Inferencia Estadística

Apuntes

Andrés Herrera Poyatos
Nuria Rodríguez Barroso
Juan Luis Suárez Díaz
Universidad de Granada

Índice

1. Introducción

En esta sección introductora se motivan los problemas de la inferencia estadística. Para ello recordamos aquellos algunos conceptos de la teoría de probabilidad que pueden no haberse estudiado en un curso básico de probabilidad.

1.1. Variables aleatorias. Vectores aleatorios

Fijemos un espacio de probabilidad $(\Omega, \mathcal{A}, P)$. En relación con este espacio de probabilidad se puede realizar un experimento aleatorio, cuyos resultados se codifican como números reales para facilitar su tratamiento. Esta codificación recibe el nombren de variable aleatoria, concepto que es estudiado en cualquier curso de probabilidad.

Definición 1.1. Una variable aleatoria es una función medible $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$.

Por ejemplo, imaginemos que Ω es el conjunto de todas las personas, $\mathcal{A}$ es el conjunto de las partes de Ω y P es uniforme, esto es, todas las personas tienen la misma probabilidad de ser escogidas. El experimento aleatorio consiste en seleccionar una persona aletoria y medir su altura. Formalmente se puede codificar como la variable aleatoria X que asigna acada persona su altura. Puede ser interesante calcular la probabilidad de que al tomar una persona su altura sea 2 metros, esto es $P(X = 2)$. En definitiva, las variables aleatorias nos permiten modelar los resultados del experimento matemáticamente. Nótese que podemos calcular la probabilidad $P(X = 2)$ gracias a que el conjunto $\{w \in \Omega : X(w) = 2\} = X^{-1}(2)$ es medible, de ahí que se exija que X lo sea en la definición.

Habitualmente para representar un experimento se necesitan múltiples valores reales. En este caso se utilizan los denominados *vectores aleatorios* o *variables aleatorias multidimensionales*, que son funciones medibles $\underset{\sim}{X} = (X_1, \ldots, X_n) : (\Omega, \mathcal{A}) \to (\mathbb{R}^n, \mathcal{B}^n)$.

A un vector aleatorio se le asocia una distribución de probabilidad en $\mathbb{R}^n$, que puede venir dada por una función de densidad o por una función de distribución. Nos referimos a un libro de texto clásico para recordar estos conceptos [1].

1.2. Distribución conjunta

Como se ha mencionado, un vector aleatorio $\underset{\sim}{X} = (X_1, \ldots, X_n)$ tiene asociada una distribución de probabilidad. Si en esta distribución interviene más de una variable ($n \geq 2$), entoces diremos que es una distribución conjunta. Cada una de las componentes del vector $\underset{\sim}{X}$ es una variable aleatoria y, por tanto, tiene asociada una distribución. Es más, la misma observación es válida para cualquier subtupla de $\underset{\sim}{X}$. La distribución de una subtupla se denomina distribución marginal. Cabe preguntarse cómo calcular una distribución marginal a partir de la distribución conjunta. Es fácil razonar que si $\underset{\sim}{X} = (Y_1, Y_2)$, donde Y_1 e Y_2 son vectores aleatorios, entonces la distribución marginal de Y_1 tiene función de densidad

$$f(y_1) = \int f(y_1, y_2)\, dy_2,$$

donde $f(y_1, y_2)$ es la función de densidad de $\underset{\sim}{X}$. Para los detalles nos referimos de nuevo a un libro de texto básico de teoría de probabilidad [1].

1.3. Muestra aleatoria simple

Sea X una variable aleatoria que se desea observar. Una muestra aleatoria simple de X es un vector aleatorio $\underset{\sim}{X} = (X_1, \ldots, X_n)$ donde las variables X_i son independientes y tienen la misma distribución que la variable X. Esta definición se corresponde con realizar n veces consecutivas el experimento definido por la variable X. Los n experimentos son independientes y siguen la misma distribución de probabilidad. En la práctica tras realizar los experimentos obtenemos un vector $\underset{\sim}{x} = (x_1, \ldots, x_n)$ con los valores observados. Este vector se denomina realización de la muestra.

Podemos calcular la distribución de la muestra $\underset{\sim}{X}$ gracias a la hipótesis de independencia. En efecto, la función de densidad en un punto $\underset{\sim}{x}$ viene dada por

$$f(\underset{\sim}{x}) = \prod_{i=1}^{n} f(x_i).$$

1.4. Familias de distribuciones paramétricas

Habitualmente observamos variables aleatorias de las que desconocemos su distribución. No obstante, intuimos que la distribución tiene una determinada forma que depende de un número finito de parámetros. Esto es, la densidad de la distribución pertenece a una familia paramétrica $\{f(x|\theta) : \theta \in \Theta\}$, donde $\Theta \subset \mathbb{R}^k$. Denotamos por $\theta_0 \in \Theta$ al verdadero valor del parámetro de la distribución de la variable aleatoria que estamos observando. La inferencia estadística se encarga de inferir propiedades de θ_0 a partir de la realización de una muestra de X.

2. Familias de distribuciones

2.1. Distribuciones discretas

En esta sección se desarrollan varias de las distribuciones discretas más importantes de la estadística.

2.1.1. Distribución uniforme

La distribución uniforme es una distribución de probabilidad que asume un número finito de valores con la misma probabilidad. Es fácil comprobar que la función masa de probabilidad es $f(x|n) = \frac{1}{n}$. Claramente $\sum_{i=1}^{n} \frac{1}{n} = 1$.

La función generatriz de momentos es fácil calcularla y viene definida por $\varphi_X(t) = \frac{e^t(1-e^tN)}{N(1-e^t)}$. De ella podemos obtener su media y varianza las cuales quedan de la siguiente forma:

$$E[X] = \frac{N+1}{2}$$
$$Var(X) = E[X^2] - (E[X])^2 = \frac{(N+1)(N-1)}{12}$$

2.1.2. Distribución de Poisson

Esta distribución expresa, a partir de una frecuencia de ocurrencia media, la probabilidad de que ocurra un determinado número de eventos durante cierto período de tiempo. La función de masa o probabilidad de la distribución de Poisson es $f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$.
Claramente $\sum_{i=1}^{n} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{i=1}^{n} \frac{\lambda^x}{x!} = 1$.

La función generatriz de momentos de dicha distribución se calcula de la siguiente manera $\varphi_X(t) = \sum_{i=0}^{n} \frac{e^{tx}e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{i=0}^{n} \frac{(e^t\lambda)^x}{x!} = e^{-\lambda}e^{e^t\lambda} = e^{\lambda(e^t-1)}$.

A partir de la función generatriz de momentos podemos fácilmente deducir la media y la varianza:

$$E[X] = \lambda$$
$$Var(X) = E[X^2] - (E[X])^2 = \lambda$$

2.1.3. Distribución binomial

Considérese un experimento de Bernoulli con probabilidad $\theta \in [0, 1]$. Repetimos el experimento n veces y nos preguntamos cuál es la probabilidad de que se hayan conseguido x aciertos, donde $x = 0, 1, \ldots, n$. Es fácil ver que esta probabilidad viene dada por $\binom{n}{x}\theta^x(1-\theta)^{n-x}$. Esta cuestión, que es habitual en la estadística, origina la distribución binomial.

Definición 2.1. Una variable aleatoria sigue una distribución binomial con parámetros $n \in \mathbb{N}$ y $\theta \in [0, 1]$ si su función masa de probabilidad viene dada por $f(x|n, \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$. En tal caso se denota $X \sim B(x|n, \theta)$.

2.1.4. Distribución multinomial

Definición 2.2. La distribución multinomial se puede ver como una generalización de la distribución binomial para variables politómicas (variables discretas con más de dos categorías).
Supongamos que se realizan n ensayos independientes que dan lugar a k resultados distintos $X = (X_1, ...X_k)$ con probabilidades $\theta_1, ..., \theta_k$ respectivamente, donde $\theta_1 + ... + \theta_k = 1$. Entonces, la función de probabilidad para dicha variable multinomial sería:

$$f(X|\theta_1, ..., \theta_k) = \frac{n!}{X_1!X_2!...X_k!}\theta^{X_1}\theta^{X_2}...\theta^{X_k}$$

2.2. Distribuciones continuas

En esta sección se desarrollan varias de las distribuciones continuas más importantes de la estadística.

2.2.1. Distribución uniforme

La distribución uniforme asigna una credibilidad uniforme a todos los puntos de un intervalo $[a, b]$. Esto es, su función de densidad viene dada por

$$f(x|a, b) = \begin{cases} \frac{1}{b-a} \text{ si } x \in [a, b], \\ 0 \text{ en otro caso.} \end{cases}$$

Claramente tenemos que $\int_{-\infty}^{\infty} f(x|a, b)dx = 1$. Además, podemos calcular fácilmente sus momentos como sigue (y, por tanto, también su varianza)

$$E[X^j] = \int_a^b \frac{x^j}{b-a}dx = \frac{b^{j+1} - a^{j+1}}{(b-a)(j+1)},$$

$$Var(X) = E[X^2] - E[X]^2 = \frac{a^2 + ab + b^2}{3} - \frac{(a+b^2)}{4} = \frac{(b-a)^2}{12}.$$

2.2.2. Distribución normal

La distribución normal, también llamada distribución gaussiana, es la distribución más importante de la estadística. Esto se debe a sus numerosas aplicaciones en análisis de poblaciones y al teorema central del límite.

Definición 2.3. Sean $\mu \in \mathbb{R}$ y $\sigma^2 > 0$. Definimos la distribución $N(x|\mu, \sigma^2)$ como la distribución que tiene función de densidad

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}, x \in \mathbb{R}.$$

La distribución normal está bien definida como consecuencia del siguiente lema.

Lema 2.1. *Sean $\mu \in \mathbb{R}$ y $\sigma > 0$. Tenemos que $\int_{-\infty}^{\infty} e^{-(x-\mu)^2/(2\sigma^2)}dx = \sqrt{2\pi}\sigma$.*

Demostración. En primer lugar, vamos a calcular la integral para $\mu = 0$ y $\sigma = 1$. La demostración consiste en reducir el problema en calcular una integral en dos variables. Para ello, elevamos al cuadrado y obtenemos

$$\left(\int_{-\infty}^{\infty} e^{-x^2/2}dx\right)^2 = \left(\int_{-\infty}^{\infty} e^{-t^2/2}dt\right)\left(\int_{-\infty}^{\infty} e^{-s^2/2}ds\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(t^2+s^2)/2}dtds.$$

Resolvemos esta última integral mediante un cambio a polares

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(t^2+s^2)/2}dtds = \int_{-\pi}^{\pi}\left(\int_0^{\infty}\rho e^{-\rho^2/2}d\rho\right)d\theta = 2\pi\int_0^{\infty}\rho e^{-\rho^2/2}d\rho = 2\pi.$$

Por último, utilizamos el cambio de variable $y = (x-\mu)/\sigma$ para obtener

$$\int_{-\infty}^{\infty} e^{-(x-\mu)^2/(2\sigma^2)}dx = \int_{-\infty}^{\infty}\sigma e^{-y^2/2}dy = \sqrt{2\pi}\sigma. \qquad \square$$

Nótese que si $X \sim N(x|\mu,\sigma^2)$, entonces $Y = (X-\mu)/\sigma$ sigue una distribución $N(x|0,1)$.

Proposición 2.2. *La función característica de la distribución $N(x|\mu,\sigma^2)$ viene dada por $\varphi_X(t) = e^{it\mu - t^2\sigma^2/2}$.*

Demostración. En primer lugar, tenemos que

$$\varphi_X(t) = E[e^{itX}] = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{itx-(x-\mu)^2/(2\sigma^2)}dx = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-((x-\mu)^2-2itx\sigma^2)/(2\sigma^2)}dx.$$

Completamos cuadrados como sigue

$$(x-\mu)^2 - 2itx\sigma^2 = (x-(it\sigma^2+\mu))^2 + t^2\sigma^4 - 2it\sigma^2\mu.$$

Esto sugiere utilizar el cambio de variable $g(y) = y + it\sigma^2$. Obtenemos

$$\begin{aligned}\sqrt{2\pi}\sigma\varphi_X(t) = \int_{-\infty}^{\infty} e^{-((x-\mu)^2-2itx\sigma^2)/(2\sigma^2)} &= e^{it\mu-t^2\sigma^2/2}\int_{-\infty}^{\infty} e^{-((x-(it\sigma^2+\mu))^2)/(2\sigma^2)}dx \\ &= e^{it\mu-t^2\sigma^2/2}\int_{-\infty}^{\infty} e^{-(y-\mu)^2)/(2\sigma^2)}dy = \sqrt{2\pi}\sigma e^{it\mu-t^2\sigma^2/2},\end{aligned}$$

como se quería. Nótese que a pesar de ser una integral de contorno compleja el cambio de variable es válido. En efecto, el cambio de variable es afín y la función a integrar es entera. Por tanto, utilizando el camino cerrado $g([0,\infty]) + [\infty, 0]$ se puede probar que el cambio es válido. $\square$

Análogamente se puede probar el siguiente resultado.

Proposición 2.3. *La función generatriz de momentos de la distribución $N(x|\mu,\sigma^2)$ viene dada por $\varphi_X(t) = e^{t\mu-t^2\sigma^2/2}$.*

Corolario 2.4. *Los momentos de la distribución $N(x|\mu,\sigma^2)$ verifican la ecuación recurrente*

$$E[X^k] = -(k-1)\sigma^2 E[X^{k-2}] + (\mu - t\sigma^2)E[X^{k-1}], \quad k \geq 2.$$

Demostración. Sabemos que $E[X^k] = \varphi_X^{(k)}(t)$. Tenemos $\varphi_X^{(1)}(t) = (\mu - t\sigma^2)\varphi_X(t)$. Consecuentemente,

$$\varphi_X^{(2)}(t) = -\sigma^2\varphi_X(t) + (\mu - t\sigma^2)\varphi_X^{(1)}(t).$$

Por inducción se extiende el resultado fácilmente para $k \geq 2$. $\square$

Corolario 2.5. *Si* $X \sim N(x|\mu,\sigma^2)$*, entonces* $E[X] = \mu$ *y* $E[X^2] = \sigma^2 + \mu^2$*. Consecuentemente,* $Var(X) = \sigma^2$*. Como consecuencia de este resultado al parámetro* μ *se le llama media y al parámetro* σ^2 *varianza.*

Podemos utilizar los dos corolarios anteriores para calcular los momentos de la distribución normal resolviendo una ecuación recurrente de segundo orden. Evidentemente, la fórmula obtenida será bastante larga. Sin embargo, esta ecuación se simplifica en el caso de los momentos centrados, como pone de manifiesto el siguiente resultado, que se puede demostrar fácilmente por inducción a partir del Corolario 2.4.

Corolario 2.6. *Si* $X \sim N(x|0,\sigma^2)$*, entonces*

$$E[X^k] = \begin{cases} 0 & \text{si } k \text{ es impar;} \\ (k-1)!!\sigma^k & \text{si } k \text{ es par;} \end{cases}$$

donde $n!!$ *denota al doble factorial, definido como el producto de los números desde* 1 *hasta* n *con la misma paridad que* n*.*

2.2.3. Distribución T de Student

2.2.4. Distribución de Dirichlet

La distribución de Dirichlet es la generalización en multivariable de la distribución Beta. Comúnmente, se utilizan las funciones de Dirichlet como funciones a priori en estadística Bayesiana. Definimos la distribución de Dirichlet de orden $n \geq 2$ con parámetros $\alpha_1, ..., \alpha_n$ tiene una función de densidad de probabilidad

$$f(x_1, ..x_n|\alpha_1, ..., \alpha_n) = \frac{\Gamma(\alpha_1 + ... + \alpha_n)}{\Gamma(\alpha_1) + ... + \Gamma(\alpha_n)} \prod_{i=1}^{n} x_i^{\alpha_i - 1}$$

definida en el simplex abierto de (n-1) dimensiones definido por:

$$x_1, ..., x_n > 0$$

$$x_1 + ... + x_{n-1} < 1$$

$$x_n = 1 - x_1 - ... - x_{n-1}$$

Teorema 2.7. *Sea* $X = (X_1, ..., X_k) \sim Dirichlet(X|\alpha_1, ..., \alpha_k, \alpha_{k+1})$ *se tiene que para cualquier* $k_1 < k$ *se verifica que* $X' = (X_1, ..., X_{k_1}) \sim D(X'|\alpha_1, ..., \alpha_{k_1}, \alpha^*_{k_1+1})$ *con* $\alpha^*_{k+1} = \sum_{j=1}^{k_1} \alpha_j$

Demostración. Pendiente □

3. Estimación de parámetros

Supongamos que estamos estudiando un fenómeno aleatorio que sabemos que sigue una distribución $f(X|\theta_0)$, donde $\theta_0 \in \Theta$ es un parámetro que no es conocido. Nuestro objetivo es estimar el parámetro θ_0 a partir de una muestra $x_1, \ldots, x_n$. Para ello buscamos una función T_n de manera que podamos decir $\theta_0 \approx T_n(x_1, \ldots, x_n)$.

Definición 3.1. Un estimador puntual es una función medible $T_n(X_1, \ldots, X_n)$ que toma valores en Θ, donde Θ es el dominio del parámetro a estimar. Una estimación es la evaluación obtenida por un estimador sobre una muestra $x_1, \ldots, x_n$, esto es, $T_n(x_1, \ldots, x_n)$.

Nótese que la nomenclatura es ambigua. Para nosotros una muestra es una secuencia finita de variables aleatorias independientes e idénticamente distribuidas. Sin embargo, a los valores $x_1, \ldots, x_n$ obtenidos en la práctica también se le denomina muestra. Algunos autores evitan esta abigüedad denominando a $x_1, \ldots, x_n$ realización de la muestra. Nosotros distinguiremos entre ambos casos mediante el uso de mayúsculas para denotar variables aleatorias y el uso de minúsculas para denotar valores concretos.

