramétricos, surge la siguiente pregunta fundamental:

> *Dado que los datos $X_1, \ldots, X_n$ provienen de una distribución $f(x \mid \theta)$, ¿cómo podemos utilizarlos para aprender sobre el valor del parámetro θ?*

La inferencia estadística se basa en cuantificar qué valores de θ son más compatibles con los datos observados. Para ello, introducimos la *función de verosimilitud*, una herramienta fundamental que nos permite evaluar qué tan plausible es un valor del parámetro dado el conjunto de datos.

2.1. Definición

Dado un conjunto de datos observados $X_1, \ldots, X_n$, la **función de verosimilitud** está definida como:

$$L(\theta \mid X_1, \ldots, X_n) = f(X_1, \ldots, X_n \mid \theta).$$

En el caso de que las observaciones sean independientes, la verosimilitud se expresa como:

$$\begin{aligned} L(\theta \mid X_1, \ldots, X_n) &= f(X_1, \ldots, X_n \mid \theta), \\ &= \prod_{i=1}^{n} f(X_i \mid \theta). \end{aligned}$$

Intuitivamente, la función de verosimilitud mide qué tan bien cada posible valor de θ explica los datos observados. En otras palabras, asigna "puntaje" a cada θ basado en la compatibilidad con los datos.

2.2. El ejemplo más sencillo

Consideremos un modelo en el que $X_1, \ldots, X_n$ es una m.a. de $X \sim \text{Bernoulli}(\theta)$, donde cada observación toma valores 0 o 1 con probabilidad θ. La función de verosimilitud es:

$$\begin{aligned} L(\theta \mid X_1, \ldots, X_n) &= \prod_{i=1}^{n} f(X_i \mid \theta), \\ &= \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i}, \\ &= \theta^{S_n}(1-\theta)^{n-S_n}, \end{aligned}$$

donde $S_n = \sum_{i=1}^{n} X_i$ representa el número total de éxitos en la muestra.

Supongamos ahora que $n = 10$ y $S_n = 2$. En este caso, la verosimilitud se expresa como:

$$L(\theta \mid X_1, \ldots, X_n) = \theta^2(1-\theta)^8, \quad \theta \in (0,1).$$

Esta función mide la plausibilidad de cada valor de θ dados los datos observados. Como se observa en la gráfica, la verosimilitud alcanza su máximo en $\theta = 0.2$, lo que sugiere que este es el valor más compatible con los datos.
Este principio es la base del *método de máxima verosimilitud* (MLE), una técnica clave en inferencia estadística.

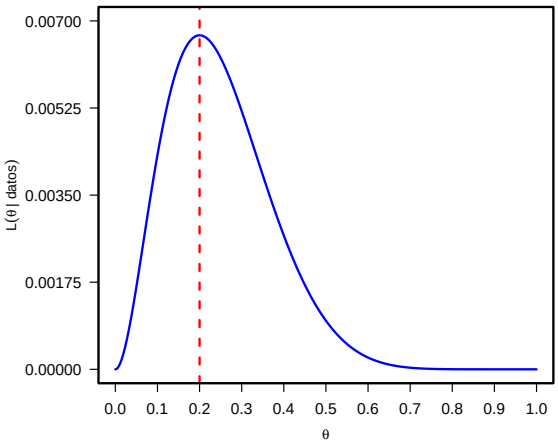

Figura 1: Función de verosimilitud para el modelo Bernoulli con $n = 10$ y $S_n = 2$.

2.3. Más sobre la función de verosimilitud

Hasta ahora, hemos definido la función de verosimilitud y presentado un ejemplo sencillo para ilustrar su comportamiento. Sin embargo, es útil resaltar algunos aspectos clave que ayudan a comprender mejor su papel en inferencia estadística.

2.3.1. Interpretación

Es importante notar que la función de verosimilitud $L(\theta \mid X_1, \ldots, X_n)$ no es una probabilidad en sí misma, sino una medida relativa de plausibilidad para distintos valores de θ. A diferencia de una función de densidad de probabilidad, la verosimilitud no está normalizada, es decir, no necesariamente suma (o integra) a 1 sobre todos los valores posibles de θ.

Lo que realmente nos interesa al trabajar con la verosimilitud no es su magnitud absoluta, sino cómo varía con respecto a θ. En particular, buscamos el valor o valores de θ que la maximizan, lo

que motiva el método de *máxima verosimilitud*.

