3 ESTIMACION PUNTUAL

3.1 INTRODUCCION

En un problema estadístico, si los datos fueron generados a partir de una distribución de probabilidad *F(x)* desconocida, los métodos de la **Inferencia Estadística** permiten decir algo respecto de esta distribución.

El problema de buscar una función de distribución de población se hace en varios tiempo. En presencia de una muestra, se supone que la distribución de donde proviene no es totalmente desconocida - por ejemplo pertenece a una determinada familia de distribuciones teóricas -entonces solamente uno o varios **parámetros** que definen la familia de distribuciones son desconocidos. En este caso la teoría de estimación tiene por objetivo dar valores a estos parámetros a partir de los valores muestrales.

Por ejemplo, *F(x)* pertenece a la familia de las distribuciones normales $N(\mu,1)$ de varianza igual a *1* y de esperanza μ desconocida. Aquí μ es el único parámetro desconocido de la distribución. Pero si se supone la varianza también desconocida, se tendrán dos parámetros desconocidos, la media μ y la varianza σ^2 de la distribución de población.

La elección de la familia de distribuciones se hace a partir de consideraciones teóricas o bien de la distribución de frecuencias empírica (o de métodos de Estimación No paramétrica que no hacen parte de este curso).

Los parámetros desconocidos son constantes que toman valores en un espacio Ω llamado **espacio de parámetros**:

$$\begin{array}{ll} N(\mu,1) & \Omega = \Re \\ N(\mu,\sigma^2) & \Omega = \Re \times]0,+\infty[\\ Exp(\beta) & \Omega =]0,+\infty[\\ Binomial(10,p) & \Omega = [0,1] \end{array}$$

Sean $X_1, X_2, ..., X_n$ valores muestrales obtenidos sobre una muestra aleatoria simple de una v.a. X de función de densidad $f(x|\theta)$, en que θ es desconocido. Hay varias maneras de decir algo sobre θ. Lo más simple consiste en dar un valor único para θ. Es **la estimación puntual**: se busca elegir un valor para θ a partir de los valores muestrales. Es decir se tiene que definir una función $\delta : \Re^n \to \Omega$, que es un estadístico llamado **estimador** de θ. El valor tomado por esta función sobre una muestra particular de tamaño *n* es una **estimación**. Otra forma de estimar un parámetro consiste en buscar no un sólo valor para θ, sino un conjunto de valores, un intervalo en general, en el cual se tiene alta probabilidad de encontrar θ. Es la el método de **estimación por intervalo**.

Procediendo así, tratamos de **estimar el valor de los parámetros**, que son considerados como constantes, a partir de estadísticos que son aleatorios. Ahora bien, frecuentemente se sabe algo más sobre los parámetros; este conocimiento obviamente no es preciso, sino no se tendría el problema de estimar estos parámetros; pero se tienen ideas sobre sus posibles valores, que pueden ser traducidas a una **función de distribución a priori** sobre el espacio de parámetro Ω. Los estimadores bayesianos toman en cuenta la distribución a priori y los valores muestrales.

El problema es que hay una regla única que permita construir estos estimadores. Por ejemplo, la esperanza y la mediana de una distribución de población simétrica son iguales: la media y la mediana empíricas –cuyos valores no son iguales en general- son entonces ambas estimaciones posibles para la esperanza.

Para elegir entre varios estimadores de un mismo parámetro hay que definir criterios de comparación. Empezaremos para presentar dos métodos de estimación (el método de los momentos y el método de máxima verosimilitud) y a continuación algunas propiedades razonables para decidir si un estimador es aceptable. Terminaremos por el método de Estimación Bayesiana.

3.2 MÉTODO DE LOS MOMENTOS

Vimos en el capítulo anterior que la media muestral $\overline{X}_n \overset{c.s.}{\to} E(X) = \mu$. Más generalmente si el momento de orden *r:* $\mu_r = E(X^r)$ existe, entonces por la ley de los grandes números:

$$m_r = \frac{1}{n}\sum_i X_i^r \overset{c.s.}{\to} \mu_r \quad (P(\lim_{n\to+\infty} m_r = \mu_r) = 1).$$

Luego una forma de estimar a μ_r consiste en tomar el momento empírico m_r.

Ejemplo: Este método produce como estimador de la media μ, $\hat{\mu} = \overline{X}_n$ y como estimador de la varianza σ^2, $\hat{\sigma}^2 = m_2 - m_1 = S_n^2$.

3.3 MÉTODO DE MÁXIMA VEROSIMILITUD

Sean $X_1, X_2, ..., X_n$ los valores muestrales de una muestra aleatoria simple de una v.a. de densidad o función de probabilidad $f(x|\theta)$ en que $\theta \in \Omega$, el espacio de parámetros.

Definición 1: Se llama función de verosimilitud a la densidad conjunta o función de probabilidad del vector aleatorio formado de los valores muestrales ($X_1, X_2, ..., X_n$), se denota $f_n(x_1, ...x_n|\theta)$.

Como los valores muestrales son independientes, se tiene: $f_n(x_1, ...x_n|\theta) = \prod_i f(x_i|\theta)$

Un estimador del parámetro θ basado en una muestra de tamaño *n* es una función δ de los valores muestrales $X_1, X_2, ..., X_n$ a valores en el espacio de parámetro Ω. El valor que toma el estimador δ sobre una muestra $x_1, x_2, ..., x_n$ se llama **estimación** o **valor estimado**.

El estimador de Máxima Verosimilitud es el estimador que hace la función de verosimilitud $f(x_1, ..., x_n|\theta)$ máxima. Tal estimador puede entonces no ser único, o bien no existir.

3.4 COMPARACIÓN DE LOS ESTIMADORES

Obviamente un buen estimador $\hat{\theta}$ para θ será aquel que tiene una diferencia $\hat{\theta} - \theta$ lo más pequeño posible. Pero como esta diferencia es aleatoria esperaremos que sea pequeña con alta probabilidad.

3.4.1 Estimador consistente

Un estimador depende del tamaño de la muestra a través de los valores muestrales; los estimadores $\hat{\theta}_n$ asociados a muestras de tamaño n ($n \in N$) constituyen sucesiones de v.a.. Un buen estimador debería converger en algún sentido hacia θ.