En múltiples situaciones encontramos estimadores de calidad de forma natural. Por ejemplo, imaginemos que el parámetro θ_0 se corresponde con la media de la distribución $f(X|\theta_0)$. En tal caso, parece claro que el mejor estimador para θ_0 será la media muestral $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$. Sin embargo, en general no sabemos qué estimador hay que utilizar. Buscamos técnicas que nos proporcionen estimadores que sean razonables. En ocasiones querremos estimar $g(\theta_0)$, donde g es determinada transformación de Θ en otro espacio más manejable.

3.1. Método de los momentos

El método de los momentos es, probablemente, el método más antiguo para estimar parámetros. Fue propuesto por Pearson al finales del siglo XIX. En muchos casos los resultados de este método son mejorables. Sin embargo, siempre es un último recurso en el caso de que no podamos aplicar otros métodos.

Sea $X_1, \ldots, X_n$ una muestra de un fenómeno con función de distribución $f(X|\theta)$ con $\theta = (\theta_1, \ldots, \theta_m) \in \Theta \subset \mathbb{R}^m$. Definimos los momentos de la muestra como $m_j = \frac{1}{n}\sum_{i=1}^n X_i^j$. En media se debería cumplir que $m_j = E_\theta X^j$ para todo j tal que $E_\theta X^j$ existe. Nótese que $E_\theta X^j = \mu_j(\theta_1, \ldots, \theta_m)$ es una función que depende de $\theta_1, \theta_2, \ldots, \theta_k$. El método de los momentos propone como estimador a una solución del sistema de ecuaciones

$$\begin{aligned} m_1 &= \mu_1(\theta_1, \ldots, \theta_k), \\ m_2 &= \mu_2(\theta_1, \ldots, \theta_k), \\ &\vdots \\ m_k &= \mu_k(\theta_1, \ldots, \theta_k). \end{aligned} \tag{1}$$

Ejemplo 3.1 (Distribución normal): Supongamos que $X_1, \ldots, X_n$ son muestras de una distribución normal $N(\theta, \sigma^2)$. En el contexto anterior, los parámetros a estimar son $\theta_1 = \theta, \theta_2 = \sigma^2$. En este caso el sistema (1) viene dado por las ecuaciones $\overline{X} = \theta$ y $m_2 = \theta^2 + \sigma^2$. La solución claramente es $\theta = \overline{X}$ y

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2.$$

En este caso, los estimadores obtenidos coinciden con nuestra intuición. Este método es más útil cuando no disponemos de un estimador intuitivo. △

3.2. Método de la máxima verosimilitud de Fisher

El método de la máxima verosimilitud es una de las técnicas más utilizada para obtener estimadores de calidad.

Definición 3.2. Sea $x_1, \dots, x_n$ una muestra de un fenómeno con función de distribución $f(X|\theta_0)$, donde $\theta_0 \in \Theta$. Se define la función de verosimilitud para cada $\theta \in \Theta$ como $L(\theta; x) = \prod_{i=1}^n f(x_i|\theta)$.

Para cada posible valor θ del parámetro a estimar, la verosimilitud proporciona la credibilidad que se le da a θ para los datos $x_1, \dots, x_n$. Buscamos una aproximación $\hat{\theta}$ de θ_0 en base a la muestra obtenida. Parece lógico que si asumimos que los datos son correctos, entonces una buena aproximación será aquella en la que los datos sean coherentes, esto es, la probabilidad de que se den datos similares a la muestra observada debe ser lo más alta posible.

Definición 3.3. Para cada elemento $x = (x_1, \dots, x_n)$ del espacio muestral, definimos $\hat{\theta}(x) \in \Theta$ como un máximo global de $L(\theta; x)$. El estimador máximo verosímil (EMV) de una muestra X se define como $\hat{\theta}(X)$.

El estimador máximo verosímil presenta principalmente dos problemas.

- Cálculo del estimador. Para calcular $\hat{\theta}(X)$ es necesario maximizar una función. Muchas veces esto es complejo incluso para funciones de densidad comunes. Es más, puede suceder que la verosimilitud presente múltiples máximos globales y, por tanto, el estimador máximo verosímil no está bien definido. Necesitaremos condiciones sobre la distribución que nos permitan asegurar la buena definición del estimador máximo verosímil.
- Sensibilidad numérica. El valor $\hat{\theta}(x)$ puede cambiar considerablemente para pequeñas variaciones de x. Nos preguntamos qué condiciones debe verificar la función de distribución para evitar este comportamiento.

Para adentrarnos en el estudio de estos problemas necesitaremos teoría general de estimadores. Antes de desarrollarla realizaremos varios ejemplos de cálculo de estimadores máximo verosímiles.

Comentario 3.2. Los máximos globales de la función $L(\theta; x)$ se corresponden con los máximos globales de la función $\log L(\theta; x) = \sum_{i=1}^n \log f(x_i|\theta)$. En múltiples ocasiones es más sencillo maximizar esta última expresión.

Ejemplo 3.3 (Distribución normal): Consideremos una muestra $X_1, \dots, X_n$ de un fenómeno con distribución $N(\theta, 1)$. En primer lugar, calculamos la función de verosimilitud

$$L(\theta; x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\sum_{i=1}^n (x_i-\theta)^2/2}.$$

En virtud del Comentario 3.2 maximizamos la función $-\sum_{i=1}^n (x_i-\theta)^2/2 - n/2 \log(2\pi)$. Maximizar esta función equivale a minimizar $h(\theta) = \sum_{i=1}^n (x_i-\theta)^2$. Derivando, obtenemos que $h'(\theta) = 0$ si, y solo si, $\theta = \overline{x}$. Además, es rutinario comprobar que $\overline{x}$ es el mínimo absoluto de h. Por tanto, $\overline{x}$ es el máximo absoluto de $L(\theta; x)$. Tenemos pues $\hat{\theta}(x) = \overline{x}$. △

A continuación pretendemos extender el método de la máxima verosimilitud para estimar $g(\theta)$, donde $g : \Theta \to \Theta'$ sobreyectiva. Si la aplicación g fuese inyectiva, entonces podemos definir de norma natural

la verosimilitud de $\eta \in \Theta'$ como $L^*(\eta|x) = L(g^{-1}(\eta)|x)$. Claramente, el valor que maximiza $L^*(\eta|x)$, que denotaremos $\hat{g}(x)$, es $g(\hat{\theta}(x))$. Sin embargo, los casos que presentan relevancia práctica son aquellos en los que g no es inyectiva ya que de esta forma conseguimos reducir la dimensionalidad del espacio de parámetros. Necesitamos extender la definición de verosimilitud para abordar esta problemática.

Definición 3.4. En el contexto anterior, definimos la verosimilitud inducida por g como

$$L^*(\eta|x) = \sup\{L(\theta; x) : \theta \in g^{-1}(\eta)\}.$$

El valor $\hat{g}(x)$ que maximiza $L^*(\eta|x)$ se denomina estimador maximo verosímil de $g(\theta)$.

La definición anterior es artificial en el sentido de que se realiza con el fin de poder mantener la propiedad de invarianza del estimador máximo verosímil, que se recoge en el siguiente teorema.

Teorema 3.4 (Invarianza de Zehna)**.** *Para cualquier aplicación sobreyectiva $g : \Theta \to \Theta'$ se tiene que $\hat{g}(X) = g(\hat{\theta}(X))$.*
Demostración. En primer lugar, la definición de la verosimilitud inducida proporciona

$$\sup_{\eta \in \Theta'} L^*(\eta|x) = \sup_{\eta \in \Theta'} \sup\{L(\theta; x) : \theta \in g^{-1}(\eta)\} = \sup_{\theta \in \Theta} L(\theta; x).$$

Por tanto, si la verosimilitud tiene un máximo global $\hat{\theta}(x)$, entonces lo tiene la verosimilitud inducida (el recíproco puede no ser cierto) y se alcanza en $g(\hat{\theta}(x))$. □

3.3. Teoría general de estimadores

En esta sección introduciremos conceptos y definiciones relacionados con estimadores arbitrarios. El objetivo de esta teoría es dotarnos de herramientas que nos permitan abordar el estudio práctico de estimadores concretos, como el estimador máximo verosímil.

3.3.1. Estadísticos suficientes

Fijemos $\{f(x|\theta) : \theta \in \Theta\}$ una familia de distribuciones. Sea $X = (X_1, \ldots, X_n)$ una muestra de la distribución $f(x|\theta_0)$. Nuestro objetivo es inferir el parámetro θ_0 a partir de la muestra. El concepto de estadístico suficiente nos permitirá separar la información contenida en X en dos partes. Una parte contiene toda la información útil sobre θ_0 mientras que la otra parte no dependerá del parámetro θ_0. Consecuentemente, podemos ignorar esta última parte.

Intuitivamente, un estadístico T es suficiente para la familia de distribuciones considerada si $T(X)$ nos permite estimar θ_0 tan bien como lo permite toda la muestra X. Procedemos a dar la definición matemática.

Definición 3.5. Un estadístico $T(X_1, X_2, \ldots, X_n)$ es suficiente si para cada $\theta \in \Theta$ y t la distribución condicional de $X_1, X_2, \ldots, X_n$ respecto de θ y $T(X) = t$ no depende de θ.

El teorema de factorización de Neyman nos proporciona un criterio práctio para ver si un estadístico es suficiente.

Teorema 3.5. *Sea $\{f(x|\theta) : \theta \in \Theta\}$ una familia de distribuciones. Sea $X = (X_1, \ldots, X_n)$ una muestra de la distribución $f(x|\theta)$. Sea $T(X_1, X_2, \ldots, X_n)$ un estadístico. Entonces, T es suficiente si y solo si la función de verosimilitud puede factorizarse de la siguiente forma*

$$L(\theta; x_1, x_2, \ldots, x_n) = h(t|\theta)g(x_1, x_2, \ldots, x_n),$$

donde $t = T(x_1, \ldots, x_n)$ *y* $g(x_1, x_2, \ldots, x_n)$ *no depende de* θ.

Nótese que si T es un estadístico suficiente, entonces a falta de una constante $h(t|\theta)$ es la función de densidad de la variable $T(X)$.

Si encontramos un estadístico suficiente, entonces podemos inferir el parámetro θ_0 utilizando solamente la función $h(t|\theta)$. Interesa pues que el codominio del estadístico suficiente sea lo más simple posible. Los estadísticos suficientes son especialmente interesantes al aplicar el método de la máxima verosimilitud.

Ejemplo 3.6 (Distribución normal, media desconocida): Sea $X = (X_1, \ldots, X_n)$ una muestra de $N(x|\mu, \sigma^2)$ donde solamente σ^2 es conocido ($\theta = \mu$). Es fácil verificar que

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - \mu)^2 = (n-1)S^2 + n(\overline{x} - \mu)^2.$$

A partir de la igualdad anterior obtenemos

$$f(x|\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\sum_{i=1}^{n}(x_i - \mu)^2/(2\sigma^2)) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-((n-1)S^2 + n(\overline{x} - \mu)^2)/(2\sigma^2)).$$

Definimos

$$g(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-(n-1)S^2/(2\sigma^2)) \text{ y } h(t|\mu) = \exp(n(t-\mu)^2)/(2\sigma^2)).$$

Tenemos que $f(x|\theta) = h(\overline{x}|\theta)g(x_1, x_2, \ldots, x_n)$ y, por tanto, $T(x) = \overline{x}$ es suficiente. △

Ejemplo 3.7 (Distribución normal, ambos parámetros son desconocidos): Sea $X = (X_1, \ldots, X_n)$ una muestra de $N(x|\mu, \sigma^2)$ donde μ y σ^2 son desconocidos ($\theta = (\mu, \sigma^2)$). Tenemos que

$$f(x|\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\sum_{i=1}^{n}(x_i - \mu)^2/(2\sigma^2)) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2)/(2\sigma^2)).$$

Por tanto, el estadístico $T(X_1, \ldots, X_n) = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$ es suficiente (tomamos $g(x_1, x_2, \ldots, x_n) = 1$). También podemos desarrollar $f(x|\theta)$ como sigue

$$f(x|\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-(\sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - \mu)^2)/(2\sigma^2)).$$

Consecuentemente, el estadístico $T(X_1, \ldots, X_n) = (\overline{x}, S^2)$ también es suficiente. Nótese que el estadístico $T(X_1, \ldots, X_n) = (X_1, \ldots, X_n)$ es trivialmente suficiente, pero no aporta ninguna información. △

3.3.2. Score, hipótesis de regularidad y función de información de Fisher

Definición 3.6. Sea $\Theta \subset \mathbb{R}^m$ un abierto y sea $\{f(X|\theta) : \theta \in \Theta\}$ una familia de funciones de densidad. Sea $X = (X_1, \ldots, X_n)$ una muestra que sigue una distribución con función de densidad $f(X|\theta_0)$. Si la función de verosimilitud para los valores $x = (x_1, \ldots, x_n)$ es diferenciable en $\theta \in \Theta$, entonces definimos el score de θ como el gradiente de la función $\log L(x; \theta)$ y lo denotamos $S(x; \theta)$.

Intuitivamente el score indica la sensibilidad de la verosimilitud en un punto. Nos centraremos en el estudio del score cuando Θ es un abierto de $\mathbb{R}$. En tal caso

$$S(x; \theta) = \frac{\partial}{\partial\theta} \log f(x|\theta) = \frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)}.$$

Supongamos que el score de θ existe para cualesquiera valores de la muestra $x_1, \ldots, x_n$. En tal caso es natural considerar la función $E_{X|\theta}[S(X;\theta)]$, que depende solamente de θ. Si θ fuese el parámetro a estimar, entonces $E_{X|\theta}[S(X;\theta)]$ mide la sensibilidad media de la verosimilitud en θ.

Lema 3.8. *Sea $\Theta \subset \mathbb{R}$ un abierto y sea $\{f(X|\theta) : \theta \in \Theta\}$ una familia de funciones de densidad que verifican las siguientes condiciones de regularidad:*

a) Para cualesquier muestra $x = (x_1, \ldots, x_n)$ la función $L(x;\theta)$ es diferenciable para todo $\theta \in \Theta$.

b) Se verifica

$$\frac{\partial}{\partial\theta}\int_X f(x|\theta)dx = \int_X \frac{\partial}{\partial\theta} f(x|\theta)dx.$$

Entonces, $E_{X|\theta}[S(X;\theta)] = 0$.
Demostración. Tenemos que

$$E_{X|\theta}[S(X;\theta)] = \int_X \frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)} f(x|\theta)dx = \int_X \frac{\partial}{\partial\theta} f(x|\theta)dx = \frac{\partial}{\partial\theta}\int_X f(x;\theta)dx = \frac{\partial}{\partial\theta}1 = 0. \qquad \square$$

En el resultado anterior aparecen por primera vez hipótesis de regularidad sobre las distribuciones a estudiar. Nótese que en la práctica normalmente vamos a trabajar con distribuciones que satisfagan estas hipótesis. La hipótesis b) se verifica si la derivada de la verosimilitud es continua [5]. Consecuentemente, todas las distribuciones continuas estudiadas, exceptuando la distribución de Laplace, cumplen estas hipótesis de regularidad (su función de densidad es de clase infinito con respecto a θ).

Definición 3.7. Sea $\Theta \subset \mathbb{R}$ un abierto y sea $\{f(X|\theta) : \theta \in \Theta\}$ una familia de funciones de densidad para la cual siempre existe el score. Dado $\theta \in \Theta$, definimos la función de información de Fisher en θ como el segundo momento de la variable aleatoria $S(X;\theta)$, donde $X = (X_1, \ldots, X_n)$ es una muestra de la distribución con función de densidad $f(X|\theta)$. Se denota $\mathcal{I}(\theta) := E_{X|\theta}[S(X;\theta)^2] \geq 0$.

Si en determinado contexto no está clara la muestra X para la cual calculamos la información de Fisher, entonces la denotamos $\mathcal{I}_X$ o $\mathcal{I}^X$.

El siguiente resultado nos permite explicar por qué se define de esta forma la información de Fisher.

Corolario 3.9. *Bajo las hipótesis de regularidad del Lema 3.8, tenemos que $\mathcal{I}(\theta) = Var_{X|\theta}(S(X;\theta))$.*
Demostración. Nótese que $Var_{X|\theta}(S(X;\theta)) = \mathcal{I}(\theta) - E_{X|\theta}[S(X;\theta)]^2$. El Lema 3.8 nos indica que $E_{X|\theta}[S(X;\theta)] = 0$. $\square$

Como consecuencia, la información de Fisher nos informa de cómo varía la sensibilidad de la verosimilitud en θ. Si la información de Fisher es pequeña, entonces la sensibilidad de la verosimilitud en θ no depende prácticamente de la muestra utilizada y, por tanto, siempre será cercana a cero. Si por el contrario la información de Fisher es muy grande, entonces la sensibilidad de la verosimilitud en θ varía mucho en función de la muestra con la que se trabaje. Si utilizamos el estimador máximo verosímil, entonces estamos maximizando el logaritmo de la verosimilitud. Buscamos pues aquellos θ que sean extremos relativos de $\log L(x;\theta)$ y, por tanto, verifiquen $S(x;\theta) = 0$. Consecuentemente, nos interesa que $\mathcal{I}(\theta)$ sea grande para todo θ ya que de esta forma podremos discriminar aquellos θ que tengan score no nulo (no son extremos relativos de $\log L(x;\theta)$). Si en determinado θ la información de Fisher es muy pequeña, obtendremos que θ es un candidato a estimador máximo verosímil para casi cualquier muestra, incluso para muestras poco probables bajo ese parámetro, lo cual dificulta el correcto cómputo del estimador.

En lo que sigue habitualmente exigiremos unas hipótesis de regularidad más fuertes, denominadas hipótesis o condiciones de regularidad de Cramer-Rao. Estas hipótesis son las siguientes:

i) Θ es un abierto de $\mathbb{R}$.

ii) Para cualquier muestra $x = (x_1, \ldots, x_n)$, la verosimilitud $L(x|\theta)$ es dos veces derivable en Θ.

iii) $\frac{\partial^i}{\partial \theta^i} \int_X f(x|\theta)dx = \int_X \frac{\partial^i}{\partial \theta^i} f(x|\theta)dx$ para $i = 1, 2$.

iv) Para cada $\theta \in \Theta$ se tiene $0 < \mathcal{I}(\theta) < +\infty$.