2.3.2. Log-verosimilitud

En muchos casos, en lugar de trabajar directamente con la función de verosimilitud, se considera su logaritmo:

$$\ell(\theta \mid X_1, \ldots, X_n) = \log\left(L(\theta \mid X_1, \ldots, X_n)\right).$$

Esta transformación es útil por varias razones:

- Convierte productos en sumas, lo que facilita los cálculos algebraicos y la diferenciación.
- Suele evitar problemas numéricos, ya que los productos de muchas probabilidades pueden volverse extremadamente pequeños y difíciles de manejar computacionalmente.
- La ubicación del máximo de la verosimilitud y de la log-verosimilitud es la misma, ya que el logaritmo es una función monótona creciente.

2.3.3. Otro ejemplo: Verosimilitud en un modelo normal

Para reforzar la intuición sobre la función de verosimilitud, consideremos el caso de una m.a. $X_1, \ldots, X_n$ de $X \sim N(\mu, \sigma^2)$, con σ^2 conocida. En este caso, la función de verosimilitud para μ está dada por:

$$\begin{aligned} L(\mu \mid X_1, \ldots, X_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right), \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}\right). \end{aligned}$$

Tomando el logaritmo:

$$\ell(\mu \mid X_1, \ldots, X_n) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2.$$

Maximizando esta expresión con respecto a μ, obtenemos que el estimador de máxima verosimilitud (MLE) es simplemente la media muestral:

$$\hat{\mu}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Este resultado concuerda con nuestra intuición: la media muestral es el mejor estimador de la media poblacional bajo el supuesto de normalidad. Este ejemplo también ilustra cómo la función de verosimilitud cambia dependiendo del modelo paramétrico que estamos considerando.

3. Estadísticas Suficientes

Un problema común en inferencia estadística es determinar si es necesario utilizar toda la muestra para hacer estimaciones sobre θ o si existe una manera más eficiente de resumir la información sin perder precisión. Idealmente, queremos encontrar una función de los datos que contenga toda la información relevante sobre θ. A esta función la llamamos **estadística suficiente**. Su existencia nos permite reducir la cantidad de datos a considerar sin sacrificar la calidad de la inferencia.

3.1. Definición de estadística suficiente

Sea $X_1, \ldots, X_n$ una m.a. de $X \sim f(x \mid \theta)$, una estadística $T(X_1, \ldots, X_n)$ es **suficiente** para θ si la distribución condicional de los datos, dado $T(X_1, \ldots, X_n)$, no depende de θ:

$$P(X_1, \ldots, X_n \mid T(X_1, \ldots, X_n), \theta) = P(X_1, \ldots, X_n \mid T(X_1, \ldots, X_n)), \quad \text{i.e. no depende de } \theta.$$

Una vez observada $T(X_1, \ldots, X_n)$, los datos originales no aportan información adicional sobre θ. Esto implica que cualquier inferencia sobre θ se puede realizar únicamente con $T(X_1, \ldots, X_n)$, sin necesidad de la muestra completa.

3.1.1. Ejemplo

Sea X_1, X_2, X_3 una m.a. de $X \sim \text{Bernoulli}(\theta)$, vamos a considerar $T_1(X_1, X_2, X_3) = X_1 + X_2 + X_3$ y $T_2(X_1, X_2, X_3) = X_1X_2 + X_3$ y aplicar la definición directamente para identificar si alguna es una estadística suficiente.

En en la Tabla 1, se presentan los valores posibles que puede tomar la m.a., así como los valores que tomaría cada estadística.

No.	X_1	X_2	X_3	T_1	T_2
1	0	0	0	0	0
2	1	0	0	1	0
3	0	1	0	1	0
4	0	0	1	1	1
5	1	1	0	2	1
6	1	0	1	2	1
7	0	1	1	2	1
8	1	1	1	3	2

Cuadro 1: Valores posibles de la muestra y las estadísticas T_1 y T_2.