Definición 2: Se dice que un estimador $\hat{\theta}_n$ de un parámetro θ es **consistente** cuando converge en probabilidad hacia θ: $P(|\hat{\theta}_n - \theta| < \varepsilon) \to 1$.

Los momentos empíricos de una v.a. real son estimadores consistentes de los momentos teóricos correspondientes. Más aún la convergencia es casi-segura y la distribución asintótica de estos estimadores es normal.

3.4.2 Estimador insesgado

Definición 3: Se dice que un estimador $\hat{\theta}$ de θ es insegado si $E(\hat{\theta}) = \theta$.

Vimos que la media muestral $\overline{X}_n$ es un estimador insesgado de la media poblacional si la muestra es aleatoria simple, pero la varianza muestral $S_n^2 = \frac{1}{n}\sum_i (X_i - \overline{X}_n)^2$ no es un estimador insesgado para la varianza poblacional σ^2: $E(S_n^2) = \frac{n-1}{n}\sigma^2$

Pero, la diferencia si $E(S_n^2) - \sigma^2 = -\frac{1}{n}\sigma^2$, que se llama **sesgo**, tiende a cero.

Definición 4: Se dice que el estimador $\hat{\theta}$ es **asintoticamente insesgado** si $E(\hat{\theta}) \to \theta$ cuando $n \to +\infty$.

Por otro lado se puede construir un estimador insesgado de σ^2 a partir de S_n^2: $\tilde{\sigma}^2 = \frac{n}{n-1} S_n^2$.

Pero observamos que $\tilde{\sigma}^2 = S_{n-1}^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X}_n)^2$, es decir que, si bien el estimador $\tilde{\sigma}^2$ es insesgado para σ^2, tiene mayor varianza que S_n^2.

Por otro lado observamos que si $\hat{\theta}$ es un estimador sesgado de θ, se tiene:

$$E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + (sesgo)^2$$

En efecto, $E[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2] = E[(\hat{\theta} - E(\hat{\theta}))^2] + [E(\hat{\theta}) - \theta)]^2$

Si $E[(\hat{\theta}-\theta)^2] \to 0$ cuando crece el tamaño de la muestra, entonces $\hat{\theta}$ converge en media cuadrática hacia θ ($\hat{\theta} \overset{m.c.}{\to} \theta$).

Definición 5: Se llama **Error Cuadrático Medio** a $E[(\hat{\theta}-\theta)^2]$

Proposición 1: $E[(\hat{\theta}-\theta)^2] \to 0 \iff Var(\hat{\theta}) \to 0$ y $E(\hat{\theta}) \to \theta$

Como la convergencia en media cuadrática implica la convergencia en probabilidad se tiene una condición suficiente para que un estimador sea asintoticamente insesgado:

Proposición 2: Si $\hat{\theta}$ es un estimador consistente de θ y $E(\hat{\theta})$ es finito, entonces $\hat{\theta}$ es asintoticamente insesgado.

y una condición suficiente para que sea consistente:

Proposición 3: Si $Var(\hat{\theta}) \to 0$ y $E(\hat{\theta}) \to \theta$, entonces $\hat{\theta}$ es un estimador consistente de θ.

Ejemplo: observamos que el error cuadrático medio de la varianza muestral $S_n^2 = \frac{1}{n}\sum (X_i - \overline{X}_n)^2$ sesgada para σ^2 y es menor que el de $\tilde{\sigma}^2 = \frac{1}{n-1}\sum (X_i - \overline{X}_n)^2$ estimador insesgado de σ^2, pero de mayor varianza (Ver figura 1).

3.4.3 Estimador eficiente

Entre dos estimadores insesgados elegiremos obviamente aquel que tiene menos varianza.

Vamos establecer una desigualdad (CRAMER-RAO), que permite dar una cota inferior a la varianza de un estimador insesgado. Esta cota se basa en la cantidad de información de Fisher.

3.4.3.1 Cantidad de información de Fisher

Sea una v.a. *X* de función de densidad o función de probabilidad $f(x|\theta)$ en donde θ es una parámetro desconocido del conjunto Ω.

Definición 6: Se llama cantidad de información de Fisher dada por *X* sobre el parámetro θ a la cantidad

$$I(\theta) = E\left[\left(\frac{\partial \ln(f)}{\partial \theta}\right)^2\right]$$

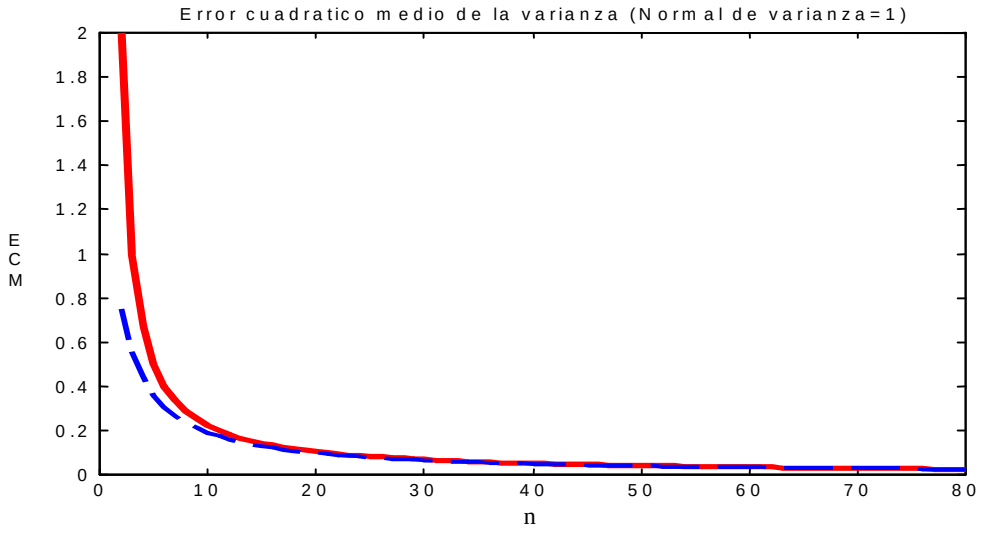

Figura 1: Comparación de los E.C.M. de $\tilde{\sigma}^2$ (———) y S_n^2 (▪▪▪▪▪▪▪).

