Ampliación de Inferencia Estadística
Facultad de Matemáticas
Tercero Grado de Estadística
Tema 2. Introducción a los conceptos de Inferencia Bayesiana.
La Información de Fisher y la a priori de Jeffrey

1. Idea Intuitiva

El objetivo básicamente es cuantificar la *información* que aporta una muestra acerca del parámetro que se pretende estimar. Obviamente hay que especificar el concepto de información. De alguna forma se puede decir que un valor de X, por ejemplo x, es *informativo* acerca de un valor particular del parámetro θ_0 cuando el valor de la función de densidad varía mucho para un valor del parámetro muy cercano a θ_0. Dicho en otros términos, el valor de la función de densidad en x es muy sensible a pequeñas perturbaciones en los valores del parámetro.

En definitiva, consideremos una familia paramétrica de distribuciones $\{f_X(x;\theta), \theta \in \Theta\}$. Fijemos un valor del paramétro θ_0 y un valor muy próximo $\theta' = \theta_0 + d\theta$. Hay muchas formas de medir la distancia entre funciones de densidad o la discrepancias entre ellas. Una puede ser el cociente entre ambas funciones de densidad:

$$r(\theta_0, \theta') = \frac{f_X(x;\theta_0)}{f_X(x;\theta')}.$$

Esta razón se puede interpretar como que si es un valor excesivamente grande signifca que la función de densidad disminuye drásticamente en un entorno cercano a θ_0 y por tanto el valor de x si aportaría *información* para θ_0. Se tendría la interpretación opuesta si dicho cociente es muy próximo a cero. Si se realiza una transformación creciente la interpretación se mantiene en función de los valores de dicha transformación. En Estadística una transformación muy usada es la logarítmica ya que *contrae* de alguna forma los datos. En este caso vamos a definir la distancia entre esas dos funciones de densidad como el logaritmo de $r(\cdot,\cdot)$. Sin embargo, tal como la estamos construyendo, esa distancia depende del valor de x que a su vez es una observación de una variable aleatoria X. Por tanto, la distancia o discrepancia entre esas dos funciones de densidad la podemos definir como el valor esperado de esa distancia para de esta forma hacerla solo dependiente del parámetro θ_0. Es decir,

$$l(\theta_0, \theta') = E_{\theta_0}\left[\log f_X(X, \theta_0) - \log f_X(X;\theta')\right] \tag{1.1}$$

Si suponemos que la función de densidad $f_X(x;\theta')$ es derivable respecto a θ al menos hasta el orden 2, podemos aproximar la discrepancia mediante su desarrollo en serie de

Taylor obteniéndose

$$l(\theta_0,\theta') \approx E_{\theta_0}\left[-\left(\frac{\partial \log f_X(X;\theta)}{\partial\theta}\right)_{\theta=\theta_0} d\theta - \frac{1}{2}\left(\frac{\partial^2 \log f_X(X;\theta)}{\partial\theta^2}\right)_{\theta=\theta_0}(d\theta)^2\right] \quad (1.2)$$

Sin embargo, se demuestra el seguiente Lema.

Lema 1.1. *Sea $f_{\mathbf{X}}(\mathbf{x};\theta)$ la función de verosimilitud de θ tal que se verifica*

$$\frac{\partial}{\partial\theta}\int f_{\mathbf{X}}(\mathbf{x};\theta)d\mathbf{x} = \int \frac{\partial}{\partial\theta} f_{\mathbf{X}}(\mathbf{x};\theta)d\mathbf{x}.$$

Entonces,

$$E_\theta\left[\frac{\partial \log f_{\mathbf{X}}(\mathbf{X};\theta)}{\partial\theta}\right] = 0.$$

Demostración. Se tiene que

$$\frac{\partial \log f_{\mathbf{X}}(\mathbf{x};\theta)}{\partial\theta} = \frac{\frac{\partial}{\partial\theta} f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{x};\theta)}.$$

Por tanto,

$$\begin{aligned} E_\theta\left[\frac{\partial \log f_{\mathbf{X}}(\mathbf{X};\theta)}{\partial\theta}\right] &= \int f_{\mathbf{X}}(\mathbf{x};\theta)\frac{\partial \log f_{\mathbf{X}}(\mathbf{x};\theta)}{\partial\theta}d\mathbf{x} \\ &= \int \frac{\frac{\partial}{\partial\theta} f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{x};\theta)} f_{\mathbf{X}}(\mathbf{x};\theta)d\mathbf{x} \\ &= \int \frac{\partial}{\partial\theta} f_{\mathbf{X}}(\mathbf{x};\theta)d\mathbf{x} \\ &= \frac{\partial}{\partial\theta}\int f_{\mathbf{X}}(\mathbf{x};\theta)d\mathbf{x} \\ &= 0. \end{aligned} \quad (1.3)$$