Para verificar si T_1 y T_2 son estadísticas suficientes, es necesario calcular las probabilidades condicionales $P(X_1, X_2, X_3 \mid T_1, \theta)$ y $P(X_1, X_2, X_3 \mid T_2, \theta)$ para cada uno de los valores en la Tabla 1. A continuación, se presenta un ejemplo ilustrativo de cómo realizar este cálculo.

$$P(X_1 = 1, X_2 = 0, X_3 = 1 \mid T_1 = 2, \theta) = \frac{P(X_1 = 1, X_2 = 0, X_3 = 1, T_1 = 2 \mid \theta)}{P(T_1 = 2 \mid \theta)}, \tag{1}$$

$$= \frac{P(X_1 = 1, X_2 = 0, X_3 = 1 \mid \theta)}{P(T_1 = 2 \mid \theta)}, \tag{2}$$

$$= \frac{\theta^2(1-\theta)}{\binom{3}{2}\theta^2(1-\theta)}, \tag{3}$$

$$= \frac{1}{\binom{3}{2}}, \tag{4}$$

Para justificar la transición de la ecuación (1) a (2), recordemos un resultado básico de probabilidad: si $A \subset B$, entonces $P(A \cap B) = P(A)$. En este caso, el conjunto $\{T_1 = 2\}$ contiene al conjunto $\{X_1 = 1, X_2 = 0, X_3 = 1\}$, lo cual puede verificarse en la Tabla 1. Esto nos permite reescribir la probabilidad conjunta en el numerador como la probabilidad de la intersección.

Lo crucial de este desarrollo es notar que la probabilidad condicional $P(X_1 = 1, X_2 = 0, X_3 = 1 \mid T_1 = 2, \theta)$ no depende de θ. Este mismo resultado se obtiene para las 7 combinaciones restantes de la Tabla 1, lo que confirma que T_1 es una estadística suficiente para θ.

De manera análoga, calculamos la probabilidad condicional cuando $T_2 = 1$:

$$P(X_1 = 1, X_2 = 0, X_3 = 1 \mid T_2 = 1, \theta) = \frac{P(X_1 = 1, X_2 = 0, X_3 = 1, T_2 = 1 \mid \theta)}{P(T_2 = 1 \mid \theta)}, \tag{5}$$

$$= \frac{P(X_1 = 1, X_2 = 0, X_3 = 1 \mid \theta)}{P(T_2 = 1 \mid \theta)}, \tag{6}$$

$$= \frac{\theta^2(1-\theta)}{P(T_2 = 1 \mid \theta)}. \tag{7}$$

En este caso, el denominador no se puede expresar de forma simple como en el caso de T_1, ya que T_2 no es una función suma de las observaciones, sino una combinación no lineal de ellas. Para calcular $P(T_2 = 1 \mid \theta)$, observamos en la Tabla 1 que $T_2 = 1$ se alcanza en cuatro configuraciones distintas:

$$P(T_2 = 1 \mid \theta) = P(0, 0, 1) + P(0, 1, 1) + P(1, 0, 1) + P(1, 1, 0).$$

Dado que los X_i son independientes y siguen una Bernoulli(θ), esto se expresa como:

$$\begin{aligned} P(T_2 = 1 \mid \theta) &= (1-\theta)(1-\theta)\theta + (1-\theta)\theta\theta + \theta(1-\theta)\theta + \theta\theta(1-\theta), \\ &= (1-\theta)^2\theta + 3\theta^2(1-\theta) = \theta(1-\theta)(1+2\theta). \end{aligned}$$

Sustituyendo en la ecuación (7):

$$\begin{aligned} P(X_1 = 1, X_2 = 0, X_3 = 1 \mid T_2 = 1, \theta) &= \frac{\theta^2(1-\theta)}{\theta(1-\theta)(1+2\theta)}, \\ &= \frac{\theta}{1+2\theta}. \end{aligned}$$

En este caso, la probabilidad condicional sí depende de θ, lo que implica que T_2 no es una estadística suficiente para θ. Esto significa que T_2 no captura toda la información relevante sobre el parámetro, y se requeriría información adicional para hacer inferencia sobre θ.

3.2. Teorema de Factorización de Fisher-Neyman

Verificar directamente la definición de suficiencia puede ser complicado. Sin embargo, a partir de ella se puede demostrar el siguiente resultado, que proporciona un criterio práctico para determinar si una estadística es suficiente.

Teorema 1. *Sea $X_1, X_2, \ldots, X_n$ una m.a. de $X \sim f(x \mid \theta)$. Una estadística $T(X_1, \ldots, X_n)$ es suficiente para θ si y solo si la función de verosimilitud se puede factorizar como:*

$$L(\theta \mid X_1, \ldots, X_n) = g(T(X_1, \ldots, X_n), \theta) h(X_1, \ldots, X_n),$$

donde g es una función de la estadística T y del parámetro θ, mientras que h es una función que depende únicamente de la muestra.