Se puede dar dos otras formas a la cantidad de Información de Fisher:

Teorema 1: $I(\theta) = Var\left(\frac{\partial \ln(f)}{\partial \theta}\right)$

Demostración:

Sea S el dominio X, entonces como $\int_S f(x\,|\,\theta)dx = 1 \;\; \forall \theta \in \Omega$, se tiene $\int_S f'(x\,|\,\theta)dx = 0 \;\; \forall \theta \in \Omega$.

Además $\frac{\partial \ln f}{\partial \theta} = \frac{f'}{f}$, luego $E\left(\frac{\partial \ln f}{\partial \theta}\right) = 0 \quad \forall \theta \in \Omega$ y $I(\theta) = Var\left(\frac{\partial \ln(f)}{\partial \theta}\right)$

El teorema siguiente nos da otra expresión para $I(\theta)$, que a menudo es más fácil de determinar.

Teorema 2: Si el dominio S de X no depende de θ, entonces:

$$I(\theta) = -E\left[\left(\frac{\partial^2 \ln(f)}{\partial \theta^2}\right)\right] \quad \text{si esta cantidad existe.}$$

Demostración:

Si $\frac{\partial^2 \ln(f)}{\partial \theta^2}$ existe $\forall \theta \in \Omega$ entonces $E\left(\frac{\partial^2 \ln f}{\partial \theta^2}\right) = 0$.

Además $\frac{\partial^2 \ln(f)}{\partial\theta^2} = \frac{ff''-(f')^2}{f^2} = \frac{f''}{f} - \left(\frac{\partial \ln f}{\partial\theta}\right)^2$. Como $\frac{\partial^2 \ln(f)}{\partial\theta^2} = \int_S f''(x\,|\,\theta)dx - I(\theta)$, se deduce que $I(\theta) = -E\left[\left(\frac{\partial^2 \ln(f)}{\partial\theta^2}\right)\right]$

Sea una m.a.s. $\{x_1, x_2, ..., x_n\}$, x_i de función de densidad o función de probabilidad $f(x\,|\,\theta)$ en donde θ es una parámetro desconocido del conjunto Ω. Sea f_n la función de verosimilitud de la muestra.

Definición 7: Se llama **cantidad de información de Fisher** de una muestra aleatoria de tamaño *n* sobre el parámetro θ a la cantidad

$$I_n(\theta) = E\left[\left(\frac{\partial\, ln(\, f_n\,)}{\partial\theta}\right)^2\right]$$

Se tienen las dos otras formas de expresar $I_n(\theta)$ como en el caso de una v.a.. *X:*

$$I_n(\theta) = Var\left(\frac{\partial\, ln(\, f_n\,)}{\partial\theta}\right) \qquad I_n(\theta) = -E\left[\left(\frac{\partial^2\, ln(\, f_n\,)}{\partial\theta^2}\right)\right]$$

Es fácil deducir de lo anterior que:

Teorema 3: Si $I(\theta)$ es la cantidad de Fisher dada por cada x_i sobre el parámetro θ, entonces

$$I_n(\theta) = nI(\theta)$$

3.4.3.2.La desigualdad

Sea una m.a.s. $\{x_1, x_2, ..., x_n\}$, x_i de función de densidad o función de probabilidad $f(x\,|\,\theta)$. Se tiene la desigualdad de CRAMER-RAO:

Teorema 4: Si el dominio *S* de *X* no depende de θ, para todo estimador *T* insesgado de θ se tiene:

$$Var(T) \geq \frac{1}{I_n(\theta)}$$

Además si *T* es un estimador insesgado de $h(\theta)$, entonces $Var(T) \geq \frac{(h'(\theta))^2}{I_n(\theta)}$.

Demostración:

Como $E\left(\frac{\partial \ln f_n}{\partial \theta}\right)=0$, $Cov\left(T,\frac{\partial \ln f_n}{\partial \theta}\right)=E\left(T\frac{\partial \ln f_n}{\partial \theta}\right)=\int t\frac{\partial \ln f_n}{\partial \theta}Ldx=\int t\frac{\partial f_n}{\partial \theta}dx$

$Cov\left(T,\frac{\partial \ln f_n}{\partial \theta}\right)=\frac{\partial}{\partial \theta}\int tf_n dx=\frac{\partial}{\partial \theta}E(T)=h'(\theta)$.

Por otro lado, de la desigualdad de Schwarz se obtiene:

$$\left(Cov\left(T,\frac{\partial \ln f_n}{\partial \theta}\right)\right)^2 \le Var(T)Var\left(\frac{\partial \ln f_n}{\partial \theta}\right)$$

Es decir que

$$(h'(\theta))^2 \le Var(T)I_n(\theta).$$

La pregunta, que se plantea entonces, es si se puede alcanzar la cota mínima de la varianza. Un tal estimador se llama eficiente. Se estudian las condiciones para obtener un estimador eficiente en Inferencia Estadística.

3.4.4 Estimador suficiente

Si se busca deducir de las observaciones de una muestra aleatoria de *n* piezas en un proceso de fabricación una información sobre la proporción θ de piezas defectuosas en el lote total, es más simple considerar el número de piezas defectuosas encontradas en la muestra en vez de la sucesión de resultados $X_{1,}X_2,...,X_n$. El conocimiento de los valores individuales no procura ninguna información suplementaria para la proporción θ que $\sum_i X_i$. Se redujo los *n* datos a un sólo valor, que es función de estos datos, sin perder información para determinar a θ.

La media muestral $\overline{X}_n$ permite simplificar la información dada por los *n* valores muestrales. Pero nos preguntamos si se pierde información usando la media muestral para estimar la media μ de la población. Esto dependerá de los supuestos que se hacen sobre la distribución de población.