Todas las distribuciones continuas estudiadas, exceptuando la distribuciónd de Laplace, verifican estas hipótesis de regularidad.

El siguiente lema profundiza en nuestro entendimiento de la función de información de Fisher.

Lema 3.10. *Bajo hipótesis de regularidad de Cramer-Rao tenemos que*

$$\mathcal{I}(\theta) = E_{X|\theta}\left[-\frac{\partial^2}{\partial \theta^2} \log f(X;\theta)\right].$$

Demostración. En primer lugar, podemos escribir

$$\frac{\partial^2}{\partial \theta^2} \log f(X|\theta) = \frac{\frac{\partial^2}{\partial \theta^2} f(X|\theta)}{f(X|\theta)} - \left(\frac{\frac{\partial}{\partial \theta} f(X|\theta)}{f(X|\theta)}\right)^2 = \frac{\frac{\partial^2}{\partial \theta^2} f(X|\theta)}{f(X|\theta)} - \left(\frac{\partial}{\partial \theta} \log f(X|\theta)\right)^2.$$

La demostración finaliza al tomar esperanzas en la expresión anterior y darse cuenta de que

$$E_{X|\theta}\left[\frac{\frac{\partial^2}{\partial \theta^2} f(X|\theta)}{f(X|\theta)}\right] = \int_X \frac{\partial^2}{\partial \theta^2} f(x|\theta)\, dx = \frac{\partial^2}{\partial \theta^2} \int_X f(x|\theta)\, dx = \frac{\partial^2}{\partial \theta^2}\, 1 = 0. \qquad \square$$

Como consecuencia, la información de Fisher también indica cuál es la curvatura media de la función $\log L(x;\theta)$, que como vemos, en media es negativa ($\mathcal{I}(\theta) \geq 0$). Para calcular el estimador máximo verosímil intentamos maximizar $\log L(x;\theta)$. Si la función de información de Fisher es habitualmente grande, entonces en media tendremos máximos relativos muy claros.

El Lema 3.10 nos permite calcular la información de Fisher de forma más sencilla, como muestran los siguientes ejemplos.

Ejemplo 3.11: Calculamos la función de información de Fisher de $X \sim B(x|n,\theta)$ donde n es conocido. Recordemos que $\log f(X|n,\theta) = \log \binom{n}{X} + X \log\theta + (n - X)\log(1-\theta)$. Derivando dos veces respecto de θ obtenemos

$$\frac{\partial^2}{\partial \theta^2} f(X|n,\theta) = \frac{-X}{\theta^2} + \frac{-(n-X)}{(1-\theta)^2}.$$

Por tanto, la función de información de Fisher responde a

$$\mathcal{I}_X(\theta) = E_{X|\theta}\left[\frac{X}{\theta^2} + \frac{(n-X)}{(1-\theta)^2}\right] = n\left(\frac{1}{\theta} + \frac{1}{1-\theta}\right) = \frac{n}{\theta(1-\theta)}. \qquad \triangle$$

Ejemplo 3.12: Calculamos la función de información de Fisher de $X \sim N(x|\mu,\sigma^2)$ para varias configuraciones de la distribución normal.

- El parámetro σ^2 es conocido. Tenemos que $\log f(X|\mu,\sigma^2) = -(X-\mu)^2/(2\sigma^2) - \log(\sqrt{2\pi}) - \log(\sigma^2)/2$. Consecuentemente, deducimos que

$$\frac{\partial^2}{\partial \mu^2} f(X|\mu,\sigma^2) = \frac{-1}{\sigma^2}.$$

Por tanto, $\mathcal{I}(\mu) = 1/\sigma^2$.

- El parámetro μ es conocido. Obtenemos que

$$\frac{\partial^2}{\partial(\sigma^2)^2} f(X|\mu,\sigma^2) = -\frac{(X-\mu)^2}{\sigma^6} + \frac{1}{2\sigma^4}.$$

Por tanto, podemos calcular $\mathcal{I}(\sigma^2)$ utilizando que $E[(X-\mu)^2] = Var(X) = \sigma^2$. Obtenemos que

$$\mathcal{I}(\sigma^2) = E_{X|\sigma^2}[\frac{(X-\mu)^2}{\sigma^6} - \frac{1}{2\sigma^4}] = \frac{1}{\sigma^6} Var((X-\mu)^2) - \frac{1}{2\sigma^4} = \frac{1}{2\sigma^4}. \qquad \triangle$$

Comentario 3.13. Bajo hipótesis de regularidad de Cramer-Rao, si $X = (X_1, \ldots, X_n)$ es una muestra de $f(X|\theta)$, entonces tenemos que

$$\frac{\partial^2}{\partial\theta^2} \log(f(X;\theta)) = \frac{\partial^2}{\partial\theta^2} \left(\sum_{i=1}^n \log(f(X_i;\theta)) \right) = \sum_{i=1}^n \frac{\partial^2}{\partial\theta^2} \log(f(X_i;\theta)).$$

Consecuentemente, $\mathcal{I}^{X_1,\ldots,X_n}(\theta) = \sum_{i=1}^n \mathcal{I}^{X_i}(\theta) = n\mathcal{I}^{X_i}(\theta)$.

Lema 3.14. *Bajo hipótesis de regularidad de Cramer-Rao, sea $T(X_1, \ldots, X_n)$ un estadístico tal que su ditribución inducida también verifica las hipótesis de regularidad de Cramer-Rao. Entonces, para cualquier $\theta \in \Theta$ se tiene*

$$\mathcal{I}_{T(X)}(\theta) \leq \mathcal{I}_X(\theta).$$

Además, la igualdad se da para todo $\theta \in \Theta$ si, y solo si, T es suficiente.

En lo que sigue necesitaremos el siguiente lema.

Lema 3.15 (Desigualdad de Jenssen)**.** *Sean X una variable aleatoria cuya imagen está contenida en un intervalo I. Sea $g : I \to \mathbb{R}$ una función.*

a) Si g es convexa, entonces $E[g(X)] \geq g(E[X])$.

b) Si g es cóncava, entonces $E[g(X)] \leq g(E[X])$

Proposición 3.16. *Sea $X = (X_1, \ldots, X_n)$ una muestra de $f(X;\theta_0)$. Entonces, para cada $\theta_1 \in \Theta$ se tiene*

$$E_{X|\theta_0} \log f(X|\theta_0) \geq E_{X|\theta_0} \log f(X|\theta_1).$$

Demostración. Por la desigualdad de Jenssen obtenemos

$$E_{X|\theta_0} \log \frac{f(X|\theta_1)}{f(X|\theta_0)} \leq \log \int_X \frac{f(X|\theta_1)}{f(X|\theta_0)} f(X|\theta_0)\, dx = \log \int_X f(X|\theta_1)\, dx = 0,$$

de donde se deduce el resultado. □

Cabe mencionar que la información de Fisher puede definirse cuando $\Theta \subset \mathbb{R}^m$. Incluimos la definición por complitud, aunque no entraremos en ella a fondo.

Definición 3.8. Sea $\Theta \subset \mathbb{R}^m$ un abierto y sea $\{f(X|\theta) : \theta \in \Theta\}$ una familia de funciones de densidad para la cual siempre existe el score. Dado $\theta \in \Theta$, definimos la función de información de Fisher en θ como

$$(\mathcal{I}(\theta))_{i,j} = E_{X|\theta} \left[\left(\frac{\partial}{\partial\theta_i} \log f(X;\theta) \right) \left(\frac{\partial}{\partial\theta_j} \log f(X;\theta) \right) \right], \ 1 \leq i,j \leq m,$$

donde $X = (X_1, \ldots, X_n)$ es una muestra de la distribución con función de densidad $f(X|\theta)$.

Bajo determinadas hipótesis de regularidad se puede probar que para cada $1 \leq i, j \leq n$ se verifica

$$\left(\mathcal{I}(\theta)\right)_{i,j} = -E_{X|\theta}\left[\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\log f(X;\theta)\right].$$

Ejemplo 3.17: Calculamos la función de información de Fisher de una variable aleatoria X que siga una distribución $Beta(x|p,q)$, donde $\theta = (p,q)$. Recordemos que

$$\log f(x|p,q) = (p-1)\log x + (q-1)\log(1-x) - \log\beta(p,q). \tag{2}$$

Recordemos que $\beta(p,q) = \Gamma(p)\Gamma(q)/\Gamma(p+q)$. Habitualmente al cociente $\Gamma'(p)/\Gamma(p)$ se le llama función digamma y se denota $\psi(p)$. Las derivadas parciales de (2) son las siguientes:

a) $\frac{\partial}{\partial p}\log f(x|p,q) = \log x + \psi(p+q) - \psi(p)$.

b) $\frac{\partial}{\partial q}\log f(x|p,q) = \log(1-x) + \psi(p+q) - \psi(q)$.

c) $\frac{\partial^2}{\partial p^2}\log f(x|p,q) = \psi'(p+q) - \psi'(p)$.

d) $\frac{\partial^2}{\partial q^2}\log f(x|p,q) = \psi'(p+q) - \psi'(q)$.

e) $\frac{\partial^2}{\partial p\partial q}\log f(x|p,q) = \psi'(p+q)$.

Nótese que las derivadas obtenidas no dependen de x y, por tanto, al tomar esperanzas obtenemos las mismas derivadas. Consecuentemente, la función de información de Fisher de X viene dada por

$$\mathcal{I}_X(p,q) = -\begin{pmatrix}\psi'(p+q)-\psi'(p) & \psi'(p+q)\\ \psi'(p+q) & \psi'(p+q)-\psi'(q)\end{pmatrix}. \qquad \triangle$$

3.3.3. Estimadores insesgados

Para comprobar cómo de bueno es un estimador T podemos definir una función de pérdida $L(\theta, T(X))$ que indique la pérdida asociada a estimar un parámetro mediante T si su verdadero valor es θ. A partir de la función de pérdida definimos la función de riesgo, que asocia a cada posible valor del parámetro la pérdida media producida por el estimador. La función de riesgo viene dada por

$$R_T^L(\theta) = E_{X|\theta}[L(\theta, T(X))].$$

Un estimador T que "minimice uniformemente" la función de riesgo hará mejores estimaciones en media. Con minimizar uniformemente queremos decir que para cada estimador T' se tiene que

$$R_T^L(\theta) \leq R_{T'}^L(\theta)\ \forall\ \theta \in \Theta.$$

En esta sección introducimos un tipo particular de estimadores que minimizan determinada función de riesgo.

Definición 3.9. Se denomina sesgo de un estimador T de $g(\theta)$ a la diferencia entre la esperanza del estimador y el verdadero valor del parámetro a estimar. Diremos que un estimador es insesgado si para cualquier posible valor del parámetro a estimar su sesgo es nulo.

Nótese que el sesgo de un estimador es la función de riesgo asociada a la pérdida $L(\theta, T(X)) = g(\theta) - T(X)$. Un estimador insesgado verifica $0 = g(\theta) - E_{X|\theta}T(X)$ y, por tanto, minimiza uniformemente la función de riesgo. Aunque esta propiedad puede parecer a priori interesante, puede suceder que en la

práctica el estimador insesgado no proporcione estimaciones de calidad si la varianza $Var_{X|\theta}(T(X))$ es muy alta.

Claramente, la media muestral es un estimador insesgado de la media de la distribución. El siguiente resultado nos muestra otro ejemplo de un estimador insesgado.

Proposición 3.18. *Sea $X_1, \dots, X_n$ una muestra de alguna población con función de densidad $f(X|\theta_0)$. Definimos la varianza muestral como*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Entonces, S^2 es un estimador insesgado de la varianza de la distribución.
Demostración. Nótese que $\sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n X_i^2 - n\overline{X}^2$. Consecuentemente tenemos

$$E\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = \sum_{i=1}^{n} E\left[X_i^2\right] - nE[\overline{X}^2] = n(E\left[X_i^2\right] - E[\overline{X}^2]).$$

Utilizando que $Var(\overline{X}) = Var(X_i)/n$ y $E[\overline{X}] = E[X_i]$ obtenemos

$$E[X_i^2] - E[\overline{X}^2] = Var(X_i) + E[X_i]^2 - Var(\overline{X}) - E[\overline{X}]^2 = \frac{n-1}{n}Var(X_i).$$

Por tanto, $E[S^2] = Var(X_i)$ como se quería. □

La función de información de Fisher juega un papel importante en el estudio de los estimadores insesgados como muestra el siguiente Teorema.

Teorema 3.19 (Cota de Cramer-Rao)**.** *Supongamos que se verifican las hipótesis de regularidad de Cramer-Rao. Sea $g : \Theta \to \mathbb{R}$ de clase 1. Sea $\hat{\theta}$ un estimador insesgado de $g(\theta)$ tal que*

$$\int_X \left|\hat{\theta}(x)\frac{\partial}{\partial\theta}f(x|\theta)\right| dx < \infty.$$

Entonces, para todo $\theta \in \Theta$

$$Var_{X|\theta}(\hat{\theta}) \geq \frac{g'(\theta)^2}{\mathcal{I}(\theta)}.$$

Demostración. Puesto que $\hat{\theta}$ es insesgado tenemos que

$$g(\theta) = \int_X \hat{\theta}(x) f(x|\theta)\, dx.$$

Podemos derivar respecto de θ la expresión anterior y utilizar que $\int_X \frac{\partial}{\partial\theta} f(x|\theta)\, dx = 0$, obteniendo

$$g'(\theta) = \int_X \hat{\theta}(x)\frac{\partial}{\partial\theta}f(x|\theta)\, dx = \int_X \left(\hat{\theta}(x) - g(\theta)\right)\frac{\partial}{\partial\theta}f(x|\theta)\, dx. \tag{3}$$

Aplicamos la desigualdad de Cauchy-Schwarz al miembro de la derecha de (3), obteniendo

$$g'(\theta)^2 \leq \left(\int \left(\hat{\theta}(x) - g(\theta)\right)^2 f(x|\theta)\, dx\right)\left(\int \left(\frac{\partial}{\partial\theta}(\log f(x|\theta))\right)^2 f(x|\theta)\, dx\right) = Var_{X|\theta}(\hat{\theta})\,\mathcal{I}(\theta). \quad \square$$

Corolario 3.20. *Supongamos que se verifican las hipótesis de regularidad de Cramer-Rao. Sea $\hat{\theta}$ un estimador insesgado de θ tal que*

$$\int_X \left| \hat{\theta}(x) \frac{\partial}{\partial \theta} f(x|\theta) \right| dx < \infty.$$

Entonces, para todo $\theta \in \Theta$

$$Var_{X|\theta}(\hat{\theta}) \geq \frac{1}{\mathcal{I}_X(\theta)} = \frac{1}{n\mathcal{I}_{X_i}(\theta)}.$$

La cota de Cramer-Rao nos dice que si la información de Ficher en θ es pequeña, entonces cualquier estimador insesgado tendrá una gran varianza y, por tanto, será inestable ante pequeños cambios en la muestra.

Definición 3.10. Un estimador se dice eficiente si alcanza la cota de Cramer-Rao para todo $\theta \in \Theta$.

3.3.4. Consistencia de sucesiones de estimadores

Nos interesa que los estimadores tiendan al parámetro de la distribución cuando el tamaño de la muestra diverge. En tal caso, podemos mejorar el resultado del estimador recurriendo a una mayor muestra de la población.

Definición 3.11. Consideremos una familia de densidades $\{f(x|\theta) : \theta \in \Theta\}$. Una sucesión de estimadores $\hat{\theta}_n$ de $g(\theta)$ es consistente para $\theta_0 \in \Theta$ si toda sucesión X_n de variables aleatorias independientes e idénticamente distribuidas con función de distribución $f(x|\theta_0)$ la sucesión $\hat{\theta}_n(X_1, \ldots, X_n)$ converge en probabilidad (P_{θ_0}) a $g(\theta_0)$. Si $\hat{\theta}_n$ es consitente para todo $\theta_0 \in \Theta$, entonces decimos que es consistente.

Teorema 3.21. *Sea $\hat{\theta}_n$ una sucesión de estimadores de $g(\theta)$ verificando*

a) $\lim_{n\to\infty} E_\theta[\hat{\theta}_n] = g(\theta)$*;*

b) $\lim_{n\to\infty} Var_\theta(\hat{\theta}_n) = 0$.

Entonces, $\hat{\theta}_n$ es consistente para θ.

Demostración. Sea $\varepsilon > 0$. La desigualdad de Markov nos proporciona

$$P_\theta[|\hat{\theta}_n - g(\theta)| \geq \varepsilon] \leq \varepsilon^{-2} E[(\hat{\theta}_n - g(\theta))^2] = \varepsilon^{-2} \left(Var(\hat{\theta}_n) + (E\hat{\theta}_n - g(\theta))^2 \right).$$

La prueba finaliza al recordar que el último término converge a 0. □

Otra propiedad interesante de un estimador cuando la muestra tiene a infinito es la siguiente.

Definición 3.12. Consideremos una familia de densidades $\{f(x|\theta) : \theta \in \Theta\}$. Una sucesión de estimadores $\hat{\theta}_n$ de $g(\theta)$ es asintóticamente normal para $\theta_0 \in \Theta$ si toda sucesión X_n de variables aleatorias independientes e idénticamente distribuidas con función de distribución $f(x|\theta_0)$ la sucesión $\sqrt{n}(\hat{\theta}_n(X_1, \ldots, X_n) - g(\theta_0))$ converge en ley a una distribución $N(X|0, \sigma^2)$ para cierto $\sigma^2 > 0$.

