Apuntes sobre la mediana y su estimación

Probabilidad y Estadística 2012

1. La mediana

Definición: Sea X una variable aleatoria con función de distribución F. Se define la mediana como

$$m = \text{mín}\{x \in \mathbb{R}/F(x) \geq 1/2\}$$

La existencia de dicho mínimo está garantizada porque $\lim_{x\to+\infty} F(x) = 1$ y F es continua por derecha.

Observaciones:

1. $P(X \geq m) \geq 1/2$ y $P(X \leq m) \geq 1/2$.

   Dem: Llamemos $A = \{x \in \mathbb{R}/F(x) \geq 1/2\}$

   $P(X \geq m) \geq 1/2$ si y solo si $P(X < m) \leq 1/2$. $P(X < m) = \lim_{x\to m^-} F(x)$, pero si $x < m$ entonces $x \notin A$ por ser m el mínimo, entonces $F(x) < 1/2$ y por lo tanto $P(X < m) \leq 1/2$.

   $P(X \leq m) = F(m) \geq 1/2$ porque m es el mínimo de A y por lo tanto pertenece a A.

2. Si F es continua entonces $P(X \geq m) = P(X \leq m) = 1/2$.

   Dem: $1/2 \leq P(X \leq m) = P(X < m) = 1 - P(X \geq m) \leq 1 - 1/2 = 1/2$

3. Si X es absolutamente continua con densidad f entonces $\int_{-\infty}^{m} f = \int_{m}^{+\infty} f = 1/2$.

   Dem: Es una consecuencia inmediata de la observación anterior, pues $P(X \leq m) = \int_{-\infty}^{m} f$.

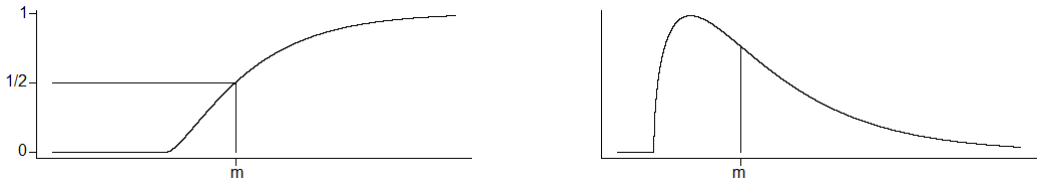

Figura 1: Función de distribución y densidad de una variable absolutamente continua y su mediana

Ejemplos:

1. X tiene distrbución de Cauchy de parámetros μ, σ si su densidad es $f(x) = \frac{\sigma}{\pi(\sigma^2+(x-\mu)^2)}$.

   Para hallar la mediana de X utilizamos la observación 3. Como f es simétrica respecto a μ (esto es $f(\mu - t) = f(\mu + t)\ \forall t$), $\int_{-\infty}^{\mu} f = 1/2$. Entonces la mediana es μ.

2. Algunos ejemplos para variables discretas:

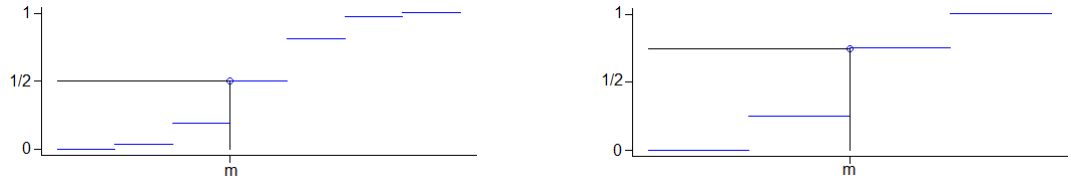

Figura 2: Funciones de distribución de variables discretas y su mediana

2. La mediana empírica

Sea $X_1, ..., X_n$ iid, una muestra de variables aleatorias con mediana m desconocida. Nuestro objetivo es estimar m. Existen varios estimadores de m conocidos como mediana empírica, nosotros daremos una definición y comentaremos luego otras posibles.

Definición: Sea $X_1^* \leq ... \leq X_n^*$ la muestra ordenada. La mediana empírica m_n se define como:

$$m_n = \begin{cases} (X^*_{\frac{n}{2}} + X^*_{\frac{n}{2}+1})/2 & \text{si } n \text{ es par} \\ \\ X^*_{\frac{n+1}{2}} & \text{si } n \text{ es impar} \end{cases}$$

Ejemplo: Si la muestra ordenada es $1, 3, 7, 10$, la mediana empírica es $\frac{3+7}{2} = 5$. Si la muestra ordenanda es $1, 3, 7, 10, 11$, la mediana empírica es 7.