En el ejemplo 2, se supone un distribución de población normal, además observamos que si suponemos la varianza conocida y igual a *1*, la función de densidad conjunta, (la función de verosimilitud) puede escribirse como función únicamente de la media muestral y del tamaño *n* de la muestra:

$$f_n(x_1,x_2,...,x_n|\theta)=\left(\frac{1}{\sqrt{2\pi}}\right)^n exp(-\frac{n}{2}(\overline{X}_n-\theta)^2)$$

Es decir que la única información relevante para estimar a θ esta dada por la media muestral. En este caso se dice que la media muestral es un estadístico suficiente. Un estadístico suficiente que se toma como estimador del parámetro θ, debería contener toda la información que llevan los valores muestrales sobre θ.

Definición 8: Un estadístico $T(x_1,x_2,...,x_n)$, función de los valores muestrales y con valor en un conjunto Ω se dice suficiente para θ si la distribución conjunta de los valores muestrales condicionalmente a $T(x_1,x_2,...,x_n)$ no depende de θ.

Definición 9: Se dice que un estadístico *T* es suficiente minimal si no se puede encontrar otro estadístico suficiente que hace una mejor reducción de los datos que *T*.

No es siempre fácil detectar si un estadístico es suficiente. Los dos siguientes teoremas permiten enunciar condiciones para que un estadístico sea suficiente.

Teorema 2 (Principio de factorización): Si $T(x_1,x_2,...,x_n)$ es suficiente para θ y $g(T(x_1,x_2,...,x_n)|\theta)$ es la densidad de $T(x_1,x_2,...,x_n)$, entonces

$$f_n(x_1,...,x_n|\theta) = g(T(x_1,...,x_n)|\theta)h(x_1,...,x_n|T(x_1,...,x_n)$$

El principio de factorización nos permite de reconocer si un estadístico es suficiente, pero no permite construir uno o saber si existe uno. El siguiente teorema permite buscar estadísticos suficientes para una clase de distribuciones (llamadas exponenciales).

Se define el concepto de estadístico suficiente minimal como un estadístico suficiente que no puede ser reducido sin destruir la propiedad de suficiencia.

Teorema 3 (Theorema de Darmois-Koopman):

Si X es una variable real cuyo dominio de variación no depende del parámetro θ. Una condición necesaria y suficiente para que existe un estadístico suficiente es que la función de densidad de *X* sea de la forma:

$$f(x|\theta) = b(x)c(\theta)exp\{a(x)q(\theta)\}$$

Además $\exists T_n(X_1,...X_n) = \sum_i a(X_i)$ es un estadístico suficiente minimal.

Si $X \sim N(\theta,1)$ y si $X_1,X_2,...,X_n$ es una muestra aleatoria de *X*

$$f_n(x_1,...,x_n|\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^n exp\{-\frac{1}{2}\sum_i x_i^2\}exp\{-\frac{n\theta^2}{2} + n\theta\,\overline{X}_n\}$$

El término $exp\{-\frac{1}{2}\sum_i x_i^2\}$ no depende de θ y el término $exp\{-\frac{n\theta^2}{2} + n\theta\,\overline{X}_n\}$ depende de θ y $\overline{X}_n$.

$n\overline{X}_n$ es un estadístico suficiente; también toda función biyectiva de $\overline{X}_n$ lo es, en particular $\overline{X}_n$.

El siguiente teorema permite de construir *"mejores*" estimadores en el sentido del Error Cuadrático Medio a partir de un estadístico suficiente:

Teorema 4: (Theorema de Rao-Blackwell)

Si *T* es un estadístico suficiente para θ y si $b(X_1, X_2, ..., X_n)$ es un estimador insesgado de θ, entonces

$$\delta(T) = E(b(X_1, X_2, ..., X_n) | T)$$

es un estimador insesgado de θ basado sobre *T* mejor que el estimador $b(X_1, X_2, ..., X_n)$, es decir que

$$E[(\delta(T) - \theta)^2] \leq E[(b(X_1, X_2, ..., X_n) - \theta)^2]$$

3.4.5 Propiedades del Estimador de Máxima Verosimilitud

No es fácil encontrar buenos estimadores -insesgado, de varianza minimal; de hecho estas dos propiedades pueden ser antagónicas en el sentido que a buscar eliminar el sesgo se aumenta la varianza. Por otro lado la búsqueda de estimadores insesgados de mínima varianza es relacionada con la existencia de estadísticos suficientes.

Cuando existe, el estimador de Máxima Verosimilitud tiene algunas propiedades interesantes:

- Generalmente es consistente;
- Es asintoticamente normal;
- No es siempre insesgado, pero lo es asintoticamente;
- Es función de un estadístico suficiente, cuando existe uno;
- Entre todos los estimadores asintoticamente insesgados, tiene la varianza asintoticamente más pequeña (es eficiente).
- Si el E.M.V. es un estadístico suficiente, entonce ses un estadístico suficiente minimal .
- Tiene la propiedad de **invarianza**.

Proposición 4 (Propiedad de Invarianza):

Si $\hat{\theta}$ es el Estimador de Máxima Verosimilitud del parámetro θ y si $g: \Omega \to \Omega$ es biyectiva, entonces $g(\hat{\theta})$ es el Estimador de Máxima Verosimilitud de $g(\theta)$.

Demostración: en efecto si $\tau = g(\theta)$, como g es biyectiva, $\theta = g^{-1}(\tau)$; si $f_n(x_1, ...x_n | \theta) = f_n(x_1, ...x_n | g^{-1}(\tau))$ es máxima para $\hat{\tau}$ tal que $g^{-1}(\hat{\tau}) = \hat{\theta}$. $\hat{\tau}$ es necesariamente el E.M.V. y como g es biyectiva, $\hat{\tau} = g(\hat{\theta})$.