Proposición 3.22. *Todo estimador asintóticamente normal es consistente.*

Demostración. Sea $\hat{\theta}_n$ un estimador asintóticamente normal de $g(\theta)$. Tenemos que $n(\hat{\theta}_n - g(\theta))^2$ converge en ley a $Gamma(X|1/2, 1/(2\sigma^2))$ por la Proposición **??**. Sea F la función de distribución de $Gamma(X|1/2, 1/(2\sigma^2))$ y sea $\varepsilon > 0$. Vamos a probar que $P[(\hat{\theta}_n - g(\theta))^2 < \varepsilon] \to 1$. En efecto, sea $1 > \alpha \geq 0$. Tomamos y tal que $F(y) = \alpha$. Tenemos que

$$P[n(\hat{\theta}_n - g(\theta))^2 < y] \to F(y) = \alpha.$$

Por tanto, para cada $\delta > 0$ existe n_0 tal que $\varepsilon > y/n$ y para cada $n \geq n_0$ tenemos

$$P[(\hat{\theta}_n - g(\theta))^2 < \varepsilon] \geq P[n(\hat{\theta}_n - g(\theta))^2 < y] \geq \alpha - \delta.$$

Deducimos que $\liminf P[(\hat{\theta}_n - g(\theta))^2 < \varepsilon] \geq \alpha - \delta$. De la arbitrariedad de δ y α se deduce el resultado. □

Como es natural, el recíproco del anterior no es cierto.

3.4. Estudio teórico del estimador máximo verosímil

En este punto nos preguntamos cuándo está bien definido el estimador máximo verosímil. En tal caso nos interesa saber si el método de la máxima verosimilitud nos proporciona un estimador consistente. Para ello aplicamos los resultados teóricos vistos en la sección anterior.

Proposición 3.23. *Si existe un estadístico suficiente T par ala familia de distribuciones $\{f(X|\theta) : \theta \in \Theta\}$ y $\hat{\theta}$ es un estimador máximo verosímil, entonces $\hat{\theta}$ depende solamente de $T(X)$.*

Demostración. Por el teorema de factorización de estimadores suficientes podemos escribir $f(X|\theta) = h(X)g(T(X), \theta)$. Maximizar $L(x;\theta) = h(x)g(T(x);\theta)$ equivale a maximizar $g(T(x);\theta)$. Por tanto, $\hat{\theta}$ depende solamente de $T(x)$. □

Teorema 3.24. *Bajo las hipótesis de regularidad de Cramer-Rao se verifican las siguientes afirmaciones:*

a) Existe n_0 tal que para cada $n \geq n_0$ la ecuación en probabilidad $0 = \sum_{i=0}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta)$ tiene solución única. A esta solución se le llama $\hat{\theta}_n(X_1, \ldots, X_n)$. En dicho punto se maximiza la verosimilitud.

b) $\hat{\theta}(X_1, \ldots, X_n)$ es consistente. De hecho, se puede probar que la convergencia a θ_0 es casi segura.

Demostración. Para cualquier muestra X de $f(X|\theta_0)$ tenemos que

$$0 > -\mathcal{I}(\theta_0) = E_{X|\theta_0}\left[\frac{\partial^2}{\partial\theta^2} \log f(X;\theta_0)\right] = \frac{\partial}{\partial\theta} E_{X|\theta_0}\left[\frac{\partial}{\partial\theta} \log f(X;\theta_0)\right].$$

Consecuentemente, $E_{X|\theta_0}[S(X;\theta)]$ es decreciente en un entorno de θ_0. Recordemos que $E_{X|\theta_0}[S(X;\theta_0)] = 0$ por el Lema 3.8. Por tanto, existe $\varepsilon > 0$ tal que

- $E_{X|\theta_0}[S(X;\theta)] > 0$ para todo $\theta \in (\theta_0 - \varepsilon, \theta_0)$;
- $E_{X|\theta_0}[S(X;\theta)] < 0$ para todo $\theta \in (\theta_0, \theta_0 + \varepsilon)$.

Esto implica que θ_0 es un máximo relativo de $E_{X|\theta_0}[\log f(X|\theta)]$. INCOMPLETO. □

Teorema 3.25. *Bajo hipótesis de regularidad de Cramer-Rao, si $\hat{\theta}(X_1, \ldots, X_n)$ es un estimador máximo verosímil consistente, entonces es asintóticamente normal. Además, la varianza de la distribución normal asociada es $1/\mathcal{I}_{X_1}(\theta_0)$.*

Demostración. Escribimos $L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i|\theta)$. Tenemos que $L_n'(\hat{\theta}_n) = 0$. El teorema del valor medio nos proporciona un θ_n entre θ_0 y $\hat{\theta}_n$ tal que $0 = L_n'(\hat{\theta}) = L_n'(\theta_0) + L_n''(\theta_n)(\hat{\theta}_n - \theta_0)$. Por tanto,

$$\sqrt{n}(\theta_0 - \hat{\theta}_n) = \frac{\sqrt{n}L_n'(\theta_0)}{L_n''(\theta_n)}. \tag{4}$$

Por el teorema central del límite tenemos que

$$\sqrt{n}L_n'(\theta_0) = \sqrt{n}(L_n'(\theta_0) - E_{\theta_0}[S(X_1;\theta_0)]) \to N(0, \mathcal{I}_{X_1}(\theta_0)),$$

donde hemos utilizado que $Var_{\theta_0}(S(X_1;\theta_0)) = \mathcal{I}_{X_1}(\theta_0)$. Estudiamos ahora el denominador de (4). Puesto que $\hat{\theta}_n$ converge en probabilidad a θ_0 y θ_n se encuentra entre ambos, tenemos que θ_n converge en probabilidad a θ_0. Además, la ley uniforme de los grandes números nos garantiza que

$$L_n''(\theta) \xrightarrow{P_{\theta_0}} E_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2} f(X_1|\theta)\right],$$

siendo la convergencia uniforme en espacios de parámetros compactos. Por tanto, tomando un compacto que contenga una cola de θ_n obtenemos la mencionada convergencia uniforme. De esta convergencia uniforme se desprende que

$$L_n''(\theta_n) \xrightarrow{P_{\theta_0}} E_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2} f(X_1|\theta_0)\right] = -\mathcal{I}_{X_1}(\theta_0).$$

Hemos obtenido pues

$$\sqrt{n}(\theta_0 - \hat{\theta}_n) = \frac{\sqrt{n}L_n'(\theta_0)}{L_n''(\theta_n)} \to N(0, \frac{1}{\mathcal{I}_{X_1}(\theta_0)}).$$ □

4. La familia exponencial

En esta sección estudiamos una amplia familia de distribuciones, denominada la familia exponencial. Veremos que gran parte de las distribuciones que hemos estudiado hasta el momento pertenecen a esta familia.

Definición 4.1. Una variable aleatoria se distribuye respecto de una *familia exponencial* si su función de distribución es de la forma

$$f(x|\theta) = h(x)\exp\left(\sum_{i=1}^{k}\theta_i T_i(x) + \Psi(\theta)\right), \tag{5}$$

donde $\theta = (\theta_1, \ldots, \theta_k)$ y $h(x) \geq 0$, $\Psi(\theta)$, $T_1(x), \ldots, T_k(x)$ son funciones reales.

Las familias exponenciales presentan características matemáticas y estadísticas muy convenientes. De estas características cabe destacar el siguiente resultado, que utiliza estadísticos suficientes introducidos en la Sección 3.3.1.

Proposición 4.1. *Sea* $\{f(X|\theta) : \theta \in \Theta\}$ *una familia exponencial y sea una muestra* $\underset{\sim}{X} = (X_1, \ldots, X_n)$. *Entonces,* $T(X) = (\sum_{j=1}^{n} T_i(X_j))_{i=1,\ldots,k}$ *es un estadístico suficiente de dimensión* k.

Demostración. En efecto, utilizando (5) basta escribir $f(\underset{\sim}{x}|\theta)$ como sigue

$$f(\underset{\sim}{x}|\theta) = \prod_{j=1}^{n} h(x_j)\exp\left(\sum_{j=1}^{n}\sum_{i=1}^{k}\theta_i T_i(x_j) + n\Psi(\theta)\right) = \prod_{j=1}^{n} h(x_j)\exp\left(\sum_{i=1}^{k}\theta_i \sum_{j=1}^{n} T_i(x_j) + n\Psi(\theta)\right). \quad \square$$

Nótese que la dimensión del estadístico suficiente encontrado no depende de la muestra. A continuación mostramos algunos ejemplos de familiais exponenciales.

Ejemplo 4.2 (Distribución binomial): La función masa de probabilidad de una distribución binomial con n fijo puede escribirse como sigue

$$f(x|p) = \binom{n}{x}p^x(1-p)^{n-x} = \binom{n}{x}\exp(x\log(p)+(n-x)\log(1-p)) = \binom{n}{x}\exp(x\log(\frac{p}{1-p})+n\log(1-p)).$$

La aplicación $f(p) = \log(\frac{p}{1-p}) = \log(\frac{1}{1-p} - 1)$ es una biyección de $(0,1)$ a $\mathbb{R}$. En este punto hacemos el cambio de variable $\theta = f(p)$. Hemos obtenido que la distribución binomial es una familia exponencial de parámetro θ con $h(x) = \binom{n}{x}$, $T_1(x) = x$ y $\Psi(\theta) = n\log(1 - f^{-1}(\theta))$. Según la Proposición 4.1 un estadístico suficiente es $T(\underset{\sim}{X}) = \sum_{i=1}^{n} X_i$ y, por tanto, la media muestral, $T(\underset{\sim}{X}) = \overline{X}$, es otro estadístico suficiente. △

En el ejemplo anterior hemos tenido que realizar un cambio de variable del espacio paramétrico para poder escribir la distribución de Bernoulli como una familia exponencial. El nuevo espacio paramétrico obtenido es el *espacio paramétrico natural* de la familia. Para evitar trabajar con cambios de variables algunos autores definen las familias exponenciales como aquellas cuya función de desidad se puede escribir de la forma

$$f(x|\theta) = h(x)\exp\left(\sum_{i=1}^{k} w_i(\theta)T_i(x) + \Psi(\theta)\right), \tag{6}$$

donde $h(x) \geq 0$, $\Psi(\theta)$, $w_1(\theta), \ldots, w_k(\theta)$ y $t_1(x), \ldots, T_k(x)$ son funciones reales. En el Ejemplo 4.2 se tendría $w_1(p) = \log(\frac{p}{1-p})$ y $\Psi(p) = n\log(1-p)$.

Ejemplo 4.3 (Distribución normal): La función de densidad de la distribución normal se puede escribir de la forma

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}+\frac{x\mu}{\sigma^2}-\frac{\mu^2}{2\sigma^2}\right)$$

y, por tanto, es una familia exponencial con $h(x)=1$, $\Psi(\mu,\sigma^2) = -\frac{\mu^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)$, $T_1(x)=x$ y $T_2(x)=-x^2/2$. El espacio paramétrico natural se corresponde con $(1/\sigma^2, \mu/\sigma^2)$. No obstante, utilizamos los parámetros (μ,σ^2) debido a la interpretación estadística de los mismos.

Como consecuencia de la Proposición 4.1 obtenemos que cualquier variable aleatoria siguiendo una distribución $N(x|\mu,\sigma^2)$ verifica que $T(\underset{\sim}{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ es un estadístico suficiente. △

La mayoría de las distribuciones estudiadas hasta el momento forman una familia exponencial. La Tabla 1 muestra una lista de ejemplos. No obstante, no toda familia de distribuciones es exponencial, como sucede con las distribuciones uniformes.

DENSIDAD	NOTACIÓN	SOPORTE
$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$N(x\|\mu,\sigma^2)$	$\mathbb{R}$
$\frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{\frac{-x}{\beta}}$	$gamma(x\|\alpha,\beta)$	$(0,\infty)$
$\frac{1}{\Gamma(\frac{p}{2})2^{\frac{p}{2}}}x^{\frac{p}{2}-1}e^{\frac{-x}{2}}$	χ^2 con p grados de libertad	$(0,\infty)$
$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$beta(x\|\alpha,\beta)$	$(0,1)$
$\binom{n}{x}\theta^x(1-\theta)^{n-x}$	$B(x\|\theta,n)$	$0,1,\ldots,n$
$\frac{e^{-\lambda}\lambda^x}{x!}$	$P(x\|\lambda)$	$0,1,\ldots,n$

Tabla 1: Ejemplos de familias exponenciales.

5. Tests de hipótesis

En la Sección 3 se estudió el problema de estimación de parámetros. En este capítulo se estudiará otro problema clásico de la inferencia, los tests de hipótesis.

Definición 5.1. Sea $\{f(X|\theta) : \theta \in \Theta\}$ una familia de distribuciones. Supongamos que una variable aleatoria sigue una distribución $f(X|\theta)$ con $\theta \in \Theta$. Una hipótesis es una afirmación $\theta \in \Theta_0$, donde Θ_0 es un subconjunto de Θ. Una hipótesis es simple si es de la forma $\theta \in \{\theta_0\}$ para cierto $\theta_0 \in \Theta$. En tal caso se escribe $\theta = \theta_0$. Si una hipótesis no es simple, entonces decimos que es compuesta. Además, dos hipótesis $\theta \in \Theta_0$ y $\theta \in \Theta_1$ son excluyentes si Θ_0 y Θ_1 son disjuntos.

El objetivo de los tests de hipótesis es, dadas dos hipótesis excluyentes, aceptar una de las dos hipótesis como verdadera tras observar una muestra de tamaño n, donde n es un entero positivo fijado de antemano. Esta acción se denomina contrastar dos hipótesis. Generalmente el tratamiento de las dos hipótesis no es simétrico, esto es, una de las dos hipótesis tiene preferencia sobre la otra y solo será rechazada cuando la evidencia en su contra sea muy clara. Esta hipótesis se llama hipótesis nula, mientras que la otra hipótesis se denomina hipótesis alternativa. Las denotamos $H_0 : \theta \in \Theta_0$ y $H_1 : \theta \in \Theta_1$ respectivamente (recordemos que $\Theta_0 \cap \Theta_1 = \emptyset$).

El contraste de hipótesis surge de forma natural en multitud de ciencias e ingenierías. Por ejemplo, supongamos que estamos estudiando cómo afecta un determinado medicamento a la presión sanguínea de los pacientes. Queremos corroborar que el medicamento no tiene ningún efecto sobre la presión, hecho que es nuestra hipótesis nula. Sea θ la variación media de la presión de los pacientes al tomar el medicamento. Esta situación puede representarse como $H_0 : \theta = 0$ y $H_1 : \theta \neq 0$. Tras tomar muestras en varios pacientes tendremos que decidir cuál de las dos hipótesis aceptamos. Otro ejemplo puede ser el estudio de la proporción media de piezas defectuosas que fabrica una empresa, que denotamos θ. Sea θ_0 el máximo valor aceptable que puede alcanzar esa proporción. Queremos combrobar si la proporción de piezas defectuosas es menor o igual que θ_0, lo que se puede representar mediante las hipótesis $H_0 : \theta \leq \theta_0$ y $H_1 : \theta > \theta_0$. Habitualmente no es factible probar cada una de las piezas y, por tanto, tenemos que hacer inferencia a partir de una muestra. En ambos problemas θ debe ser el parámetro de una distribución. Por ejemplo, podríamos utilizar una distribución normal de media θ.

Nótese que en los dos ejemplos anteriores el espacio paramétrico es unión disjunta de Θ_0 y Θ_1. Podemos suponer que siempre se da esta situación. En efecto, el espacio paramétrico siempre se puede restringir a $\Theta_0 \cup \Theta_1$. En este contexto la hipótesis alternativa es la hipótesis complementaria a la hipótesis nula.

Definición 5.2. Consideremos una muestra aleatoria simple $\underset{\sim}{X} = (X_1, \ldots, X_n)$ de X. Un test de hipótesis es una regla que permite dividir el espacio muestral de $\underset{\sim}{X}$ en dos subconjuntos medibles disjuntos. Estos conjuntos se corresponden con aquellas muestras que aceptan la hipótesis nula como cierta (región de aceptación) y aquellas muestras que rechazan la hipótesis nula y aceptan la hipótesis alternativa (región crítica).

De aquí en adelante supondremos que n está fijado de antemano. Además, supondremos que la región crítica se corresponde con la imagen inversa de un boreliano, esto es, es de la forma $\underset{\sim}{X} \in R$ con $R \in \mathcal{B}^n$ ya que el resto de casos no tiene interés práctico. En ocasiones podremos expresar la región crítica en términos de un estadístico $T(X_1, \ldots, X_n)$. Por ejemplo, este estadístico puede ser la media de la muestra y la región crítica aquellas muestras con media mayor que un determinado valor.

5.1. Errores de los tests de hipótesis

Consideremos un test de hipótesis con región crítica $\underset{\sim}{X} \in R$. En dicho test podemos cometer dos tipos de errores:

- **Error de tipo 1.** Rechazar la hipótesis nula cuando es cierta. Si $\theta \in \Theta_0$, entonces la probabilidad de cometer un error de tipo 1 es $P_\theta(\underset{\sim}{X} \in R)$.
- **Error de tipo 2.** Rechazar la hipótesis alternativa cuando es cierta. Si $\theta \in \Theta_0^c$, entonces la probabilidad de cometer un error de tipo 2 es $P_\theta(\underset{\sim}{X} \in R^c)$.

Definición 5.3. En el contexto actual, se define la función de potencia del test de hipótesis como

$$\eta(\theta) = P_\theta(\underset{\sim}{X} \in R) = \begin{cases} \text{Probabilidad de cometer un error de tipo 1} & \text{si } \theta \in \Theta_0; \\ \text{1 - Probabilidad de cometer un error de tipo 2} & \text{si } \theta \in \Theta_0^c. \end{cases}$$

Nuestro objetivo es desarrollar tests de hipótesis tales que la probabilidad de cometer errores de tipo 1 y tipo 2 sea lo más pequeña posible para cualquier valor de θ. Esto es, queremos minimizar η en Θ_0 y maximizar η en Θ_0^c. Por tanto, la función de potencia ideal es aquella que toma el valor 0 en Θ_0 y el valor 1 en Θ_0^c. Esta función solo se obtiene en situaciones triviales. Generalmente obtendremos funciones potencia mucho más complejas. Nótese que si reducimos la región crítica de un test entonces disminuye la probabilidad de cometer un error de tipo 1 pero aumenta la probabilidad de cometer un error de tipo 2. Consecuentemente, encontrar una región crítica apropiada no es una tarea sencilla.