□

A partir de ahora vamos a suponer que las condiciones de regularidad son la posibilidad de intercambiar el orden de las operaciones de derivación e integración en los casos como se ha hecho en la demostración previa y que el recorrido de la distribución no dependa de ningún parámetro. Por tanto, queda

$$\begin{aligned} l(\theta_0,\theta') &\approx E_{\theta_0}\left[-\frac{1}{2}\frac{\partial^2 \log f_X(X;\theta)}{\partial\theta^2}\,|_{\theta=\theta_0}\right](d\theta)^2 \\ &= \frac{1}{2} I_X(\theta_0)(d\theta)^2 \end{aligned} \quad (1.4)$$

Esto quiere decir que el valor de $I_X(\theta_0)$ es el coeficiente de $(d\theta)^2$ en la medida aproximada de la discrepancia entre las dos densidades. Cuando $I_X(\theta_0)$ es grande significa que una alteración de $d\theta$ en el valor del parámetro da lugar a dos distribuciones muy separadas, y cada observación es muy informativa. El caso opuesto sería cuando $I_X(\theta_0)$ tomase valores muy próximos a cero. Entonces, ambas distribuciones serían (hasta términos de segundo orden) iguales, y las observaciones de X serían nulamente infomnativas (si los dos valores del parámetro, θ y θ', dan lugar a distribuciones idénticas, el observar los valores que toma X no permite discriminar entre una y otra). Al valor $I_X(\theta)$ se le conoce con el nombre de Información de Fisher de la variable aleatoria X sobre el parámetro θ.

2. Resultados importantes

Lema 2.1. *Bajo las condiciones de regularidad, se tiene que*

$$var_\theta\left(\frac{\partial \log f_{\mathbf{X}}(\mathbf{X};\theta)}{\partial\theta}\right) = I_X(\theta) \tag{2.1}$$

Demostración. Se tiene que

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}0 \\ &= \frac{\partial}{\partial\theta}E_\theta\left[\frac{\partial \log f_{\mathbf{X}}(\mathbf{X};\theta)}{\partial\theta}\right] \\ &= \int \frac{\partial}{\partial\theta}\left(\frac{\partial \log f_{\mathbf{X}}(\mathbf{x};\theta)}{\partial\theta} f_{\mathbf{X}}(\mathbf{x};\theta)\right) d\mathbf{x} \\ &\quad \text{lo cual derivando y expresándolo con esperanzas sale} \\ &= E_\theta\left[\frac{\partial^2 \log f_{\mathbf{X}}(\mathbf{X};\theta)}{\partial\theta^2}\right] + E_\theta\left[\frac{\partial \log f_{\mathbf{X}}(\mathbf{X};\theta)}{\partial\theta}\right]^2. \end{aligned} \tag{2.2}$$

A partir de esta igualdad a cero y aplicando el teorema de König sale inmediatamente el resultado. □

Lema 2.2. *La información de Fisher asociada a una muestra aleatoria simple* $\mathbf{X}$ *formada por* n *observaciones es* $nI_X(\theta)$.

Demostración. Para una muestra aleatoria simple se tiene que

$$f_{\mathbf{X}}(\mathbf{x};\theta) = \prod_{i=1}^{n} f_X(x_i;\theta),$$

y por consiguiente

$$\frac{\partial \log f_{\mathbf{X}}(\mathbf{x};\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \log f_X(x_i;\theta)}{\partial \theta}.$$

Ahora bien, desarrollando el cuadrado de dicho sumatorio y tomando esperanzas, bajo las condiciones de regularidad se llega a que

$$I_{\mathbf{X}}(\theta) = nI_X(\theta).$$

□

Ejemplo 2.3. *Sea $X_1, \ldots, X_n$ una muestra aleatoria de una exponencial negativa de parámetro λ. La información de Fisher de la muestra valdría*

$$I_{\mathbf{X}}(\lambda) = nI_X(\lambda),$$

donde

$$I_X(\lambda) = -E_\lambda \left[\frac{\partial^2(\log \lambda - \lambda X)}{\partial \lambda^2} \right].$$

Por tanto,

$$I_{\mathbf{X}}(\lambda) = \frac{n}{\lambda^2}.$$

Luego, como era de esperar, a mayor tamaño muestral hay más información acerca del parámetro y, para un tamaño muestral fijo, la muestra aportará mayor información cuanto menor sea el valor del parámetro.

Ejemplo 2.4. *Supongamos X una variable aleatoria con distribución normal de media conocida μ_0 y varianza desconocida σ^2. Se calcula de manera inmediata que*

$$\log f_X(x;\sigma^2) = -\log(\sqrt{2\pi})\sigma - \frac{(x-\mu_0)^2}{2\sigma^2}.$$

Después de hacer las derivadas correspondientes respecto a σ y calcular la esperanza en X se obtiene que

$$I_{\mathbf{X}}(\sigma) = \frac{3n}{\sigma^2}.$$

Lo cual podemos interpretar que para tamaños fijos cuanto menor sea la varianza de la normal la muestra aleatoria simple aportará más información acerca de dicho parámetro.