Veremos en el ejemplo 2, que el Estimador de Máxima Verosimilitud de σ se puede obtener directamente o como la raíz del Estimador de Máxima Verosimilitud de σ^2. Eso se debe de la propiedad de **invarianza** del Estimador de Máxima Verosimilitud por transformación funcional biyectiva:

3.5 EJEMPLOS

Ejemplo 1: Una máquina produce diariamente un lote de piezas. Un criterio basado sobre normas de calidad vigente permite clasificar cada pieza fabricada como defectuosa o no defectuosa. El cliente aceptara el lote si la proporción de piezas θ defectuosas contenidas en el lote no sobrepasa el valor θ_o. El fabricante tiene que controlar entonces la proporción θ de piezas defectuosas contenidas en cada lote que fabrica. Pero si la cantidad de piezas N de cada lote es muy grande, no podrá examinar cada una para determinar el valor de θ. El fabricante efectúa entonces el control de calidad de una muestra aleatoria pequeña con *n* piezas. Se define la v.a. *X* que toma el valor 1 si la pieza es defectuosa y 0 en el caso contrario. Sean $X_{1,}X_2,...,X_n$ los valores obtenidos sobre la muestra. Calculamos el Estimador de Máxima Verosimilitud (E.M.V.):

Como $X_i \sim Bernouli(\theta)$ $(0 \le \theta \le 1)$, entonces la función de Verosimilitud es:

$$f_n(x_1,...,x_n|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$\underset{\theta\in\Omega}{Max}\, f_n(x_1,...,x_n|\theta) \Leftrightarrow \underset{\theta\in\Omega}{Max}\, Log(f_n(x_1,...,x_n|\theta))$$

$$Log(f_n(x_1,...,x_n|\theta)) = \sum_i [x_i Log(\theta) + (1-x_i)Log(1-\theta)]$$

$$\frac{\partial Log(f_n)}{\partial\theta} = \frac{\sum x_i}{\theta} - \frac{n-\sum x_i}{1-\theta}$$

Luego el E.M.V. $\hat{\theta}$ de θ es la proporción de piezas defectuosas observada $\frac{\sum x_i}{n}$.

Ejemplo 2: El ministerio de la salud quiere conocer la talla promedia μ de las mujeres chilenas adultas. Si $X_{1,}X_2,...,X_N$ son las tallas de todas las chilenas adultas, $\mu = \sum_{i=1}^{N} X_i / N$. Dado el tamaño grande *N* de esta población, se obtiene la talla de una muestra aleatoria de tamaño pequeño *n.* Sean $Y_{1,}Y_2,...,Y_n \in \{X_{1,}X_2,...,X_N\}$ los valores de las tallas muestrales. Si suponemos que los valores muestrales normales: $Y_i \sim N(\mu,\sigma^2)$ con los parámetros μ y σ^2 desconocidos:

$$f_n(y_1,...,y_n|\mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^n exp\{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\}$$

Luego $Log(f_n)$ es máximo cuando $\mu = \overline{X}_n$ la media muestral y $\sigma^2 = S_n^2$ la varianza muestral. El estimador $\overline{X}_n$ de la media poblacional μ es insesgado y consistente. El estimador S_n^2 de la varianza de la población es sesgado, pero asintoticamente insesgado y consistente. Además el par $(\overline{X}_n, S_n^2)$ es suficiente para (μ,σ^2).

Notas:

- Si se supone la varianza poblacional σ^2 conocida, el E.M.V. de μ sigue igual a la media muestral $\overline{X}_n$.
- Se puede buscar el estimador de la varianza o bien de su raíz σ. El resultado no cambia.

Ejemplo 3: Sea valores muestrales obtenidos de la distribución *Uniforme[0,θ], θ >0* desconocido *:*

$$f_n(x_1,...,x_n|\theta)=\frac{1}{\theta^n} \quad \text{si} \quad 0\le x_i\le\theta \quad (\forall i)$$

Cuando $\theta \ge x_i \ \ (\forall i)$, $f_n(x_1,...,x_n|\theta)$ es no nulo y es decreciente en θ; luego $f_n(x_1,...,x_n|\theta)$ es máxima para el valor más pequeño de θ que hace $f_n(x_1,...,x_n|\theta)$ no nulo: el E.M.V. de θ es entonces $\hat{\theta}=Max\{x_1,x_2,...,x_n\}$.

El método de los momentos produce un estimador bien diferente. En efecto, como $E(X)=\theta/2$, el estimador de los momentos es entonces $\tilde{\theta}=2\overline{X}_n$.

En este ejemplo, una dificultad se presenta cuando se toma el intervalo *]0,θ [* abierto, dado que no se puede tomar como estimador el máximo; en este caso no existe E.M.V. Puede ocurrir que no es único también: si se define el intervalo *[θ ,θ +1]*, la función de verosimilitud es:

$$f_n(x_1,...,x_n|\theta)=1 \quad \text{si} \quad \theta\le x_i\le\theta+1 \quad (\forall i)$$

es decir: $f_n(x_1,...,x_n|\theta)=1$ si $Max\{x_1,...,x_n\}-1\le\theta\le Min\{x_1,...,x_n\}$

Por lo cual todo elemento del intervalo $[Max\{x_1,...,x_n\}-1, Min\{x_1,...,x_n\}]$ es E.M.V. No tenemos unicidad. Aquí el estimador de los momentos, que es igual a $\overline{X}_n - 1/2$, es bien diferente también.

3.6 ESTIMADORES DE BAYES

3.6.1 Distribuciones a priori

En el problema de estimación de un parámetro de una distribución de función de densidad $f(x|\theta)$, es frecuente tener algunas ideas sobre los valores que puede tomar θ; en este caso conviene tomar en cuenta este conocimiento o **creencia** que se puede traducir en una distribución de probabilidad sobre el espacio de parámetros Ω, sea $\pi(\theta)$. Es decir que ahora θ ya no es un parámetro constante, sino una variable aleatoria. Esta distribución no depende de los valores muestrales. Está definida previo al muestreo.

Por ejemplo, en un proceso de fabricación se tiene la proporción θ desconocida de piezas efectuosas. Si no se sabe nada respecto a θ, se puede suponer que todos los valores son equiprobables: $\theta \sim Uniforme(0,1)$. Pero uno puede sospechar que los valores alrededor de 0.10 son más probables; en este caso se podrá tomar una distribución *beta* más concentrada en 0.10.

Definición 10: Se llama distribución a priori a la distribución atribuida a un parámetro poblacional, antes de tomar alguna muestra.

La distribución a priori es siempre totalmente especificada.