Recordemos que la hipótesis nula generalmente tiene preferencia frente a la hipótesis alternativa. Por tanto, el error de tipo 1 es más grave que el error de tipo 2. Para asegurarnos que se respeta esta preferencia podemos buscar tests de hipótesis que garanticen que la probabilidad de cometer un error de tipo 1 es menor que un valor α fijado de antemano.

Definición 5.4. Un test de hipótesis es de tamaño $\alpha \in [0,1]$ si $\sup\{\eta(\theta) : \theta \in \Theta_0\} = \alpha$.

Definición 5.5. Un test de hipótesis tiene nivel de significación $\alpha \in [0,1]$ si $\eta(\theta) \leq \alpha$ para todo $\theta \in \Theta_0$.

Si nos restringimos a estudiar los tests de tamaño α, entonces nuestro objetivo se reduce a buscar entre todos éstos tests aquel que minimice la probabilidad de cometer un error de tipo 2. Este hecho queda formalizado en la siguiente definición.

Definición 5.6. Considérese un test con función potencia η. Decimos que es un test uniformemente más potente de tamaño α (resp. de nivel de significación α) si para cualquier otro test de tamaño α (resp. de nivel de significación α) con función potencia η' se verifica $\eta(\theta) \geq \eta'(\theta)$ para todo $\theta \in \Theta_0^c$. Habitualmente abreviaremos uniformemente más potente por UMP.

En lo que sigue buscaremos obtener los tests UMP de tamaño α en caso de que sea posible. Cabe decir que en la práctica se utilizan valores de α pequeños, como 0,01, 0,05 o 0,1.

5.2. Tests de Neyman-Pearson

En esta sección estudiamos cuáles son los tests UMP de tamaño y significación α para contrastar dos hipótesis simples. Esta cuestión es resuelta por el Lema de Neyman-Pearson. Además, utilizaremos estos resultados para obtener tests UMP que contrasten hipótesis compuestas.

Teorema 5.1 (Lema de Neyman-Pearson)**.** *Supóngase que se desea contrastar dos hipótesis simples,* $H_0 : \theta = \theta_0$ *y* $H_1 : \theta = \theta_1$*. Para* $k \in [0, 1]$ *consideramos la región crítica*

$$C_k = \left\{ \underset{\sim}{x} : \frac{L(\theta_1; \underset{\sim}{x})}{L(\theta_0; \underset{\sim}{x})} \geq k \right\}.$$

Sea $\alpha = P_{\theta_0}(C_k)$*. Entonces, el test de región crítica* C_k *es un test UMP de significación* α*. Además, cualquier test UMP de significación* α *es de tamaño* α *y su región crítica* C' *verifica* $\{\underset{\sim}{x} : L(\theta_1; \underset{\sim}{x}) > kL(\theta_0; \underset{\sim}{x})\} \subset C' \subset C_k$ *salvo a lo sumo un subconjunto de probabilidad nula.*

Demostración. En primer lugar, nótese que el test dado por C_k es de tamaño α ya que $\sup\{\eta(\theta) : \theta \in \Theta_0\} = \eta(\theta_0) = \alpha$. Consideremos otro test de significación α cuya función potencia es η' y su región crítica es C' y veamos que $\eta'(\theta_1) \leq \eta(\theta_1)$. Por distinción de casos es fácil razonar que para cualquier muestra x se verifica

$$(1_{C_k}(\underset{\sim}{x}) - 1_{C'}(\underset{\sim}{x}))(L(\theta_1; \underset{\sim}{x}) - kL(\theta_0; \underset{\sim}{x})) \geq 0, \tag{7}$$

donde 1_A denota la función indicadora del conjunto A. Integrando la desigualdad (7) obtenemos

$$0 \leq \int (1_{C_k}(\underset{\sim}{x}) - 1_{C'}(\underset{\sim}{x}))(L(\theta_1; \underset{\sim}{x}) - kL(\theta_0; \underset{\sim}{x}))\, d\underset{\sim}{x} = \eta(\theta_1) - \eta'(\theta_1) - k(\eta(\theta_0) - \eta'(\theta_0)). \tag{8}$$

Aplicando $\eta(\theta_0) - \eta'(\theta_0) = \alpha - \eta'(\theta_0) \geq 0$ a (8) deducimos que $0 \leq \eta(\theta_1) - \eta'(\theta_1)$ como se quería. Por último, si un test de significación α con función potencia η' y región crítica C' es UMP, entonces $\eta(\theta_1) = \eta'(\theta_1)$ y, por tanto, la desigualdad (8) nos indica que $0 \leq -k(\eta(\theta_0) - \eta'(\theta_0)) \leq 0$, esto es, $\eta'(\theta_0) = \eta(\theta_0) = \alpha$. Puesto que la hipótesis es simple hemos obtenido que el test asociado a C' tiene tamaño α. Consecuentemente, en (8) se da la igualdad, cosa que solo puede suceder si la función no negativa a integrar es constantemente 0 salvo en un conjunto de probabiliad nula. La prueba finaliza al darse cuenta de que este hecho equivale a que $\{x : L(\theta_1; \underset{\sim}{x}) > kL(\theta_0; \underset{\sim}{x})\} \subset C' \subset C_k$ salvo en un subconjunto de probabilidad nula. □

En el Teorema 5.1 el valor de α se determina al seleccionar una región crítica C_k. La cuestión que uno se hace en este punto es si para cualquier $\alpha \in [0, 1]$ existe k_α tal que $\alpha = P_{\theta_0}(C_{k_\alpha})$. En tal caso podríamos encontrar tests más potentes para cualquier tamaño o significación α, que es el valor que nosotros queremos fijar de antemano. La respuesta a esta pregunta es negativa. En efecto, basta con considerar distribuciones discretas ya que para estas distribuciones no podemos asegurar que el valor α sea suma de los valores de la función masa de probabilidad. Por ejemplo, en una distribución binomial $B(x|\theta, n)$ con θ racional no se puede alcanzar el tamaño $\alpha = 1/\pi$, que es irracional. No obstante, este problema desaparece en el caso de las distribuciones continuas como muestra el siguiente corolario del Lema de Neyman-Pearson.

Corolario 5.2 (Lema de Neyman-Pearson para distribuciones continuas)**.** *Sea* $\{f(x|\theta) : \theta \in \Theta\}$ *una familia de distribuciones continuas para la cual se desean contrastar dos hipótesis simples,* $H_0 : \theta = \theta_0$ *y* $H_1 : \theta = \theta_1$*. Supongamos además que las funciones de densidad* $f(x|\theta_0)$ *y* $f(x|\theta_1)$ *coinciden a lo sumo en un subconjunto de medida nula. Entonces, para cada* $\alpha \in [0, 1]$ *existe* $k_\alpha \in [0, 1]$ *tal que el test dado por la región crítica*

$$C_{k_\alpha} = \left\{ \underset{\sim}{x} : \frac{L(\theta_1; \underset{\sim}{x})}{L(\theta_0; \underset{\sim}{x})} \geq k_\alpha \right\}$$

es UMP de tamaño α*.*

Demostración. Gracias al Lema de Neyman-Pearson la prueba se reduce a comprobar que existe k_α tal que $\alpha = P_{\theta_0}(C_k)$. En efecto, la continuidad de las funciones de densidad nos asegura que la función

$f(x|\theta_1)/f(x|\theta_0)$ es continua. Definimos $\varphi : [0,1] \to [0,1]$ dada por

$$\varphi(k) = P_{\theta_0}(\frac{L(\theta_1; \underset{\sim}{X})}{L(\theta_0; \underset{\sim}{X})} \geq k).$$

Esta función es continua y verifica $\varphi(0) = 1$ y $\varphi(1) = 0$. El resultado es consecuencia del teorema del valor intermedio aplicado a φ. □

Los tests proporcionados por el lema de Neyman-Pearson se conocen como tests de Neyman-Pearson. Los estadísticos suficientes son de especial ayuda al aplicar este tipo de tests. En efecto, si T es un estadístico suficiente y $h(t|\theta)$ es su función de densidad, entonces

$$\frac{L(\theta_1; \underset{\sim}{x})}{L(\theta_0; \underset{\sim}{x})} = \frac{h(T(\underset{\sim}{x})|\theta_1)}{h(T(\underset{\sim}{x})|\theta_0)}.$$

Por tanto, podemos calcular la región crítica del test simplemente conociendo el estadístico suficiente y su distribución. Esta observación hace que los cálculos sean más sencillos.

A continuación estudiamos varios ejemplos de los tests de Neyman-Pearson.

Ejemplo 5.3 (Distribución normal de varianza conocida): Consideramos una variable aleatoria $X \sim N(\mu, \sigma^2)$, donde σ^2 es conocido. Vamos a contrastar las hipótesis $H_0 : \mu = \mu_0$ y $H_1 : \mu = \mu_1$ con $\mu_1 > \mu_0$. A priori uno intuye que habrá que rechazar H_0 cuando la media muestral $\overline{x}$ se encuentre más cerca de μ_1 que de μ_0. Veamos que el test de Neyman-Pearson sigue esta intuición.

Recordemos que en el Ejemplo 3.6 se demostró que la media muestral $\overline{X}$ es un estadístico suficiente de cualquier distribución normal de varianza conocida. Además, obtuvimos que $\overline{X} \sim N(\mu, \sigma^2/n)$. Esta observación permite calcular la región crítica del test de Neyman-Pearson con facilidad. En efecto,

$$\frac{L(\mu_1; \underset{\sim}{x})}{L(\mu_0; \underset{\sim}{x})} = \exp\left(-\frac{(\overline{x}-\mu_1)^2 - (\overline{x}-\mu_0)^2}{2\sigma^2/n}\right) = \exp\left(-\frac{\mu_1^2 - \mu_0^2 + 2\overline{x}(\mu_0 - \mu_1)}{2\sigma^2/n}\right).$$

Como consecuencia, obtenemos que $L(\mu_1; \underset{\sim}{x})/L(\mu_0; \underset{\sim}{x}) \geq k$ si, y solo si,

$$\overline{x} \geq \frac{\mu_0 + \mu_1}{2} + \frac{\sigma^2 \log k}{n(\mu_1 - \mu_0)} = A_k.$$

Dado $\alpha \in [0,1]$ buscamos $k_\alpha \in [0,1]$ tal que $P_{\mu_0}(\overline{X} \geq A_{k_\alpha}) = \alpha$. Esta ecuación podemos resolverla para valores concretos de α. Bajo H_0 tenemos $\overline{X} \sim N(\mu_0, \sigma^2/n)$ y, por tanto, sabemos que $A_{k_\alpha} = \mu_0 + z_\alpha \sigma/\sqrt{n}$, donde z_α verifica $F(z_\alpha) = 1 - \alpha$ para la función de distribución F de $N(0,1)$. Los valores z_α que más se utilizan aparecen habitualmente en las tablas de la distribución normal. También se pueden calcular mediante un ordenador.

Por ejemplo, supongamos que $\mu_0 = 0$, $\mu_1 = 1$, $\sigma = 1$ y $\alpha = 0{,}05$. Además, consideremos que se toman muestras de tamaño $n = 100$. En este caso tenemos que $z_\alpha = 1{,}645$. Por tanto, $A_{k_\alpha} = z_\alpha/\sqrt{100} = 0{,}1645$. Esto es, rechazaremos la hipótesis H_0 si $\overline{x} \geq 0{,}1645$.

Habitualmente se denota $Z = \sqrt{n}(\overline{X} - \mu_0)/\sigma$. Bajo H_0 esta variable sigue la distribución $N(0,1)$. El test que hemos obtenido nos dice que rechazaremos H_0 cuando $z = \sqrt{n}(\overline{x} - \mu_0)/\sigma \geq z_\alpha$. Debido al uso del estadístico Z este test se conoce comúnmente como *test Z*. En este caso, $z = 10\overline{x}$. △

5.3. Descripción de un test mediante p-valores

En el Ejemplo 5.3 hemos realizado el test para un determinado valor de α, obteniendo una desigualdad que nos indica si aceptamos o rechazamos la hipótesis nula en función de la muestra. No obstante, algunas muestras tienen más evidencia en su contra que otras. Por tanto, a la hora de realizar un test es conveniente obtener un indicador de la evidencia que tiene una muestra en contra de la hipótesis nula. En este punto entran en juego los p-valores.

Definición 5.7. En un contraste de hipótesis un p-valor $p(\underset{\sim}{X})$ es un estadístico que verifica $0 \leq p(\underset{\sim}{x}) \leq 1$ para cualquier punto $\underset{\sim}{x}$ del espacio muestral. Un p-valor es válido si para cualquier $\theta \in \Theta_0$ y $\alpha \in [0,1]$ se tiene $P_\theta(p(\underset{\sim}{X}) \leq \alpha) \leq \alpha$.

A patir de un p-valor válido $p(\underset{\sim}{X})$ podemos construir un test con nivel de significación α para cualquier $\alpha \in [0,1]$. Este test rechaza H_0 si y solo si $p(\underset{\sim}{X}) \leq \alpha$.

Definición 5.8. Supongamos que un test de hipótesis se puede formular para cualquier nivel de significación α. Sea $p(\underset{\sim}{X})$ un p-valor válido. Decimos que el test es descrito por el p-valor $p(\underset{\sim}{X})$, y solo si, la región crítica del test con significación α es $p(\underset{\sim}{X}) \leq \alpha$.

En un test descrito por un p-valor podemos elegir el valor de α que consideremos apropiado y simplemente compararlo con $p(\underset{\sim}{x})$ para saber cuál es el resultado del test. Nótese que valores pequeños de $p(\underset{\sim}{X})$ dan evidencia de que H_1 es cierta y, por tanto, el p-valor mide la evidencia a favor de la hipótesis nula. Por tanto, realizar un test descrito por un p-valor es más informativo que simplemente elegir entre "aceptar H_0" o "rechazar H_0".

El siguiente resultado nos informa sobre cómo construir un p-valor válido.

Teorema 5.4 ([2, Teorema 8.3.27])**.** *Consideramos el contraste de hipótesis dado por $H_0 : \theta \in \Theta_0$ y $H_1 : \theta \in \Theta_1$. Sea $W(\underset{\sim}{X})$ un estadístico. Entonces, la función*

$$p(\underset{\sim}{x}) = \sup\{P_\theta(W(\underset{\sim}{X}) \geq W(\underset{\sim}{x})) : \theta \in \Theta_0\}$$

es un p-valor válido.

La pregunta que uno se hace en este punto es, dado un test que se puede formular para cualquier valor de α, cómo obtener un p-valor que lo describa. Esto no va a ser siempre factible ya que habrá regiones críticas que no sepamos o no se puedan escribir de la forma $p(\underset{\sim}{X}) \leq \alpha$. No obstante, en múltiples casos prácticos sí es posible como muestra el siguiente resultado.

Teorema 5.5. *Sea $W(\underset{\sim}{X})$ un estadístico. Supongamos que para cada $\alpha \in [0,1]$ disponemos de un test de hipótesis de tamaño α cuya región crítica es $W(\underset{\sim}{X}) \geq k_\alpha$ para cierta constante k_α. Entonces, el test se puede describir a partir del p-valor*

$$p(\underset{\sim}{x}) = \sup\{P_\theta\left(W(\underset{\sim}{X}) \geq W(\underset{\sim}{x})\right) : \theta \in \Theta_0\}.$$

Demostración. Nótese que $p(\underset{\sim}{x})$ es un p-valor gracias al Teorema 5.4. Sea $\alpha \in [0,1]$ y sea k_α como en el enunciado. Tenemos que $p(\underset{\sim}{x}) \leq \alpha = \sup\{P_\theta\left(W(\underset{\sim}{X}) \geq k_\alpha\right) : \theta \in \Theta_0\}$ si, y solo si, para cada $\theta \in \Theta_0$ se tiene

$$P_\theta\left(W(\underset{\sim}{X}) \geq W(\underset{\sim}{x})\right) \leq P_\theta\left(W(\underset{\sim}{X}) \geq k_\alpha)\right),$$

esto es, $W(\underset{\sim}{x}) \geq k_\alpha$. □

Corolario 5.6. *Consideremos un test de Neyman-Pearson con función de densidad continua. Entonces, el test se puede describir a partir del p-valor*

$$p(\underset{\sim}{x}) = P_{\theta_0}\left(\frac{L(\theta_1;\underset{\sim}{X})}{L(\theta_0;\underset{\sim}{X})} \geq \frac{L(\theta_1;\underset{\sim}{x})}{L(\theta_0;\underset{\sim}{x})}\right).$$

El p-valor que hemos construido en el Teorema 5.5 otorga a una muestra $\underset{\sim}{x}$ evidencia contra la hipótesis nula cuando el valor de $W(\underset{\sim}{x})$ es muy alto. El valor obtenido al aplicar este p-valor a una muestra puede interpretarse como la probabilidad de observar una muestra que sea tan poco favorable a la hipótesis nula como ella. Esto es lo que sucede en el Corolario 5.6, donde el p-valor es efectivamente una probabilidad. Algunos usuarios de los tests de hipótesis interpretan el p-valor como la probabilidad de que la hipótesis nula sea falsa para la muestra obtenida. Esto es erróneo y en ningún momento deberíamos utilizar esta interpretación. En la Sección 6 estudiaremos tests de hipótesis desde el punto de vista bayesiano. En estos tests sí obtendremos probabilidades de que la hipótesis nula sea cierta o no.

Ejemplo 5.7 (Continuación del Ejemplo 5.3): Recordemos que obtuvimos que la región crítica del test Z está determinada por $\overline{X} \geq A_{k_\alpha}$. Por tanto, el Teorema 5.5 nos dice que el p-valor del test es $p(\underset{\sim}{x}) = P_{\mu_0}(\overline{X} \geq \overline{x})$. También probamos que podemos describir la región crítica en términos de la variable Z. De hecho, $P_{\mu_0}(\overline{X} \geq \overline{x}) = P_{\mu_0}(Z \geq \sqrt{n}(\overline{x} - \mu_0)/\sigma)$, lo que nos permite calcular los p-valores gracias a que $Z \sim N(0,1)$.