Nota 1. *Uno de los problemas que presenta la información de Fisher es que no es invariante bajo transformación del parámetro. En particular, sea X una variable aleatoria con función de densidad $f(x;\theta)$, donde el valor del parámetro θ es desconocido pero*

pertenece a un espacio paramétrico Θ. Sea $I_o(\theta)$ la información de Fisher de X. Supongamos ahora que transformamos el parámetro por μ donde $\theta = \phi(\mu)$, con ϕ derivable. Es fácil demostrar que

$$I_1(\mu) = [\phi^{'}(\mu)]^2 I_0[\phi(\mu)].$$

Sin embargo, de esta desventaja se crea la a priori de Jeffrey no informativa a partir de la Información de Fisher:

$$\pi(\theta) \propto \sqrt{I(\theta)}.$$

El uso de esta a priori no informativa viene dada porque sí es invariante bajo transformación del parámetro (es por ello que se toma la raíz cuadrada) evitando así el problema que presenta la a priori uniforme.

Nota 2. *Otra utilidad de la información de Fisher es la selección de estimadores óptimos. Por ejemplo, sea T un estimador de θ tal que $E(T) = \psi(\theta)$. Entonces, se demuestra que*

$$var(T) \geq \frac{\psi^{'}(\theta)}{I_{\mathbf{X}}(\theta)}.$$

Es decir, la varianza de los estimadores de una función del parámetro tiene una cota inferior que depende de la información de Fisher. Esa cota es conocida con el nombre de Cota de Cramer-Rao (C-R). En particular, será interesante estudiar que estimador alcanza dicha cota dentro del conjunto de estimadores insesgados de una función paramétrica. Por ejemplo, en el caso de X según una distribución normal de parámetros μ y σ^2. La Información de Fisher para la media vale

$$I_{\mathbf{X}}(\mu) = \frac{n}{\sigma^2}.$$

Luego la cota de C-R vale σ^2/n ya que en este caso se tiene que $\psi(\theta) = \theta$. Fijémonos que los estimadores de la media del tipo:

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

con ω_i un conjunto de ponderaciones, son estimadores insesgados. Un caso particular sería la media aritmética que se corresponde con $\omega_i = 1/n$ para todo i. También se tiene mediante un cálculo sencillo que

$$var(\bar{X}) = \frac{\sigma^2}{n},$$

es decir, entre todos los estimadores insesgados de la media poblacional, la media muestral es el que alcanza la cota de C-R y por tanto, no habrá otro estimador insesgado con menor variabilidad muestral. Luego, en este sentido es el óptimo.

Veamos el siguiente teorema que caracteriza un estadístico suficiente mediante la información de Fisher.

Teorema 2.5. *Sea* $\mathbf{X}$ *una muestra aleatoria de una distribución* $f(x;\theta)$ *y* $T = T(\mathbf{X})$ *un estadístico. Entonces se cumple que*

$$I_{\mathbf{X}}(\theta) \geq I_T(\theta) \text{ para todo } \theta \in \Theta,$$

dándose la igualdad si, y sólo si T *es suficiente para* θ.

Demostración. Por comodidad en la demostración vamos a suponer que X es una variable aleatoria absolutamente continua. Sea $T = T(\mathbf{X})$ un estadístico, entonces la función de densidad de $\mathbf{X}$ condicionada a $T = t$ viene dada por

$$h_{\mathbf{X}|T=t}(\mathbf{x};\theta) = \begin{cases} 0, & \text{si } T(\mathbf{x}) \neq t \\ \frac{f(\mathbf{x};\theta)}{g_T(t;\theta)}, & \text{si } T(\mathbf{x}) = t. \end{cases}$$

En el caso particular que $T(\mathbf{x}) = t$ se tiene que

$$f(\mathbf{x};\theta) = g_T(t;\theta) h_{\mathbf{X}|T=t}(\mathbf{x};\theta).$$

Luego, tomando logaritmo y derivando dos veces respecto al parámetro se obtiene

$$\frac{\partial^2}{\partial\theta^2} \log f(\mathbf{x};\theta) = \frac{\partial^2}{\partial\theta^2} \log g_t(t;\theta) + \frac{\partial^2}{\partial\theta^2} \log h_{\mathbf{X}|T=t}(\mathbf{x};\theta).$$

En definitiva, tomando ahora esperanzas se llega a que

$$I_{\mathbf{X}}(\theta) = I_T(\theta) + I_{\mathbf{X}|T(\mathbf{X})}(\theta) \text{ para todo } \theta,$$

y como la función de información de Fisher es no negativa se tiene que

$$I_{\mathbf{X}}(\theta) \geq I_T(\theta) \text{ para todo } \theta.$$

La igualdad en la expresión anterior se cumple si, y solo si

$$0 = I_{\mathbf{X}|T(\mathbf{X})}(\theta).$$

O equivalentemente,

$$0 = E\left[\left(\frac{\partial}{\partial\theta} \log h_{\mathbf{X}|T=t}(\mathbf{x};\theta)\right)^2\right],$$

Es decir,

$$\frac{\partial}{\partial\theta} \log h_{\mathbf{X}|T=t}(\mathbf{x};\theta) = 0, \text{ para todo } \mathbf{x} \text{ tal que } T(\mathbf{x}) = t.$$

Lo cual implica que la distribución de $\mathbf{X}$ condicionada a $T(\mathbf{X}) = t$ no depende de θ, por lo que T es suficiente. □