3.6.2 Distribuciones a posteriori

Ahora hay que relacionar los valores muestrales con la distribución a priori $\pi(\theta)$. La función de v verosimilitud $f_n(x_1,...,x_n|\theta)$ es ahora una densidad condicional y $h(x_1,...,x_n,\theta) = f_n(x_1,...,x_n|\theta)\pi(\theta)$ es la densidad conjunta de $(x_1,...,x_n,\theta)$. De la cual se puede deducir la distribución condicional de θ dado los valores muestrales $(x_1,...,x_n)$:

Definición 11: La distribución condicional de θ dada la muestra $X_1, X_2,...,X_n$ se llama **distribución a posteriori** y su densidad es igual a $\xi(\theta|x_1,...,x_n) = \dfrac{f_n(x_1,...,x_n|\theta)\pi(\theta)}{g_n(x_1,...,x_n)}$, en que

$g_n(x_1,...,x_n) = \int_{\Omega} h(x_1,...,x_n,\theta)d\theta$ es la densidad marginal de los valores muestrales $(x_1,...,x_n)$.

La distribución a posteriori representa la actualización de la información a priori $\pi(\theta)$ en vista de la información contenida en los valores muestrales, $f_n(x_1,...,x_n|\theta)$. Podemos entonces estudiar esta distribución a posteriori de θ dando la moda, la media, la mediana, la varianza, etc. Un estimador natural en este caso es tomar la moda de $\xi(\theta|x_1,...,x_n)$, que aparece como el máximo de la distribución corregida de θ.

Ejemplo 4: Sean $X \sim Bernoulli(p)$ y la distribución a priori $p \sim Beta(\alpha,\beta)$, con α y β dados.

$$f_n(x_1,...,x_n|p) = p^{n\overline{X}_n}(1-p)^{n-n\overline{X}_n}$$

$$\pi(p) = p^{\alpha-1}(1-p)^{\beta-1} / B(\alpha,\beta) \quad (0 \le p \le 1)$$

en que $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

La densidad a posteriori de *p* es entonces:

$$\xi(p|x_1,...,x_n) = p^{\alpha+n\overline{X}_n-1}(1-p)^{\beta+n-n\overline{X}_n-1} / B(\alpha+n\overline{X}_n, \beta+-n\overline{X}_n)$$

que es la distribución $Beta(\alpha+n\overline{X}_n, \beta+n-n\overline{X}_n)$. El máximo está en la moda de esta distribución, cuando está definida; en este caso es igual a $(\alpha+n\overline{X}_n-1)/(\alpha+\beta+n)$.

Ejemplo 5: Sea $X \sim N(\theta,1)$ y la distribución a priori $\theta \sim N(0,10)$.
$\xi(\theta|x_1,...,x_n) \propto f_n(x_1,...,x_n|\theta)\pi(\theta)$ ($\propto$ designa la proporcionalidad con respecto a θ).

$$\xi(\theta \mid x_1,...,x_n) \propto exp\left(-\frac{\sum_i (x_i - \theta)^2}{2} - \frac{\theta^2}{20} \right) \Rightarrow \xi(\theta \mid x_1,...,x_n) \propto exp\left(n\theta \overline{X}_n - \frac{-11\theta^2}{20} \right) \Rightarrow$$

$$\xi(\theta \mid x_1,...,x_n) \propto exp\left(-\frac{-11}{20}(\theta - (\frac{10}{11} n\overline{X}_n))^2 \right)$$

La distribución a posteriori de θ es entonces $N(\frac{10}{11} n\overline{X}_n, \frac{10}{11})$. La moda de la distribución es igual a la media $\frac{10}{11} n\overline{X}_n$.

3.6.3 Funciones de pérdida

Los métodos de estimación propuestos hasta ahora no toman en cuenta un aspecto importante del problema, que son las consecuencias de tales estimaciones.

Dado que los estimadores son la base de una decisión final, es importante poder comparar los procedimientos que conducen a estas decisiones mediante algún criterio de evaluación, que mide las consecuencias de cada estimación en función de los valores del parámetro θ.

Definición 12: Se llama **función de pérdida** o **función de costo** a la función $L : \Omega \times \Omega \to [0,+\infty[$, en que $L(\theta,\delta)$ es creciente con el error entre el parámetro θ y un estimador δ.

No es siempre fácil definir esta función de pérdida, que es específica de cada problema y puede tener algún aspecto subjetivo (noción de utilidad). Sin embargo, se puede elegir entre diversas funciones de pérdida clásicas, cuando no se puede construir una propia:

- Función de pérdida cuadrática: Es la función de pérdida más utilizada y más criticada:
  $$L(\theta,\delta) = (\theta - \delta)^2$$
  que penaliza demasiado los errores grandes.

- Función de pérdida absoluta: Una solución alternativa a la función cuadrádica es usar el valor absoluto:
  $$L(\theta,\delta) = |\theta - \delta|$$

- o bien una función afín por parte:
  $$L(\theta,\delta) = \begin{cases} k_1(\theta - \delta) & si\ \theta > \delta \\ k_2(\theta - \delta) & si\ \theta \leq \delta \end{cases}$$

- Función de perdida "*0-1*":
  Sea $I_\varepsilon(\delta)$ el intervalo de centro δ y largo 2ε
  $$L(\theta,\delta) = \begin{cases} 0 & si\ \theta \in I_\varepsilon(\delta) \\ 1 & si\ \theta \notin I_\varepsilon(\delta) \end{cases}$$

3.6.4 Estimadores de Bayes

La función de pérdida $L(\theta,\delta)$ es una función de θ considerada como aleatoria con la distribución a posteriori $\xi(\theta \mid x_1,...,x_n)$. Luego es natural de buscar un estimador $\delta(x_1,...,x_n)$ de θ tal que la pérdida promedio sea mínima.

Definición 13: El estimador de Bayes es solución de $\underset{\delta}{Min}\{E[L(\theta,\delta) \mid x_1,...,x_n]\}$

La solución depende de la función de perdida elegida.

- Función de pérdida cuadrática: $L(\theta,\delta)=(\theta-\delta)^2$, el estimador de Bayes es simple de encontrar: $E\{(\theta-\delta)^2 \mid x_1,...,x_n\}$ es mínimo para $\delta(x_1,...,x_n)=E(\theta \mid x_1,...,x_n)$. es decir es la media de la distribución a posteriori ξ.