Veamos un ejemplo numérico. Los datos osn los mismos que los del Ejemplo 5.3. Supongamos que tras observar una muestra con $n = 100$ hemos obtenido $\overline{x} = 0{,}2$. Entonces, $p(\underset{\sim}{x}) = P_{\mu_0}(Z \geq 2) = 1 - F(2) = 1 - 0{,}9772 = 0{,}0328$, donde F es la función de distribución de N(0,1). Por tanto, rechazaremos la hipótesis nula para cualquier nivel de significación mayor que 0,0328. △

5.4. Tests de la razón de verosimilitud

En esta sección estudiamos los tests de la razón de la verosimilitud, que están íntimamente ligados con los estimadores máximo verosímiles. Además, veremos que en el caso de hipótesis simples estos tests coinciden con el test de Neyman-Pearson.

Definición 5.9. Considérese una hipótesis $H : \theta \in \Theta_0$. El ratio de verosimilitud de H para la muestra $\underset{\sim}{x}$ es

$$\lambda(\underset{\sim}{x}) = \frac{\sup\{L(\theta;\underset{\sim}{x}) : \theta \in \Theta_0\}}{\sup\{L(\theta;\underset{\sim}{x}) : \theta \in \Theta\}}.$$

Fijado un contraste de hipótesis, $H_0 : \theta \in \Theta_0$ y $H_1 : \theta \in \Theta_1$, un test de la razón de verosimilitud es cualquier test cuya región crítica sea de la forma $\{\underset{\sim}{x} : \lambda(\underset{\sim}{x}) \leq c\}$, donde $0 \leq c \leq 1$.

Para motivar la definición de este tipo de test, supongamos que estamos trabajando con distribuciones discretas. En tal caso, tanto el numerador como el denominador de $\lambda(\underset{\sim}{x})$ se corresponden con la máxima probabilidad posible de la muestra $\underset{\sim}{x}$ si variamos θ en Θ_0 y Θ respectivamente. Si el cociente de ambos valores es pequeño ($\lambda(\underset{\sim}{x}) \leq c$), entonces es razonable rechazar la hipótesis nula puesto que hay elementos de Θ_0^c para los cuales la muestra es más probable.

Supongamos ahora que existen los estimadores máximo verosímiles de θ_0 en Θ_0 y Θ. Denotamos a estos estimadores $\hat{\theta_0}(x)$ y $\hat{\theta}(x)$ respectivamente. Entonces, $\lambda(\underset{\sim}{x}) = L(\hat{\theta_0}(x);\underset{\sim}{x})/L(\hat{\theta}(x);\underset{\sim}{x})$. El test de la razón de verosimilitud nos dice que rechazaremos la hipótesis nula cuando $\hat{\theta}(x)$ tenga una credibilidad considerablemente mayor que la de $\hat{\theta_0}(x)$, esto es, $\hat{\theta}(x)$ es mucho mejor estimador. Como caso particular,

si a partir de una muestra x hemos realizado una estimación de θ mediante un estimador máximo verosímil $\hat{\theta}$, entonces todo test de la razón de verosimilitud acepta la hipótesis $H_0 : \theta = \hat{\theta}$ para la muestra x. Esto es, la filosofía de los tests de la razón de verosimilitud es coherente con la filosofía de los estimadores máximo verosímiles.

Ejemplo 5.8 (**Test t de Student** – distribución normal con μ y σ^2 desconocidos): Consideramos una variable $X \sim N(\mu, \sigma^2)$ con μ y σ^2 desconocidos y buscamos contrastar las hipótesis $H_0 : \mu = \mu_0$ y $H_1 : \mu \neq \mu_0$ aplicando el test de la razón de verosimilitud. Esto es, el parámetro es $\theta = (\mu, \sigma^2)$, $\Theta = \mathbb{R} \times \mathbb{R}^+$ y $\Theta_0 = \{0\} \times \mathbb{R}^+$. Este test se denomina test t de Student.

Tenemos que calcular $\lambda(\underset{\sim}{x})$ para cualquier muestra $\underset{\sim}{x}$. Recordemos que en el Ejemplo 3.7 se demostró que la función de verosimilitud en este contexto responde a

$$L(\mu, \sigma; \underset{\sim}{x}) = (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) = (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-1}{2\sigma^2}((n-1)S^2 + n(\overline{x} - \mu))\right),$$

donde S^2 es la varianza muestral. En primer lugar calculamos

$$\lambda_0(\underset{\sim}{x}) = \sup\{L(\theta; \underset{\sim}{x}) : \theta \in \Theta_0\} = \sup\{L(\mu_0, \sigma^2; \underset{\sim}{x}) : \sigma^2 \in \mathbb{R}^+\}.$$

Es fácil ver que el estimador máximo verosímil de la varianza de una normal con media conocida μ_0 es $\hat{\sigma}_0^2(\underset{\sim}{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2$. Por tanto, obtenemos

$$\lambda_0(\underset{\sim}{x}) = L(\mu_0, \hat{\sigma}_0^2; \underset{\sim}{x}) = (2\pi\hat{\sigma}_0^2)^{-n/2} \exp(-n/2).$$

Por otro lado, en el caso de una normal con media y varianza desconocida el estimador máximo verosímil $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$ viene dado por $\hat{\mu}(\underset{\sim}{x}) = \overline{x}$ y $\hat{\sigma}^2(\underset{\sim}{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = (n-1)S^2/n$. Por tanto, obtenemos

$$\lambda_1(\underset{\sim}{x}) = \sup\{L(\theta; \underset{\sim}{x}) : \theta \in \Theta\} = (2\pi\hat{\sigma}^2)^{-n/2} \exp(-n/2).$$

El ratio de verosimilitud de H_0 es

$$\begin{aligned}\lambda(\underset{\sim}{x}) = \frac{\lambda_0(\underset{\sim}{x})}{\lambda_1(\underset{\sim}{x})} = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right)^{-n/2} = \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2}{(n-1)S^2/n}\right)^{-n/2} = \\ \left(\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - \mu_0)^2}{(n-1)S^2}\right)^{-n/2} = \left(1 + \frac{t^2}{n-1}\right)^{-n/2},\end{aligned} \tag{9}$$

donde $t = \sqrt{n}(\overline{x} - \mu_0)/S$. Recordemos en este punto que bajo H_0 la variable $T = \sqrt{n}(\overline{X} - \mu_0)/S$ sigue una distribución t de Student con $n-1$ grados de libertad. A partir de (9) nótese que λ) es decreciente respecto de $|t|$. Por tanto, el test de la razón de verosimilitud tiene como región crítica

$$\frac{\sqrt{n}(\overline{X} - \mu_0)}{S} \geq b$$

para cierta constante $b \in [0, 1]$. La región crítica no depende de σ^2. Por tanto, tenemos que $\eta(\theta) = P_{\mu_0}(T > b)$ para todo $\theta \in \Theta_0$. Consecuentemente el tamaño del test es $\sup\{\eta(\theta) : \theta \in \Theta_0\} = P_{\mu_0}(T \geq b)$. En la práctica determinamos b de manera que el tamaño del test sea igual a α. Para ello utilizamos la función de distribución de la distribución t de Student con $n-1$ grados de libertad. Por último, cabe hablar del p-valor del test t de Student. Utilzando el Teorema 5.5 obtenmos que el p-valor viene dado por

$$p(\underset{\sim}{x}) = P_{\mu_0}\left(\frac{\sqrt{n}(\overline{X} - \mu_0)}{S} \geq t\right). \qquad \triangle$$

6. Estadística bayesiana

En esta sección estudiaremos el modelo de inferencia estadística desde el punto de vista bayesiano. La estadística bayesiana proporciona resultados similares a la estadística clásica en el problema de estimación. Las ventajas de los modelos bayesianos serán remarcables al realizar tests de hipótesis, donde nos permitirán hablar de la probabilidad de que una hipótesis sea cierta o no. Recordemos que esta afirmación es imposible en la inferencia clásica ya que los parámetros de un modelo no son variables aleatorias.

6.1. Introducción

En primer lugar recordamos uno de los teoremas clásicos de la probabilidad, el Teorema de Bayes, que es la base de la inferencia Bayesiana. Supongamos que en el espacio de probabilidad $(\Omega, \mathcal{A}, P)$ tenemos una partición de Ω dada por los sucesos $A_1, \ldots, A_n$, todos ellos con probabilidad no nula. Sea B un suceso no nulo del que conocemos sus probabilidades condicionadas a cada suceso A_i. Entonces, utilizando la definición de probabilidad condicionada obtenemos la probabilidad de cada A_i condicionada al suceso B como sigue

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{P(B)}. \tag{10}$$

A su vez, la ley de la probabilidad total establece que $P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$ y, por tanto, aplicando esta a igualdad a (10) deducimos

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}, \tag{11}$$

donde los valores $P(B|A_i)$ eran conocidos. La ecuación (11) se conoce como Teorema de Bayes. A los valores $P(A_i)$ los llamamos *probabilidades a priori*, mientras que a los valores $P(A_i|B)$ los denominamos las *probabilidades a posteriori*. El Teorema de Bayes se puede deducir de igual forma para distribuciones de probabilidad.

La estadística bayesiana se basa en la interpretación subjetiva de la probabilidad. Utiliza la percepción existente por parte del investigador para otorgar una credibilidad a cada parámetro del modelo en forma de una distribución de probabilidad (distribución a priori). Posteriormente aplica el Teorema de Bayes para obtener una distribución de los parámetros condicionada a la muestra (distribución a posteriori), con la que formular inferencias con respecto al parámetro de interés.

Consideremos un problema de inferencia estadística en el que las observaciones se toman de una variable aleatoria X que sigue una distribución $f(x|\theta)$, con $\theta \in \Theta$. Disponemos de información previa sobre θ, que podemos recoger definiendo una distribución de probabilidad sobre el espacio Θ, la distribución a priori, dando así a θ el carácter de variable aleatoria, con la peculiaridad de que no es observable. Sin embargo, sí observamos la variable aleatoria X condicionada al verdadero valor que toma θ, que llamaremos θ_0. Así, el estudio de las observaciones de X aporta información sobre el valor de θ, información que debemos combinar con la distribución a priori para modificarla. El resultado de esta modificación es de nuevo una distribución sobre Θ, que llamaremos la distribución a posteriori de θ, una vez observada la variable aleatoria X. Estos son los planteamientos básicos que conforman el enfoque bayesiano de la estadística.

En lo que sigue definiremos formalmente la distribución a posteriori, recordando los conceptos que sean necesarios.

Definición 6.1. Sea X una variable aleatoria que sigue una distribución $f(x|\theta)$, con $\theta \in \Theta$. A una distribución $\pi(\theta)$ sobre el espacio Θ establecida con información previa conocida sobre θ se le llama distribución a priori de la variable aleatoria θ.

Comentario 6.1. Dada una muestra aleatoria simple $\underset{\sim}{X} = (X_1, \ldots, X_n)$ de X y una distribución a priori $\pi(\theta)$, podemos calcular la distribución conjunta de $\underset{\sim}{X}$ y θ utilizando la definición de condicionamiento [1], que es el análogo a (10) para variables aleatorias. En efecto, en este caso obtenemos que la ditribución conjunta tiene la función de densidad

$$f(\underset{\sim}{x}, \theta) = f(\underset{\sim}{x}|\theta)\pi(\theta).$$

A partir de la distribución conjunta podemos calcular la distribución marginal de $\underset{\sim}{X}$, que denotamos $m(\underset{\sim}{X})$. Esta distribución tiene función de densidad

$$m(\underset{\sim}{x}) = \int_\Theta f(\underset{\sim}{x}, \theta)d\theta = \int_\Theta f(\underset{\sim}{x}|\theta)\pi(\theta)d\theta.$$

Definición 6.2. Dada una muestra aleatoria simple $\underset{\sim}{X} = (X_1, \ldots, X_n)$ de X, una realización de la muestra $\underset{\sim}{x} = (x_1, \ldots, x_n)$ y una distribución a priori $\pi(\theta)$, se define la distribución a posteriori de θ como la ley de θ condicionada a $\underset{\sim}{X} = \underset{\sim}{x}$. La denotamos $\pi(\theta|\underset{\sim}{x})$.

La función de densidad de la distribución a posteriori se puede calcular a partir de la distribución a priori y $f(\underset{\sim}{x}|\theta)$ utilizando de nuevo la definición de condicionamiento. En efecto, tenemos que

$$\pi(\theta|\underset{\sim}{x})m(\underset{\sim}{x}) = f(\underset{\sim}{x}, \theta) = f(\underset{\sim}{x}|\theta)\pi(\theta).$$

En consecuencia, podemos expresar la distribución a posteriori de esta forma

$$\pi(\theta|\underset{\sim}{x}) = \frac{f(\underset{\sim}{x}, \theta)}{m(\underset{\sim}{x})}.$$

A veces es conveniente omitir el uso de la distribución marginal $m(\underset{\sim}{x})$, en cuyo caso escribimos

$$\pi(\theta|\underset{\sim}{x}) = \frac{f(\underset{\sim}{x}|\theta)\pi(\theta)}{\int_\Theta f(\underset{\sim}{x}|\theta)\pi(\theta)\,d\theta}.$$

Una vez hemos calculado la distribución a posteriori los problemas de la inferencia clásica se vuelven muy sencillos. En la Sección 7 se estudian los tests de hipótesis desde una perspectiva bayesiana. En el caso de la estimación las complicaciones son incluso menores. Podemos estimar el parámetro θ como la moda de la distribución $\pi(\theta|\underset{\sim}{x})$, siguiendo la filosofía de los estimadores máximo verosímiles. Si θ fuese un valor real, entonces podemos realizar su estimación mediante una esperanza, esto es, proponemos como estimador $\hat{\theta}(\underset{\sim}{x}) = E_{\pi(\theta|\underset{\sim}{x})}\theta$. En la mayoría de las aplicaciones ambos estimadores presentan un comportamiento similar al de los estimadores máximo verosímiles. Estudiaremos esta similitud a lo largo de la multitud de ejemplos que se realizan en esta sección. Es más, en la Sección 6.5 demostraremos que, bajo determinadas condiciones, la distribución a posteriori degenera en el verdadero valor del parámetro θ cuando el tamaño de la muestra diverge. Esta propiedad es análoga a la propiedad de convergencia del estimador máximo verosímil. No obstante, la principal ventaja de los estimadores bayesianos es su buen comportamiento para muestras pequeñas, sobre las que los estimadores máximo verosímiles podían no funcionar correctamente debido a la falta de información. Este buen comportamiento se debe

a la información introducida por la distribución a priori, que permite decidir el valor del parámetro en el caso de que la muestra no sea relevante.

Para finalizar la introducción, cabe destacar es que es posible no exigirle a la distribución de probabilidad a priori que integre, es decir, podrían distribuir una probabilidad infinita sobre Θ. En tal caso se dice que la distribución es *impropia*. Pese a su carácter impropio estas distribuciones nos pueden permitir hacer inferencias correctas.

6.2. Estadística clásica vs bayesiana

Veamos ahora las diferencias entre la inferencia clásica y la bayesiana. En la inferencia clásica destacan las siguientes características:

- El concepto de probabilidad está limitado a aquellos sucesos en los que se pueden definir frecuencias relativas.
- θ es un valor fijo, pero desconocido.
- Se usa el concepto de intervalo de confianza.
- El método de muestreo es muy importante.
- Se pueden usar estimadores de máxima verosimilitud o estimadores insesgados.
- Los tests de hipótesis se construyen fijando un tamaño α y minimizando los errores de tipo 2.

Por su parte, en la inferencia bayesiana destacan:

- Podemos establecer probabilidades previas para cualquier suceso.
- θ es una variable que sigue una distribución de probabilidad.
- Se usa el concepto de intervalo de credibilidad para θ.
- El método de muestreo no importa; solo importan los datos.
- Se utilizan estimadores diferentes según la utilidad; la estimación es un problema de decisión.
- En tests de hipótesis se puede hablar de probabilidad de que una hipótesis sea cierta.

Uno de los aspectos más criticados de la estadística bayesiana es el grado de subjetividad a la que se expone la inferencia por el hecho de que es el experimentador quien define la distribución a priori. En cualquier caso, en lo que hay coincidencia es en que si hay información sobre θ, entonces ésta tiene que ser utilizada en la inferencia.

Como acabamos de decir, una parte muy importante en la inferencia bayesiana es la selección de la distribución a priori. En muchos casos, si no disponemos de una distribución clara para modelar θ es posible considerar distribuciones específicas que permitan simplificar los cálculos de la distribución a posteriori. A continuación estudiaremos distintos medios para seleccionar estas distribuciones.

6.3. Familias conjugadas

La principal dificultad que surge en los problemas de inferencia bajo la perspectiva bayesiana es tanto la confianza que se pueda esperar de la distribución a priori como el cálculo de la distribución a posteriori.

La primera cuestión es importante ya que la inferencia que se realice puede depender de la elección de la distribución inicial, razón por la cual en muchos casos se recurre a distribuciones no informativas, que no imponen unas condiciones muy fuertes sobre el parámetro. Otra tendencia en la elección de las distribuciones a priori es aprovechar la información que proporciona la muestra para mejorar la distribución inicial.

En cuanto al cálculo de la distribución a posteriori, no todas las distribuciones a priori conducen a cómputos asequibles ni a una distribución tratable y, en ocasiones, hay que recurrir a métodos numéricos para poder trabajar con ellas. Por tanto, es deseable obtener distribuciones a priori que nos faciliten este proceso.

En esta sección nos centramos en este último problema. Buscamos familias de distribuciones a priori cuyas distribuciones a posteriori asociadas sean de fácil cálculo. En este sentido surge el concepto de familias a priori conjugadas.