El teorema de factorización nos permite: 1) Identificar una estadística suficiente sin necesidad de proponer un candidato a priori. 2) Evitar el cálculo explícito de la distribución condicional de los datos.

En esencia, si la parte de la verosimilitud que depende de θ solo involucra $T(X_1, \ldots, X_n)$, entonces $T(X_1, \ldots, X_n)$ es suficiente para θ.

3.2.1. Ejemplo: Modelo Bernoulli

Sea $X_1, \ldots, X_n$ una m.a. de $X \sim \text{Bernoulli}(\theta)$. Como vimos anteriormente, la función de verosimilitud está dada por:

$$L(\theta \mid X_1, \ldots, X_n) = \theta^{S_n}(1-\theta)^{n-S_n},$$

donde $S_n = \sum_{i=1}^n X_i$ representa el número total de éxitos en la muestra.

Comparando con la factorización del teorema, identificamos:

$$g(T(X_1, \ldots, X_n), \theta) = \theta^{S_n}(1-\theta)^{n-S_n}, \quad h(X_1, \ldots, X_n) = 1.$$

Por lo tanto, por el Teorema de Factorización de Fisher-Neyman, la estadística $T(X_1, \ldots, X_n) = S_n$ es suficiente para θ.

En términos prácticos, esto significa que, en el modelo Bernoulli, para hacer inferencia sobre θ, no es necesario conocer los n valores de la muestra individualmente, sino solo el número total de éxitos S_n.

3.2.2. Ejemplo: Modelo Normal

Consideremos ahora el caso donde $X_1, \ldots, X_n$ es una m.a. de $X \sim N(\mu, \sigma^2)$. La función de verosimilitud está dada por:

$$\begin{aligned} L(\mu \mid X_1, \ldots, X_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(X_i - \mu)^2\right), \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right). \end{aligned}$$

Utilizando la identidad:

$$\begin{aligned} \sum_{i=1}^{n}(X_i - \mu)^2 &= \sum_{i=1}^{n}(X_i \underbrace{-\bar{X}_n + \bar{X}_n}_{=0} - \mu)^2, \\ &= \sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + 2\cancel{\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)\left(\bar{X}_n - \mu\right)} + n(\bar{X}_n - \mu)^2, \\ &= \sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + n(\bar{X}_n - \mu)^2. \end{aligned}$$

Podemos reescribir la verosimilitud como:

$$L(\mu \mid X_1, \ldots, X_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right) \exp\left(-\frac{n}{2\sigma^2}(\bar{X} - \mu)^2\right). \qquad (8)$$

Caso 1: Varianza conocida, media desconocida. Si suponemos que σ^2 es conocida, observamos que la factorización nos permite definir:

$$\begin{aligned} g(T(X_1, \ldots, X_n), \mu) &= \exp\left(-\frac{n}{2\sigma^2}(\bar{X}_n - \mu)^2\right), \\ h(X_1, \ldots, X_n) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right). \end{aligned}$$

Por el Teorema de Factorización de Fisher-Neyman, $\bar{X}_n$ es suficiente para μ. Es decir, si conocemos $\bar{X}_n$, el resto de los datos no aporta información adicional sobre μ.

Caso 2: Media conocida, varianza desconocida. Si en cambio suponemos que μ es conocida, de la ecuación (8) podemos escribir:

$$\begin{aligned} g(T(X_1, \ldots, X_n), \sigma^2) &= \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right) \exp\left(-\frac{n}{2\sigma^2}(\bar{X}_n - \mu)^2\right), \\ h(X_1, \ldots, X_n) &= \left(\frac{1}{2\pi}\right)^{n/2}. \end{aligned}$$

Por el Teorema de Factorización de Fisher-Neyman, $\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$ es una estadística suficiente para σ^2 (ya que esta cantidad contiene toda la información sobre σ^2).

También es sencillo verificar que $\sum_{i=1}^{n} X_i^2$ y $\bar{X}_n$ son suficientes para σ^2.

Caso 3: Ambos parámetros desconocidos. Finalmente, si tanto μ como σ^2 son desconocidos, la factorización en (8) nos indica que las estadísticas $\bar{X}_n$ y $\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ son suficientes para μ y σ^2.