Bajo ciertas hipótesis la mediana empírica es un estimador consistente de la mediana. Enunciemos un teorema muy general al respecto:

Proposición: $X_1, ..., X_n \sim F$ iid y m su mediana. Si F es tal que $\forall \varepsilon > 0$ $F(m + \varepsilon) > 1/2$ entonces $m_n \xrightarrow{c.s.} m$

Por motivos de simplicidad vamos a demostrar una proposición que es un caso particular de la anterior:

Proposición: $X_1, ..., X_n \sim F$ iid y m su mediana. Si F es continua y estrictamente creciente entonces $m_n \xrightarrow{c.s.} m$

Dem: Para esta demostración vamos a necesitar recordar que si F_n es la distribución empírica asociada a la muestra $X_1, ..., X_n$, entonces casi seguramente $F_n(x) \xrightarrow{n} F(x)$ $\forall x \in \mathbb{R}$.

Sea $\varepsilon > 0$, casi seguramente

$$F_n(m - \varepsilon) \xrightarrow{n} F(m - \varepsilon) < F(m) \text{ pues } F \text{ estrictamente creciente.}$$

$$F_n(m + \varepsilon) \xrightarrow{n} F(m + \varepsilon) > F(m) \text{ pues } F \text{ estrictamente creciente.}$$

$$F_n(m_n) = \begin{cases} \frac{1}{n}\frac{n}{2} = \frac{1}{2} & \text{si } n \text{ es par} \\ \\ \frac{1}{n}\frac{n+1}{2} \xrightarrow{n} \frac{1}{2} & \text{si } n \text{ es impar} \end{cases}$$

$F(m) = \frac{1}{2}$, pues F es continua. Como $F_n(m_n) \xrightarrow{n} \frac{1}{2}$, tenemos que, casi seguramente

$$\lim_n F_n(m - \varepsilon) < \lim_n F_n(m_n) < \lim_n F_n(m + \varepsilon)$$

Entonces casi seguramente $\exists n_0$ tal que, $F_n(m-\varepsilon) < F_n(m_n) < F_n(m+\varepsilon)\ \forall n \geq n_0$. Pero entonces

$$m - \varepsilon < m_n < m + \varepsilon\ \forall n \geq n_0$$

Por lo tanto $m_n \xrightarrow{c.s.} m$.

Observación: Otros estimadores de la mediana (también conocidos como mediana empírica) son:

$$\underline{m_n} = \begin{cases} X^*_{\frac{n}{2}} & \text{si } n \text{ es par} \\ X^*_{\frac{n+1}{2}} & \text{si } n \text{ es impar} \end{cases}$$

$$M_n = \begin{cases} X^*_{\frac{n}{2}+1} & \text{si } n \text{ es par} \\ X^*_{\frac{n+1}{2}} & \text{si } n \text{ es impar} \end{cases}$$

Repase la demostración de consistencia de la mediana empírica, para verificar que, prácticamente con la misma prueba, se obtiene la consistencia de $\underline{m_n}$ y M_n.

Observación: La proposición anterior permite estimar el parámetro μ de una distribución de Cauchy, cuando no lo conocemos.

Ejemplo: Volvamos al ejemplo de la distribución de Cauchy. A continuación se presenta el gráfico de m_n en función de n para una simulación de 1000 variables con distribución de Cauchy de parámetros $\mu = 0$ y $\sigma = 1$ (puntos azules). Observe que, a medida que crece n, la mediana empírica se aproxima a la mediana.

Además, en el gráfico se presenta el promedio en función de n para esa misma muestra (puntos negros). Observe que a veces el promedio se ve enormemente perturbado. Esto se debe a la aparición de un dato muy lejano a 0. Sin embargo, la mediana no presenta esa sensibilidad. Puede parecer sorprendente que el promedio no se aproxime a ningún valor, pero observe que en este caso no se aplica la ley de los grandes números pues el valor esperado de una variable de Cauchy no existe.

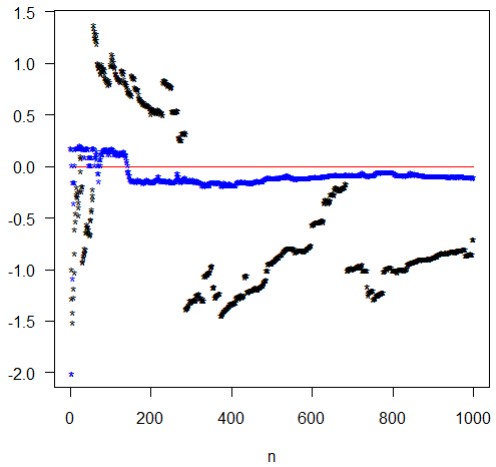

Figura 3: Mediana empírica (m_n en azul) y promedio ($\bar{X}_n$ en negro) en función de n, de una muestra de variables de Cauchy($\mu = 0, \sigma = 1$).