Definición 6.3. Sea $\mathcal{F} = \{\pi_i(\theta) : i \in I\}$ una familia de distribuciones a priori. Se dice que $\mathcal{F}$ es conjugada respecto de la familia de densidades $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$ si para cualquier $\pi(\theta) \in \mathcal{F}$ y $f(x|\theta) \in \mathcal{P}$ se verifica que $\pi(\theta|x) \in \mathcal{F}$. Es decir, una familia $\mathcal{F}$ de distribuciones a priori es conjugada respecto a la familia dada si y solo si las distribuciones a posteriori pertenecen de nuevo a $\mathcal{F}$.

Recordemos que $\pi(\theta|x) = f(x|\theta)\pi(\theta)/m(x)$. El denominador $m(x)$ es una constante ya que x está fijo. Como con secuencia, obtenemos la siguiente observación.

Comentario 6.2. Sean $\pi(\theta), \Pi(\theta) \in \mathcal{F}$. Se tienen las siguientes condiciones equivalentes:

a) $f(x|\theta)\pi(\theta) \propto \Pi(\theta)$;

b) $\pi(\theta|x) = \Pi(\theta)$.

Por tanto, en la definición de familias conjugadas, la afirmación $\pi(\theta|x) \in \mathcal{F}$ equivale a decir que $f(x|\theta)\pi(\theta) \propto \Pi(\theta)$ para cierta distribución $\Pi(\theta) \in \mathcal{F}$. Este hecho se enuncia en la siguiente proposición.

Proposición 6.3. *Una familia de distribuciones a priori $\mathcal{F} = \{\pi_i(\theta) : i \in I\}$ es conjugada respecto de la familia de densidades $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$ si, y solo si, el producto de cualesquiera dos distribuciones de ambas familias vuelve a ser, salvo constante, una distribución de la familia de distribuciones a priori.*

Tener una familia de distribuciones conjugadas a priori nos permite simplificar en gran medida el cálculo de la distribución a posteriori. En efecto, solo tenemos que identificar la distribución de $\mathcal{F}$ que es proporcional a $f(x|\theta)\pi(\theta)$. Esa distribución coincidirá con $\pi(\theta|x)$. De esta forma evitamos tener que realizar el cálculo de la distribución marginal de x, que suele hacerse mediante una integral. Además, en caso de necesitar el valor $m(x)$ basta darse cuenta de que $m(x) = f(x|\theta)\pi(\theta)/\pi(\theta|x)$.

Es posible calcular las distribuciones conjugadas para las familias de distribuciones clásicas, obteniendo de nuevo otras distribuciones clásicas.

Proposición 6.4.

- *La familia de distribuciones Beta es una familia de distribuciones conjugada para las distribuciones de Bernouilli, binomiales y binomiales negativas.*
- *La familia de distribuciones Gamma es una familia de distribuciones conjugada para las distribuciones de Poisson.*
- *La familia de distribuciones normales es una familia de distribuciones conjugada para la familia de distribuciones normales con varianza conocida.*

- *La familia de distribuciones de Dirichlet es una familia de distribuciones conjugada para la familia de distribuciones multinomiales.*

Veamos algún ejemplo de los proporcionados por la proposición anterior.

Ejemplo 6.5: Vamos a considerar una distribución de Poisson de parámetro $\lambda > 0$ y, como distribución a priori, una Gamma de parámetros α y β. Dada una muestra $\underset{\sim}{x} = (x_1, \ldots, x_n)$ se tiene

$$f(\underset{\sim}{x}|\lambda) = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!} \text{ y } \pi(\lambda) = \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}.$$

Multiplicando ambas distribuciones obtenemos

$$f(\underset{\sim}{x}|\lambda)\pi(\lambda) = \frac{e^{-n\lambda}\lambda^{\sum x_i}\lambda^{\alpha-1}e^{-\beta\lambda}\beta^\alpha}{\prod x_i!\Gamma(\alpha)}$$

$$= \frac{\beta^\alpha}{\prod x_i!\Gamma(\alpha)}\lambda^{\sum x_i+\alpha-1}e^{-(\beta+n)} \propto Gamma\left(\lambda|\alpha+\sum x_i, \beta+n\right).$$

Es decir, para variables aleatorias de Poisson de parámetro λ, escogiendo una distribución a priori Gamma de parámetros α y β obtenemos como distribución de λ a posteriori una nueva Gamma, esta vez de parámetros $\alpha + \sum_{i=1}^n x_i$ y $\beta + n$, donde n es el tamaño de la muestra.

Notemos que la conjugación nos ha permitido evitar el cálculo de la distribución marginal de x. Si optamos por calcularla, obtendríamos:

[PROXIMAMENTE]

Llegando de nuevo al mismo resultado. △

Ejemplo 6.6: Consideremos ahora una distribución multinomial dada por la función masa de probabilidad

$$f(x_1, \ldots, x_k|\theta_1, \ldots, \theta_k) = \frac{n!}{x_1! \ldots x_k!}\theta_1^{x_1} \ldots \theta_k^{x_k},$$

con $\sum_{i=1}^k x_i = n$, $0 < \theta_i < 1$ y $\sum_{i=1}^k \theta_i = 1$.

Llamamos $\theta = (\theta_1, \ldots, \theta_k)$. Elegimos como distribución a priori la distribución de Dirichlet de parámetros $\alpha_1, \ldots, \alpha_k$, cuya función de densidad viene dada por

$$\pi(\theta) \propto \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} \prod_{i=1}^k \theta_i^{\alpha_i - 1}.$$

Entonces, dada una muestra $\underset{\sim}{x} = (x_1, \ldots, x_k)$, se tiene

$$\pi(\theta|\underset{\sim}{x}) \propto \pi(\theta) f(\underset{\sim}{x}|\theta) \propto \left(\prod_{i=1}^k \theta_i^{\alpha_i - 1}\right)\left(\prod_{i=1}^k \theta_i^{x_i}\right)$$

$$= \prod_{i=1}^k \theta_i^{x_i+\alpha_i-1} \propto Dirichlet(\theta|\alpha_1 + x_1, \ldots, \alpha_k + x_k),$$

obteniendo que la distribución a posteriori sigue una Dirichlet. △

Ejemplo 6.7: Consideremos $X \sim \mathcal{N}(\mu, \sigma^2)$, con σ^2 conocido. Fijamos una muestra $\underset{\sim}{x} = (x_1, \ldots, x_n)$. La función de densidad está condicionada únicamente a μ. Tenemos que

$$f(\underset{\sim}{x}|\mu) = (2\pi\sigma^2)^{-n/2} \exp\left(-\sum \frac{(x_i-\mu)^2}{2\sigma^2}\right).$$

Elegimos una distribución a priori $\mu \sim \mathcal{N}(\eta, \tau^2)$, es decir,

$$\pi(\mu) = (2\pi\tau^2)^{-1/2} \exp\left(-\frac{(\mu-\eta)^2}{2\tau^2}\right).$$

Calculamos la distribución conjunta, obteniendo

$$f(\underset{\sim}{x}|\mu)\pi(\mu) \propto \exp\left(-\sum \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{(\mu-\eta)^2}{2\tau^2}\right).$$

Llamamos $\overline{x} = \frac{1}{n}\sum x_i$ y $s^2 = \frac{1}{n}\sum (xb_i - \overline{x})^2$. Notemos que ninguna de las dos expresiones depende de μ. Usando que $\sum (x_i-\mu)^2 = \sum (x_i - \overline{x} + \overline{x} - \mu)^2 = \sum (x_i-\overline{x})^2 + 2(\overline{x}-\mu)\sum (x_i - \overline{x}) + \sum (\overline{x}-\mu)^2 = ns^2 + n(\overline{x}-\mu)^2$, tenemos

$$f(\underset{\sim}{x}|\mu)\pi(\mu) \propto \exp\left(-\frac{n}{2\sigma^2}[s^2 + (\overline{x}-\mu)^2] - \frac{(\mu-\eta)^2}{2\tau^2}\right)$$

$$= \exp\left(-\frac{ns^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2\sigma^2\tau^2}[n\tau^2(\overline{x}-\mu)^2 + \sigma^2(\mu-\eta)^2]\right) \propto \exp\left(-\frac{1}{2\sigma^2\tau^2}[n\tau^2(\overline{x}-\mu)^2 + \sigma^2(\mu-\eta)^2]\right).$$

Ahora, desarrollamos la expresión $n\tau^2(\overline{x}-\mu)^2 + \sigma^2(\mu-\eta)^2$. Podemos separarla según los sumandos que dependan o no de μ. A continuación, dividimos la exponencial en dos partes, una con los sumandos independientes de μ y otra con los dependientes. La expresión resultante es

$$f(\underset{\sim}{x}|\mu)\pi(\mu) \propto \exp\left(-\frac{n\overline{x}^2\tau^2 + \sigma^2\eta^2}{2\sigma^2\tau^2}\right) \exp\left(-\frac{\mu^2(n\tau^2+\sigma^2) - 2\mu(n\overline{x}\tau^2 + \sigma^2\eta)}{2\sigma^2\tau^2}\right).$$

$$\propto \exp\left(-\frac{1}{2\sigma^2\tau^2}\left[\mu^2(n\tau^2+\sigma^2) - 2\mu(n\overline{x}\tau^2+\sigma^2\eta)\right]\right).$$

Ahora ajustamos cuadrados sobre el exponente. Procedemos como sigue

$$-\frac{1}{2\sigma^2\tau^2}\left[\mu^2(n\tau^2+\sigma^2) - 2\mu(n\overline{x}\tau^2+\sigma^2\eta)\right] = -\frac{n\tau^2+\sigma^2}{2\sigma^2\tau^2}\left[\mu^2 - 2\mu\frac{n\overline{x}\tau^2+\sigma^2\eta}{n\tau^2+\sigma^2}\right]$$

$$= -\frac{n\tau^2+\sigma^2}{2\sigma^2\tau^2}\left[\mu^2 - 2\mu\frac{n\overline{x}\tau^2+\sigma^2\eta}{n\tau^2+\sigma^2} + \left(\frac{n\overline{x}\tau^2+\sigma^2\eta}{n\tau^2+\sigma^2}\right)^2\right] + \frac{n\tau^2+\sigma^2}{2\sigma^2\tau^2}\left(\frac{n\overline{x}\tau^2+\sigma^2\eta}{n\tau^2+\sigma^2}\right)^2$$

$$= -\frac{n\tau^2+\sigma^2}{2\sigma^2\tau^2}\left[\mu - \frac{n\overline{x}\tau^2+\sigma^2\eta}{n\tau^2+\sigma^2}\right]^2 + \frac{\left(n\overline{x}\tau^2+\sigma^2\eta\right)^2}{2\sigma^2\tau^2\left(n\tau^2+\sigma^2\right)}.$$

Volviendo a la distribución conjunta, hemos obtenido

$$f(\underset{\sim}{x}|\mu)\pi(\mu) \propto \exp\left(\frac{\left(n\overline{x}\tau^2+\sigma^2\eta\right)^2}{2\sigma^2\tau^2\left(n\tau^2+\sigma^2\right)}\right) \exp\left(-\frac{n\tau^2+\sigma^2}{2\sigma^2\tau^2}\left[\mu - \frac{n\overline{x}\tau^2+\sigma^2\eta}{n\tau^2+\sigma^2}\right]^2\right)$$

$$\propto \exp\left(-\frac{n\tau^2+\sigma^2}{2\sigma^2\tau^2}\left[\mu-\frac{n\overline{x}\tau^2+\sigma^2\eta}{n\tau^2+\sigma^2}\right]^2\right) = \exp\left(-\left[\mu-\frac{n\overline{x}\tau^2+\sigma^2\eta}{n\tau^2+\sigma^2}\right]^2 / \left[2\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}\right]\right)$$
$$\propto \mathcal{N}\left(\mu \,\middle|\, \frac{n\overline{x}\tau^2+\sigma^2\eta}{n\tau^2+\sigma^2}, \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}\right),$$

obteniendo finalmente una distribución a posteriori también normal. Hemos demostrado que la familia normal es conjugada. △

6.4. Distribuciones objetivas. Distribución de Jeffreys

Aunque las distribuciones conjugadas permiten facilitarnos los cálculos, no siempre tienen un comportamiento adecuado de cara a la inferencia. En algunas aplicaciones el uso de las distribuciones a priori conjugadas puede estar introduciendo una forma a $\pi(\theta)$ que no se adecúe a la realidad. Además, como hemos visto en los ejemplos anteriores, las distribuciones conjugadas tienen parámetros que de todas formas deberían ser seleccionados utilizando el conocimiento experto del problema a resolver. En condiciones en las que no se conozca información alguna sobre el parámetro θ estas propiedades pueden ser perjudiciales.

En esta sección estudiamos distribuciones a priori objetivas o no informativas, que no introducen ninguna información sobre el parámetro θ. De estas distribuciones una de las más famosas es la distribución de Jeffreys.

Definición 6.4. Sea $\{f(x|\theta) : \theta \in \Theta\}$ una familia de distribuciones con parámetro $\theta \in \Theta$. La distribución a priori de Jeffreys se define como $\pi^J(\theta) \propto \sqrt{\mathcal{I}_X(\theta)}$.

Ejemplo 6.8 (Distribución binomial): Vamos a estudiar la distribución a priori de Jeffreys para la distribución binomial. En el Ejemplo 3.11 se calculó la función de información de Fisher para la distribución binomial. A partir de los resultados obtenidos tenemos que $\pi^J(\theta) \propto \theta^{-1/2}(1-\theta)\theta^{-1/2}$. Por tanto, $\pi^J(\theta)$ sigue una distribución $beta(1/2, 1/2)$. La distribución a posteriori para $\underset{\sim}{x} = (x_1, \ldots, x_k)$ viene dada por

$$\pi(\theta; x) \propto \pi(\theta)\prod_{i=1}^{k} f(x_i;\theta) \propto \theta^{\sum x_i - 1/2}(1-\theta)^{\sum(n-x_i)-1/2},$$

esto es, $\pi(\theta; x)$ sigue una distribución $beta(k\overline{x}+1/2, k(n-\overline{x})+1/2)$. Recordando el Corolario **??** podemos calcular la esperanza y la varianza de la distribución a posteriori, obteniendo

$$E[\pi(\theta; x)] = \frac{k\overline{x}+1/2}{kn+1} = \frac{\overline{x}+1/(2k)}{n+1/k};$$

$$Var(\pi(\theta; x)) = \frac{(k\overline{x}+1/2)(k(n-\overline{x})+1/2)}{(kn+1)^2(kn+2)} = \frac{(\overline{x}+1/2k)((n-\overline{x})+1/(2k))}{(kn+2)(n+1/k)^2}.$$

Para k lo suficientemente grande $E[\pi(\theta; x)] \approx \overline{x}/n$, que es el estimador máximo verosímil. Por tanto, cuando $k \to \infty$ obtenemos que $Var(\pi(\theta; x)) \to 0$ y $E[\pi(\theta; x)] \to \theta_0$. △

Notemos que la distribución de Jeffreys podría ser impropia, es decir, no integrable. En cualquier caso se puede utilizar para realizar estimación. Veámoslo con el siguiente ejemplo.

Ejemplo 6.9: Consideramos la familia de distribuciones de Poisson con parámetro λ, con funciones de densidad $f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$, con $\lambda > 0$ y $x \in \mathbb{N} \cup \{0\}$. Recordemos que si $X \sim f(x|\lambda)$, entonces $I_X(\lambda) = \frac{1}{\lambda}$, y en consecuencia la distribución de Jeffreys sería $\pi^{\mathcal{J}}(\lambda) \propto \lambda^{-\frac{1}{2}}$. Sin embargo, esta función no integra

en $\mathbb{R}^+$, el dominio de λ, pues $\int_0^\infty \lambda^{-\frac{1}{2}} d\lambda = [2\lambda^{\frac{1}{2}}]_0^\infty = \infty$. Estamos por tanto ante una distribución a priori impropia. En esta situación "normalizamos" la distribución, tomando $\pi^{\mathcal{J}}(\lambda) = c\lambda^{-\frac{1}{2}}$, con $c > 0$ arbitrario. Si conseguimos que la distribución a posteriori no dependa de c podremos utilizarla para hacer inferencia.

Pasemos a calcular la distribución a posteriori. Consideramos una muestra aleatoria simple $\underset{\sim}{X} = (X_1, \ldots, X_n)$ de X, y el vector de observaciones de la muestra $\underset{\sim}{x} = (x_1, \ldots, x_n)$. Entonces $f(\underset{\sim}{x}|\lambda) = \prod_{i=1}^n f(x_i|\lambda) = e^{-n\lambda}\lambda^{\sum x_i} / \prod x_i!$. La distribución conjunta es proporcional a

$$f(\underset{\sim}{x}|\lambda)\pi^{\mathcal{J}}(\lambda) = \frac{c}{\prod x_i!} e^{-n\lambda}\lambda^{\sum x_i - \frac{1}{2}}.$$

Por otro lado la distribución marginal de $\underset{\sim}{X}$ viene dada por

$$m(\underset{\sim}{x}) = \frac{c}{\prod_{i=1}^n x_i!} \int_0^\infty e^{-n\lambda}\lambda^{\sum x_i - \frac{1}{2}} d\lambda = \left[\begin{array}{l} y = n\lambda \\ dy = nd\lambda \end{array} \right] =$$

$$= \frac{c}{n^{\sum x_i + \frac{1}{2}} \prod_{i=1}^n x_i} \int_0^\infty e^{-y} y^{\sum x_i - \frac{1}{2}} dy = \frac{c}{n^{\sum x_i + \frac{1}{2}} \prod x_i!} \Gamma(\sum x_i + 1/2).$$

Podemos comprobar que ambas funciones están indeterminadas por c. Obtenemos la distirbución a priori realizando el cociente $f(\underset{\sim}{x}|\lambda)\pi^{\mathcal{J}}(\lambda)/m(\underset{\sim}{x})$. En efecto, obtenemos

$$\pi(\lambda|\underset{\sim}{x}) = \left(\frac{c}{\prod x_i!} e^{-n\lambda}\lambda^{\sum x_i - \frac{1}{2}} \right) / \left(\frac{c}{\prod x_i!} \frac{1}{n^{\sum x_i + \frac{1}{2}}} \Gamma(\sum x_i + 1/2) \right) = \frac{e^{-n\lambda}\lambda^{\sum x_i - \frac{1}{2}} n^{\sum x_i + \frac{1}{2}}}{\Gamma(\sum x_i + 1/2)}.$$

Esto es, la distribución a posteriori es una $Gamma(\lambda, \sum x_i + 1/2, n)$. △

Hemos visto por tanto que una distribución a priori impropia también nos permite realizar inferencia. Un último detalle que podemos observar es que $m(\underset{\sim}{x}) \neq \prod m(x_i)$. Las variables X_i son independientes cuando se condicionan a θ, pero no mantienen la independencia cuando se consideran todos los posibles parámetros. Decimos que las muestras no son incondicionalmente independientes. Esta es otra de las diferencias entre la estadística clásica y la bayesiana. En la primera bajo cualquier concepto las muestras son independientes ya que el parámetro es un valor fijo, no una variable aleatoria.