- Función de pérdida absoluta: $L(\theta,\delta)=|\theta-\delta|$; el estimador de Bayes es la mediana de la distribución a posteriori ξ. Mostramos un resultado más general:

Proposición 5: El estimador de Bayes asociado a la distribución a posteriori ξ y a la función de perdida

$$L(\theta,\delta)=\begin{cases}k_1(\theta-\delta) & si\ \theta>\delta\\ k_2(\theta-\delta) & si\ \theta\le\delta\end{cases}$$

es el fractil $\dfrac{k_1}{k_1+k_2}$ de la distribución a posteriori ξ.

Demostración: Se tiene

$$E[L(\theta,\delta) \mid x_1,...,x_n] = k_2\int_{-\infty}^{\delta}(\delta-\theta)\xi(\theta \mid x_1,...,x_n)d\theta + k_2\int_{\delta}^{+\infty}(\theta-\delta)\xi(\theta \mid x_1,...,x_n)d\theta$$

Derivando con respecto a δ, se obtiene:

$$k_2P(\theta<\delta \mid x_1,...,x_n)-k_1P(\theta>\delta \mid x_1,...,x_n)=0$$

Es decir: $$P(\theta<\delta \mid x_1,...,x_n)=\frac{k_1}{k_1+k_2}$$

o sea δ es el fráctil de orden $\dfrac{k_1}{k_1+k_2}$. En particular si $k_1=k_2$, se obtiene la mediana de la distribución a posteriori de θ.

- Función de pérdida "0-1": $E[L(\theta,\delta) \mid x_1,...,x_n]$ es mínimo cuando $\int_{I_\varepsilon(\delta)}\xi(\theta \mid x_1,...,x_n)d\theta$ es máximo.

  Si $\varepsilon\to 0$, entonces $E[L(\theta,\delta) \mid x_1,...,x_n]$ es mínimo cuando $\xi(\theta \mid x_1,...,x_n)$ es máximo. El estimador de Bayes es la moda de la distribución a posteriori ξ.

Nota: Como un E.M.V., si un estimador de Bayes es un estadístico suficiente, entonces es un estadístico suficiente minimal.

3.6.5 Estimadores de Bayes para muestras grandes

Se muestra aquí, a través de un ejemplo, los efectos de la distribución a priori y de la función de pérdida sobre el estimador de Bayes, para muestras grandes. Sea θ la proporción de defectuosos. Tomamos dos distribuciones a priori y dos funciones de pérdida:
$\pi_1(\theta)=1$ para $\theta\in[0,1]$ y $\pi_2(\theta)=2(1-\theta)$ para $\theta\in[0,1]$.

$L_1(\theta,\delta)=(\theta-\delta)^2$ y $L_2(\theta,\delta)=|\theta-\delta|$. Para la función de perdida quadrática, las distribuciones a posteriori son respectivamente

$$\xi_1(\theta\mid x_1,...,x_n)\propto\theta^{n\overline{X}_n}(1-\theta)^{n-n\overline{X}_n}$$

que es una distribución $Beta(1+n\overline{X}_n, n+1-n\overline{X}_n)$

$$\xi_2(\theta\mid x_1,...,x_n)\propto\theta^{n\overline{X}_n}(1-\theta)^{n-n\overline{X}_n+1}$$

que es una distribución $Beta(1+n\overline{X}_n, n+2-n\overline{X}_n)$.

Los estimadores de Bayes para la pérdida cuadrática son las respectivas esperanzas de las distribuciones ξ_1 y ξ_2:

$$\delta_1=\frac{1+n\overline{X}_n}{n+2} \quad \text{y} \quad \delta_2=\frac{1+n\overline{X}_n}{n+3}$$

Los estimadores de Bayes para la pérdida absoluta son las respectivas medianas de las distribuciones ξ_1 y ξ_2, que se obtienen resolviendo la ecuación:

$$K\int_0^{\delta}\theta_{\alpha-1}(1-\theta)^{\beta-1}d\theta=0.5$$

en que $\alpha=1+n\overline{X}_n$ y $\beta=n+1-n\overline{X}_n$ para ξ_1 y $\alpha=1+n\overline{X}_n$ y $\beta=n+2-n\overline{X}_n$ para ξ_2.

Si $n=100$ y $n\overline{X}_n=10$ entonces $\delta_1=11/102=0.108$ y $\delta_2=11/103=0.107$ para la pérdida cuadrática. Se observara cómo la muestra corrige la distribución a priori, con las medias a priori $E(\theta)=0.5$ con ξ_1 y $E(\theta)=1/3$ con ξ_2.

Encontramos ambos estimadores de Bayes a posteriori muy cercanos con $n=100$ y cercanos de la media muestral $\overline{X}_n=10/100=0.100$.

En este ejemplo observamos que el estimador de Bayes cuadrático es consistente. No se puede siempre asegurar que el estimador de Bayes es consistente, pero bajo condiciones bastante generales es cierto.

3.7 EJERCICIOS

1. Sea X_i, ($i=1,...,n$) una muestra aleatoria simple de una v.a. X de función de distribución $Gamma(\alpha,\beta)$.

Estime *E(X)* por Máxima Verosimilitud. Muestre que el estimador resultante es insesgado, convergente en media cuadrática y es consistente.

2. Sea una m.a.s. $\{ X_1, X_2, ... X_n \}$ de una v.a. X de función de densidad $f(x|\theta) = \theta\, x^{\theta-1} \quad \forall x \in [0,1]$. Encuentre el estimador de Máxima Verosimilitud $\hat{\theta}$ de θ y pruebe que $\hat{\theta}$ es consistente y asintóticamente insesgado.

3. Sea Y una v.a. de Bernoulli de parámetro θ. Considere una m.a.s. $\{ Y_1, ..., Y_n \}$ y una distribución a priori *Beta(a,b)* para θ. Obtenga el estimador de Bayes, $\hat{\theta}$ para θ, usando una función de pérdida cuadrática. Muestre que $\hat{\theta}$ es sesgado, asintóticamente insesgado, convergente en media cuadrática y consistente.