3.3. Comentarios adicionales sobre estadísticas suficientes

3.3.1. Toda la muestra es una estadística suficiente

Desde la definición de estadística suficiente, es claro que la muestra completa $X_1, \ldots, X_n$ es una estadística suficiente para cualquier parámetro θ. Esto se debe a que al condicionar sobre todos los datos observados, la distribución condicional trivialmente no depende de θ.

Sin embargo, el concepto de suficiencia es más útil cuando nos permite encontrar una forma más compacta de resumir la información sobre θ. En la práctica, buscamos estadísticas suficientes que reduzcan la dimensión de los datos sin perder información inferencial relevante.

3.3.2. Funciones uno a uno de una estadística suficiente también son suficientes

Si $T(X_1, \ldots, X_n)$ es una estadística suficiente para θ y $S(X_1, \ldots, X_n) = k(T(X_1, \ldots, X_n))$ es una transformación invertible (uno a uno), entonces $S(X_1, \ldots, X_n)$ también es una estadística suficiente para θ.

Este resultado se deduce fácilmente del Teorema de Factorización de Fisher-Neyman: si la verosimilitud se factoriza en términos de $T(X_1, \ldots, X_n)$, entonces también se puede expresar en términos de cualquier transformación invertible $S(X_1, \ldots, X_n)$.

Por ejemplo, en el modelo Bernoulli, si $S_n = \sum_{i=1}^{n} X_i$ es suficiente para θ, cualquier transformación uno a uno de S_n, como $\bar{X}_n$, S_n^2 o e^{S_n}, también será una estadística suficiente para θ.

3.3.3. Estadística suficiente minimal

Dado que existen muchas estadísticas suficientes (incluyendo la muestra completa), en la práctica nos interesa identificar una que sea minimal, es decir, que no retenga más información de la necesaria.

Formalmente, una estadística suficiente $T(X_1, \ldots, X_n)$ se dice minimal si es una función de cualquier otra estadística suficiente. Es decir, si $S(X_1, \ldots, X_n)$ es otra estadística suficiente, entonces debe existir una función k tal que:

$$T(X_1, \ldots, X_n) = k(S(X_1, \ldots, X_n)).$$

Esto implica que no existe una estadística suficiente que reduzca más la información que $T(X_1, \ldots, X_n)$ sin perder información sobre θ.

3.3.4. Resumen sin pérdida de información y eficiencia en inferencia paramétrica

El concepto de suficiencia permite reducir la cantidad de datos a procesar sin perder información sobre θ. Esto tiene importantes implicaciones en la inferencia paramétrica, especialmente en conjuntos de datos grandes. Cuando se dispone de una estadística suficiente, se puede realizar inferencia de manera más eficiente al reducir la cantidad de información necesaria sin sacrificar precisión.

En contraste, los métodos no paramétricos, al no asumir una estructura específica en los datos, generalmente requieren analizar la muestra completa, lo que puede ser menos eficiente en términos computacionales y de almacenamiento. Así, el uso de estadísticas suficientes es una de las razones por las cuales la inferencia paramétrica puede ser más eficiente en conjuntos de datos grandes.

4. Familia Exponencial

4.1. Definición

En inferencia estadística, trabajamos con modelos paramétricos $f(x \mid \theta)$, donde θ es un parámetro desconocido. En muchos casos, estos modelos pueden expresarse en la forma:

$$f(x \mid \theta) = h(x) \exp\left(\eta(\theta) T(x) - A(\theta)\right), \tag{9}$$

donde:

- $\eta(\theta)$ es el **parámetro natural o canónico**.
- $T(x)$ es la **estadística suficiente canónica**.
- $A(\theta)$ es una **función de normalización**, que garantiza que $f(x \mid \theta)$ define una distribución válida.
- $h(x)$ es una **función base**, que depende únicamente de los datos y no del parámetro.

Si la función de densidad (o de masa de probabilidad) de una variable aleatoria puede escribirse en la forma (9), entonces decimos que la distribución pertenece a la **familia exponencial**. Esta estructura es fundamental en inferencia estadística, ya que facilita la identificación de estadísticas suficientes y simplifica el análisis de la verosimilitud.

4.2. Ejemplo: Modelo Bernoulli

Consideremos una muestra aleatoria $X_1, \ldots, X_n$ de una variable $X \sim \text{Bernoulli}(\theta)$, cuya función de masa de probabilidad es:

$$\begin{aligned} f(x \mid \theta) &= \theta^x (1-\theta)^{1-x}, \\ &= \left(\frac{\theta}{1-\theta}\right)^x (1-\theta), \\ &= \exp\left(x \log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta)\right). \end{aligned}$$

Comparando con la forma general de la familia exponencial en (9), identificamos los términos:

- $\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$.
- $T(x) = x$.
- $A(\theta) = \log(1-\theta)$.
- $h(x) = 1$.