6.5. Convergencia de distribuciones a posteriori

Nos planteamos en este punto la convergencia de las distribuciones a posteriori cuando el tamaño de las muestras crece.

Supongamos que queremos hacer inferencia sobre un fenómeno que sigue una distribución $f(x|\theta_0)$, con una medida de probabilidad P_{θ_0}. Conocemos la familia de distribuciones $\{f(x|\theta)|\theta \in \Theta\}$. Queremos ver, dada una distribución a priori $\pi(\theta)$, si converge la distribución a posteriori $\pi(\theta|X_1, \ldots, X_n)$, para las variables i.i.d. $X_1, \ldots, X_n \sim f(x|\theta_0)$, cuando $n \to \infty$.

En general, si el espacio paramétrico no es discreto, el estudio de la convergencia de la distribución marginal no es sencillo. Estudiaremos la convergencia de la distribución a posteriori cuando el espacio paramétrico es $\Theta = \{\theta_1, \ldots, \theta_k\}$ discreto.

Teorema 6.10. *En las condiciones anteriores, se tiene que*

$$\pi(\theta|X_1, \ldots, X_n) \xrightarrow[n\to\infty]{P_{\theta_0}} \theta_0.$$

Demostración. Supongamos $\Theta = \{\theta_1, \dots, \theta_k\}$. La distribución a priori viene determinada por las probabilidades de cada θ_j y vendrá dada por $\pi(\theta_j) = p_j$, con $p_j \in [0,1]$ y $\sum_{i=1}^k p_i = 1$, para $j = 1, \dots, k$. Podemos suponer también que $t \in \{1, \dots, k\}$ es el índice del parámetro que corresponde al verdadero valor, es decir, $\theta_t = \theta_0$.

Consideramos las variables aleatorias i.i.d. $X_1, \dots, X_n$ con distribución $f(x|\theta_t)$ y una muestra $x_1, \dots, x_n$. La distribución a posteriori en este caso discreto vendrá dada, para cada θ_i, por:

$$\pi(\theta_i|x_1, \dots, x_n) = \frac{p_i \prod_{j=1}^n f(x_j|\theta_i)}{\sum_{r=1}^k p_r \prod_{j=1}^n f(x_j|\theta_r)}$$

Multiplicando numerador y denominador por $\left(\prod_{j=1}^n f(x_j|\theta_t)\right)^{-1}$ obtenemos:

$$\pi(\theta_i|x_1, \dots, x_n) = \frac{p_i \prod_{j=1}^n \frac{f(x_j|\theta_i)}{f(x_j|\theta_t)}}{\sum_{r=1}^k p_r \prod_{j=1}^n \frac{f(x_j|\theta_r)}{f(x_j|\theta_t)}} \tag{12}$$

Estudiemos la convergencia de $\prod_{j=1}^n \frac{f(x_j|\theta_i)}{f(x_j|\theta_t)}$. Si $i = t$, claramente tenemos que el resultado es 1. En caso contrario, tomando logaritmos, obtenemos:

$$\log \prod_{j=1}^n \frac{f(x_j|\theta_i)}{f(x_j|\theta_t)} = \sum_{j=1}^n log \frac{f(x_j|\theta_i)}{f(x_j|\theta_t)} = n \left(\frac{1}{n} \sum_{j=1}^n log \frac{f(x_j|\theta_i)}{f(x_j|\theta_t)} \right) \tag{13}$$

Ahora, las variables aleatorias $Z_i \sim \log \frac{f(x_j|\theta_i)}{f(x_j|\theta_t)}$ son i.i.d, luego por las leyes de los grandes números el término $\frac{1}{n} \sum_{j=1}^n log \frac{f(x_j|\theta_i)}{f(x_j|\theta_t)}$ converge en probabilidad P_{θ_t} a la esperanza de cualquiera de ellas, $E\left[\log \frac{f(x_j|\theta_i)}{f(x_j|\theta_t)}\right]$. Además, como consecuencia de la desigualdad de la información (proposición 3.16), dicha esperanza es un valor estrictamente negativo. En consecuencia, a partir de la expresión obtenida en (13), podemos concluir que:

$$\log \prod_{j=1}^n \frac{f(x_j|\theta_i)}{f(x_j|\theta_t)} \xrightarrow[n\to\infty]{P_{\theta_t}} -\infty$$

La continuidad del logaritmo nos asegura que $\prod_{j=1}^n \frac{f(x_j|\theta_i)}{f(x_j|\theta_t)} \xrightarrow[n\to\infty]{P_{\theta_t}} 0$.

Finalmente, aplicando lo que acabamos de obtener a (12), concluimos que:

- Si $i \neq t$, $\pi(\theta_i|X_1, \dots, X_n) \xrightarrow[n\to\infty]{P_{\theta_t}} \frac{0}{p_t + \sum 0} = 0$
- Si $i = t$, $\pi(\theta_i|X_1, \dots, X_n) \xrightarrow[n\to\infty]{P_{\theta_t}} \frac{p_t}{p_t + \sum 0} = 1$

Pero esto es equivalente a decir que la distribución a posteriori degenera a θ_t en probabilidad P_{θ_t}. □

Comentario 6.11. Observemos que en la convergencia de la distribución a posteriori no ha influido la distribución a priori escogida. Esto nos indica que cuando el tamaño de la muestra es grande, la distribución que hayamos elegido a priori no va a tener mucha influencia sobre la distribución que utilizaremos para realizar inferencia.

7. Test de Hipótesis Bayesianos

En este apartado vamos a realizar un estudio similar al que se hizo con los test de hipótesis clásicos, pero haciendo uso de las herramientas que nos proporciona la estadística bayesiana. Para ello, recordemos que si queremos considerar un modelo bayesiano, tenemos que dotarnos de una familia de densidades y de una distribución a priori, como se sigue : ¿$\mathcal{M} = \{\pi_i(\theta)|i \in I\, \pi(\theta)\}$?.

Entonces consideramos dos posibles situaciones, con las que obtenemos dos modelos $M_1 : \{f(x|\theta_1, M_1), \pi(\theta_1, M_1)\}$ y $M_2 : \{f(x|\theta_2, M_2), \pi(\theta_2, M_2)\}$

Ahora bien, sabiendo que $\pi(\theta_i, M_i) = \pi(M_i)\pi(\theta_i|M_i)$ para $i = 1, 2.$, lo aplicamos a nuestros dos modelos y obtenemos que $M_1 : \{f(x|\theta_1, M_1), \pi(M_1)\pi(\theta_1|M_1)$ y $M_2 : \{f(x|\theta_2, M_2), \pi(M_2)\pi(\theta_2|M_2)$.

Supongamos que tenemos una muestra (independiente e idénticamente distribuida) de tamaño n que proviene de alguno de los modelos, esto es $\underset{\sim}{x} = (x_1, ..., x_n)$. Con ella ,se quiere calcular la probabilidad a posteriori del modelo M_1. Para ello, necesitamos tomar la verosimilitud de la muestra respecto de θ_i y $M_i (i = 1, 2)$, es decir, $\prod_{i=1}^n f(x_i|\theta_1, M_1)$ y $\prod_{i=1}^n f(x_i|\theta_2, M_2)$.

Con ello, la probabilidad a posteriori del modelo M_1 se obtiene de la siguiente forma:

$$P(\theta_1, M_1|\underset{\sim}{x}) = \frac{\prod_{i=1}^n f(x_i|\theta_1, M_1)\pi(\theta_1, M_1)}{\pi(M_1)m(\underset{\sim}{x}|M_1) + \pi(M_2)m(\underset{\sim}{x}|M_2)}$$

Por tanto, la probabilidad del modelo M_1 condicionado a la muestra $\underset{\sim}{x}$ es $P(M_1|\underset{\sim}{x}) = \frac{m(\underset{\sim}{x}|M_1)\pi(M_1)}{m(\underset{\sim}{x}|M_1)\pi(M_1)+m(\underset{\sim}{x}|M_2)\pi(M_2)}$.

Esta expresión se pude reescribir usando el factor de Bayes $B_{21}(\underset{\sim}{x}) = \frac{m(\underset{\sim}{x}|M_2)}{m(\underset{\sim}{x}|M_1)}$, por lo que la expresión anterior quedaría, dividiendo todo entre el numerador como: $P(M_1|\underset{\sim}{x}) = \frac{1}{1+B_{21}(\underset{\sim}{x})\frac{\pi(M_2)}{\pi(M_1)}}$.

Por tanto, el cociente entre las probabilidades de que se da cada modelo condicionado a la muestra es: $\frac{P(M_2|\underset{\sim}{x})}{P(M_1|\underset{\sim}{x})} = B_{21}(\underset{\sim}{x})\frac{\pi(M_2)}{\pi(M_1)}$.

7.1. Método de Leamer

En la sección de estadística bayesiana vimos que podíamos escoger distribuciones a priori no integrables (impropias) que aun así nos permitían obtener buenas distribuciones a posteriori sobre las que realizar inferencia. ¿Qué ocurre si cogemos como distribución a priori una distribución impropia? La respuesta se basa en que al calcular el factor de Bayes de una muestra, dicho factor está indeterminado por una constante multiplicativa que proviene de las dos distribuciones a priori impropias, esto es, si $\pi(\theta_i|M_i) = c_i h_i(\theta_i)$, con h_i no integrable y $c_i > 0$ arbitrario, entonces $\pi(\theta_1|M_1) = c_1 h_1(\theta_1)$ y lo mismo para la el segundo modelo, obtenemos que $B_{21}(\underset{\sim}{x}) = \frac{c_2 \int f(\underset{\sim}{x}|\theta_2, M_2)h_2(\theta_2)d\theta_2}{c_1 \int f(\underset{\sim}{x}|\theta_1, M_1)h_1(\theta_1)d\theta_1}$. Por ello, debemos coger distribuciones propias.

El método de Leamer consiste en obtener una submuestra de una muestra, que llamaremos muestra de entrenamiento, de forma que al calcular la distribución a posteriori esté bien definida. Para ello, tomamos $\underset{\sim}{x_1}$ submuestra de $\underset{\sim}{x}$. Para que $\pi(\theta_1|\underset{\sim}{x_1}, M_1)$ esté bien definida, se tiene que verificar que $0 < m(\underset{\sim}{x_1}|M_1) < \infty$ (análogo para el modelo M_2).

Definición 7.1. Dado un modelo M y una muestra $\underset{\sim}{x}$, una muestra de entrenamiento es un subconjunto de la muestra $\underset{\sim}{x_1} \subset \underset{\sim}{x}$. Se dice que la muestra es propia si $0 < m(\underset{\sim}{x_1}|M) < \infty$.

Una muestra de entrenamiento $\underset{\sim}{x_1}$ se dice que es minimal si es propia y ningún subconjunto suyo distinto de $\underset{\sim}{x_1}$ lo es.

Como acabamos de ver, la selección de una muestra de entrenamiento minimal nos permite poder continuar con distribuciones a priori impropias al inicio, que pasarán a ser propias tras entrenarlas con dicha muestra. El principal inconveniente de este método es determinar qué datos son los más convenientes para entrenar la distribución a priori, es decir, qué subconjunto $\underset{\sim}{x_1}$ minimal escoger.

Ejemplo 7.1: Vimos (Ejemplo 6.9) que la distribución de Jeffreys para una Poisson de parámetro λ es $\pi^{\mathcal{J}}(\lambda) = c\lambda^{-1/2}$ impropia. Sin embargo, para cualquier muestra $\underset{\sim}{x} = (x_1, \dots, x_n)$, teníamos que $m(\underset{\sim}{x}) = \frac{c}{n^{\sum x_i + \frac{1}{2}} \prod x_i!}\Gamma(\sum x_i + \frac{1}{2}) < \infty$. Por tanto, $\underset{\sim}{x}$ es una muestra de entrenamiento propia para cualquier tamaño, y en particular lo es para un solo dato. Por tanto, solo necesitamos un dato para hacer la distribución a priori propia, y la muestra de entrenamiento minimal consta de un solo dato. △

Ejemplo 7.2: En una distribución normal $\mathcal{N}(\mu, \sigma^2)$ la distribución de Jeffreys es $\pi^{\mathcal{J}}(\mu, \sigma^2) = \frac{c}{\sigma^2}\chi_{\mathbb{R}\times\mathbb{R}^+}(\mu, \sigma^2)$, no integrable. Para un solo dato la distribución a posteriori sigue sin ser integrable, pero con dos ya sí lo es. Por tanto, la muestra de entrenamiento minimal es de tamaño 2. △

Ejemplo 7.3: Supongamos que tenemos una distribución normal con varianza conocida, que podemos tomarla como 1, esto es, $\mathcal{N}(x|\mu, 1)$. Nuestro test va a consistir si μ es 0. Para ello, tomamos $\mathcal{H}_0 : \mu = 0$ como la hipótesis nula y $\mathcal{H}_1 : \mu \in R$.

Consideramos los dos modelos asociados $\mathcal{M}_0 : \mathcal{N}(x|0, 1), \pi(M_0)$ y $\mathcal{M}_1 : \mathcal{N}(x|\mu, 1), \mathcal{N}(\mu|0, 2)\pi(M_1)$, donde la distribución normal de media 0 y varianza 2 se ha obtenido haciendo la distribución intrínseca de μ, pero se puede usar también una familia de distribuciones conjugada y el resultado será el mismo.

Nuestro objetivo es calcular la probabilidad de que se de el primer modelo condicionado a una muestra de n datos. Entonces, $P(M_0|\underset{\sim}{x} = (1 + B_{10}(\overline{x}, n)\frac{\pi(M_1)}{\pi(M_0)})^{-}1)$.

Recordemos que $\prod_{i=1}^n \mathcal{N}(x_i|\mu, 1) = \mathcal{N}(\overline{x}|\mu, 1/n)$.

Para ello, calculamos el factor de Bayes anterior: $B_{10}(\overline{x}, n) = \frac{\int_{-\infty}^{\infty} \mathcal{N}(\overline{x}|\mu,1/n)\mathcal{N}(\mu|0,2)}{\mathcal{N}(\overline{x}|0,1/n)} = \frac{\mathcal{N}(\overline{x}|0,2+1/n)}{\mathcal{N}(\overline{x}|0,1/n)}$.

Entonces, reescribiendo esta última fracción nos queda que el factor de Bayes se expresa de la siguiente forma:

$B_{10}(\overline{x}, n) = \frac{1}{\sqrt{2n+1}} \exp^{\frac{n^2\overline{x}^2}{2n+1}}$.

Estudiemos el comportamiento probabilístico en el límite.

Si queremos ver qué ocurre en probabilidad M_1 cuando n diverge, nos fijamos en el exponente de la función exponencial lo primero y vemos que se comporta como $(\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu) + \sqrt{n}\mu)^2$, pero el primer término va a una normal mientras que el segundo, va a infinito, por lo tanto,

$B_{10}(\overline{x}, n) = \frac{1}{\sqrt{2n+1}} \exp^{\frac{(n\overline{x})^2}{2n+1}} \xrightarrow[n\to\infty]{M_1} \infty$.

Por tanto, $P(M_0|\underset{\sim}{x}) \xrightarrow[n\to\infty]{0}$ y, en consecuencia, $P(M_1|\underset{\sim}{x}) \xrightarrow[n\to\infty]{1}$. Recordemos que estamos viendo que ocurre cuando suponemos que queremos ver qué pasa cuando tomamos el modelo M_1.

Ahora, realizamos el mismo proceso, pero tomando como referencia el modelo M_0. En este caso, el exponente del factor de Bayes se comporta como $(\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i))^2$, que bajo el modelo M_0 converge a algo finito. Con lo cual, $B_{10}(\overline{x}, n) = \frac{1}{\sqrt{2n+1}} \exp^{\frac{(n\overline{x})^2}{2n+1}} \xrightarrow[n\to\infty]{M_0} 0$, lo que nos indica que la probabilidad

de que se me el modelo M_0 con esa muestra $\underset{\sim}{x}$ es 1, es decir, $P(M_0|\underset{\sim}{x} \overset{1}{\to})$ (bajo el modelo M_0) y, en consecuencia, $P(M_1|\underset{\sim}{x} \overset{0}{\to})$.

Referencias

[1] Probability theory, M. Loève, 1977, Springer-Verlag.

[2] Statistical Inference, G. Casella, R. L. Berger, segunda edicón (2002), Duxbury Advanced Series.

[3] Proof Wiki, Euler's Reflection Formula, https://proofwiki.org/wiki/Euler%27s_Reflection_Formula.

[4] Wikipedia, Residue theorem, https://en.wikipedia.org/wiki/Residue_theorem#Example.

[5] Wikipedia, Leibniz integral rule, https://en.wikipedia.org/wiki/Leibniz_integral_rule.

[6] Davide Giraudo, No first moment and differentiable characteristic function, http://math.stackexchange.com/questions/793788/continuous-probability-distribution-with-no-first-moment-but-the-characteristic