4.. Sean dos preguntas complementarias: *Q="vota por Pedro"* y *Q*="no vota por Pedro"*. Se obtiene una m.a.s. de *n* personas que contestan a la pregunta *Q* o *Q**; lo único que se sabe es que cada persona ha contestado a *Q* con probabilidad θ conocida y *Q** con probabilidad $1-\theta$. Se definen:

- *p*: la probabilidad que una persona contesta "SI" a la pregunta (*Q o Q**);
- π : la proporción desconocida de votos para Pedro en la población.

a) Dé la proporción π en función de *p* y θ.
b) Dé el estimador de Máxima Verosimilitud de p y deduzca un estimador $\hat{\pi}$ para π. Calcule la esperanza y la varianza de $\hat{\pi}$.
c) Estudie las propiedades de $\hat{\pi}$; estudie en particular la varianza $\hat{\pi}$ cuando $\theta = 0.5$.

5. Suponga que X tiene una función de densidad $f(x|\theta)$ y que $T(X_1, ... X_n)$ es un estimador de Bayes insesgado para θ con la función de pérdida cuadrática y una distribución a priori $\pi(\theta)$.

a) Demuestre que $E[(\theta - T(X_1, ... X_n))^2] = 0$

b) Asuma que $f(x|\theta)$ es una $N(\theta,1)$. Pruebe que $E[(\theta - \overline{X}_n)^2] = 1/n$. Concluya si $\overline{X}_n$ puede ser un estimador de Bayes para pérdida cuadrática.

6. Sea $\{ X_1, X_2, ..., X_n \}$ m.a.s. de una distribución tal que $P(X_i \in [a,b]) = \theta$.

Se define $Y_i = \begin{cases} 1 & si\ X_i \in [a,b] \\ 0 & sino \end{cases}$

a) Dé la distribución de Y_i.

b) Dé el estimador de máxima verosimilitud $\hat{\theta}$ de θ.

c) Dé la esperanza y la varianza de $\hat{\theta}$.

d) Sean las distribuciones a priori de θ :

- $\pi_1(\theta) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)\beta - 1, \quad \theta \in [0,1]$ (distribución $Beta(\alpha,\beta)$);
- $\pi_2(\theta) = 2(1-\theta), \quad \theta \in [0,1]$.

Dé los estimadores de Bayes y sus varianzas cuando se usa una función de pérdida cuadrática.
e) Aplicación numérica: dé las soluciones a las preguntas anteriores con los valores: $n=10$, $\alpha = 2$, $\beta = 2$; X_i : *1.2, 3.5, 2.4, 1.5, 6.3, 2.8, 4.2, 4.5, 3.8, 5.1* y *[a,b]=[2,4]*.

7. Sea $\theta \in [0,1]$ una m.a.s. de una v.a. *X* con función de densidad $f(x|\theta)$. Sea $Y = \delta(X_1,...,X_n)$ un estimador de θ. Se define Y_{-i} al estimador δ calculado sobre la muestra salvo la observación *i (i=1,2,...,n)*, e $\hat{Y}_{-i} = nY - (n-1)Y_{-i}$ y $\hat{Y} = \frac{1}{n}\sum_i \hat{Y}_{-i}$.

a) Calcule la varianza S^2 de $\hat{Y}$ cuando $Y = \overline{X}_n$ la media muestral y $E(X) = \theta$.

b) Deducir la distribución de $\hat{Y} - \theta$ cuando $Y = \overline{X}_n$ y $X \sim N(\theta, \sigma^2)$.

8. Sea *X* una v.a. real con densidad $f(x|\theta)$, $\theta \in \Omega = \{\theta_1, \theta_2, ..., \theta_N\}$ (Ω finito).
Sean π una distribución de probabilidad a priori sobre Ω y la función de pérdida:

$$L(\theta, \delta) = \begin{cases} 0 & si\ \theta = \delta \\ c & si\ \theta \neq \delta \end{cases}$$

con *c>0.*

a) Pruebe que la pérdida esperada se escribe como $E(L(\theta,\delta)) = c(1 - \xi(\delta|x))$, en donde ξ es la distribución a posteriori sobre Ω.
b) Deduzca la condición que debe satisfacer δ para ser el estimador de Bayes de θ asociado a π. Pruebe que el estimador no depende de *c*.
c) Si π es la distribución uniforme sobre Ω, pruebe que el estimador de Bayes de θ y el estimador de máxima verosimilitud coinciden.

9. Se considera la distribución discreta:
$P(X = x) = a_x \frac{\theta^x}{h(\theta)}$, con *x=0,1,2,...*, y en donde *h* es diferenciable y a_x puede ser nulo para algunos *x*.
Sea $X_1, X_2, ..., X_n$ una m.a.s. de esta distribución.

a) Dé las expresiones de $h(\theta)$ y $h'(\theta)$.
b) Dé el estimador de máxima verosimilitud de θ en función de *h* y *h'*.
c) Muestre que el estimador de máxima verosimilitud es el mismo que el del método de los momentos.
d) Aplique lo anterior para los casos siguientes:
i. $X \sim Binomial(N, p)$ (*N* conocido);
ii. $X \sim Poisson(\lambda)$.

10. Sean T_i, *i=1,..., I* estimadores del parámetro θ tales que : $E(T_i) = \theta + b_i$, $b_i \in \Re$. Se define un nuevo estimador *T* de θ como $T = \sum_{i=1}^{I} \lambda_i T_i$.
a) Dé una condición sobre los λ_i para que *T* sea insesgado.
b) Suponga que $b_i = 0$ $(\forall i)$ (estimadores insesgados). Plantee el problema de encontrar los coeficientes λ_i para que la varianza de *T* sea mínima.
c) Suponiendo que los T_i son no correlacionados , resuelva el problema planteado antes.
d) Sean X_{ij}, *i=1,2,…,M, j=1,2,…,n_i*, *M* m.a.s. independientes entre si, de variables aleatorias X^i con distribuciones normales de varianza común σ^2.

Sea $S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2$, el estimador insesgado de la varianza calculado en la muestra *i*.

Sea $S^2 = \frac{1}{\sum_{i=1}^{M} n_i - M}\sum_{i=1}^{M}(n_i - 1)S_i^2$

Demuestre que S^2 es el estimador lineal insesgado de varianza mínima para σ^2.