Por lo tanto, la distribución Bernoulli pertenece a la familia exponencial.

4.3. Motivación: ¿Por qué nos interesa la familia exponencial?

Hasta ahora, hemos estudiado la función de verosimilitud y el concepto de estadística suficiente, dos elementos clave en inferencia paramétrica. Sin embargo, la función de verosimilitud puede volverse compleja al analizar distintos modelos probabilísticos. La familia exponencial permite estructurar la verosimilitud de manera general y unificada, lo que facilita su estudio y aplicación.

La relevancia de la familia exponencial radica en que muchas distribuciones utilizadas en estadística, como la Bernoulli, la binomial, la normal y la Poisson, pueden expresarse en esta forma. Esto la convierte en un marco natural para el análisis de modelos paramétricos. Sus propiedades permiten:

- Expresar la función de verosimilitud en una forma más manejable.
- Identificar estadísticas suficientes de manera sencilla.
- Facilitar la inferencia estadística aprovechando sus propiedades generales.

4.3.1. Conexión con la función de verosimilitud

Sea $X_1, \ldots, X_n$ una m.a. de $X \sim f(x \mid \theta)$. Si la distribución pertenece a la familia exponencial, entonces su función de verosimilitud se puede escribir como:

$$\begin{aligned} L(\theta \mid X_1, \ldots, X_n) &= \prod_{i=1}^{n} h(X_i) \exp\left(\eta(\theta) T(X_i) - A(\theta)\right), \\ &= \left(\prod_{i=1}^{n} h(X_i)\right) \exp\left(\eta(\theta) \sum_{i=1}^{n} T(X_i) - nA(\theta)\right). \end{aligned} \tag{10}$$

Esta estructura no solo permite organizar la función de verosimilitud de manera más clara, sino que también facilita su manipulación y análisis. En particular, permite identificar estadísticas suficientes de manera automática. De hecho:

- Si una distribución pertenece a la familia exponencial, la estadística suficiente está dada por $\sum_{i=1}^{n} T(X_i)$.
- Esto se deriva directamente del Teorema de Factorización de Fisher-Neyman.

De esta manera, al expresar la función de densidad o masa de probabilidad en la forma de la familia exponencial, la estadística suficiente se obtiene de forma inmediata. Por ejemplo, en el caso de la distribución Bernoulli, la estadística suficiente es $S_n = \sum_{i=1}^{n} X_i$, lo mismo que ya habíamos obtenido.

5. Propiedades inferenciales de la familia exponencial

Uno de los beneficios clave de la familia exponencial es que su estructura facilita la estimación por máxima verosimilitud (MLE). En particular, la función de verosimilitud adopta una forma que permite obtener estimadores de manera más sencilla y con buenas propiedades asintóticas. Algunas de las razones por las que esto ocurre son:

- La función de verosimilitud puede expresarse en términos de la estadística suficiente $T(X_1, \ldots, X_n)$, lo que reduce la dimensión del problema de estimación sin pérdida de información relevante.
- En muchos casos, la función log-verosimilitud es cóncava en el parámetro, lo que garantiza que cualquier punto crítico encontrado al derivar e igualar a cero sea automáticamente un máximo global, sin necesidad de verificar la segunda derivada.
- Los estimadores de máxima verosimilitud en la familia exponencial cumplen, en general, las condiciones necesarias para garantizar su consistencia y normalidad asintótica, conceptos que serán abordados más adelante en el curso. Esto se debe a que la estructura regular de la familia exponencial evita problemas como derivadas no definidas o funciones de verosimilitud multimodales.

5.1. Conclusión

La familia exponencial no solo simplifica la formulación de la verosimilitud y la identificación de estadísticas suficientes, sino que también facilita la inferencia estadística. Sus propiedades permiten obtener estimadores de máxima verosimilitud de manera sencilla, garantizan que dichos estimadores sean consistentes y asintóticamente normales, y además ofrecen ventajas en el contexto Bayesiano mediante la existencia de priors conjugadas (lo veremos más adelante). Estas características justifican su importancia dentro de la inferencia paramétrica.