ÍNDICE

CAPÍTULO 3
ESTIMACIÓN

El problema de la estimación consiste en realizar una inferencia sobre el valor de los parámetros de un modelo estadístico a partir de los datos contenidos en una muestra. Existen tres métodos fundamentales: método de los momentos, mínimos cuadrados y máxima verosimilitud. Este curso trata fundamentalmente del método de máxima verosimilitud.

En general, un parámetro cualquiera se va a designar por θ y su correspondiente estimador por $\hat{\theta}$. El estimador $\hat{\theta}$ es una función de los datos muestrales, cuyo valor depende de la muestra concreta. Por tanto, el parámetro θ toma un valor fijo mientras que el estimador $\hat{\theta}$ es aleatorio, su valor depende de la muestra.

3.1. Estimación por el método de los momentos

Supongamos que se desean estimar k parámetros $(\theta_1,\theta_2,...,\theta_k)$ de una distribución $f(X;\theta_1,\theta_2,...,\theta_k)$. El método de los momentos consiste en buscar los valores de los parámetros que igualan los k primeros momentos con relación al origen en la muestra y en la población. Por tanto, no se utiliza la distribución completa para obtener estimadores sino únicamente los momentos indicados.

En la población el momento de orden i se ha definido:

$$\alpha_i = \int_{-\infty}^{\infty} X^i f(X) dX$$

el correspondiente momento muestral es:

$$a_i = \frac{\sum_{i=1}^{n} X^i}{n}$$

De esta forma se obtiene el siguiente sistema de *k* ecuaciones con *k* incógnitas:

$$\left.\begin{array}{rcl} \alpha_1 & = & a_1 \\ \alpha_2 & = & a_2 \\ & \vdots & \\ \alpha_k & = & a_k \end{array}\right\}$$

Los valores de los parámetros que satisfacen el sistema de ecuaciones constituyen los estimadores ($\hat{\theta}_1$, $\hat{\theta}_2$,..., $\hat{\theta}_k$).

Ejemplo 3.1. (Distribución de Poisson).

Se obtiene una muestra X_1,..., X_n *de una distribución de Poisson. Estimar* λ *por el método de los momentos.*

En esta distribución sólo hay un parámetro que además cumple $\alpha = E(X) = \lambda$. *El primer momento muestral es* $a_1 = \bar{X}$. *El estimador es el valor que resuelve:* $\alpha_1 = a_1$, *es decir* $\hat{\lambda} = \bar{X}$.

Ejemplo 3.2. (Regresión logística).

Sea Y una variable dicotómica cuya distribución depende de una variable independiente X . Se desea estimar el parámetro β *de la regresión logística:*

$$f(Y_i \mid X) = \frac{\exp(Y\beta X)}{1 + \exp(\beta X)}$$

Dado que Y es dicotómica, para cada valor de X el primer momento poblacional es su valor esperado:

$$E(Y \mid X) = 0 \times f(Y = 0 \mid X) + 1 \times f(Y = 1 \mid X) = f(Y = 1 \mid X)$$

El correspondiente momento muestral es $\overline{Y}(X)$*, que representa la media de Y para todas las observaciones en las que X toma un valor determinado. Por tanto, si X toma los valores* x_1, x_2,..., x_n *para cada uno de ellos se dispone de un momento poblacional:* $E(Y|x_i)$, $E(Y|x_2)$,...,$E(Y|x_n)$ *y un momento muestral:* $\overline{Y}(x_i)$, $\overline{Y}(x_2)$,...,$\overline{Y}(x_n)$*. El estimador de* β *es el valor que iguala los momentos en la población y la muestra, es decir:*

$$\sum_{i=1}^{n} \frac{\exp(\beta x_i)}{1 + \exp(\beta x_i)} = \sum_{i=1}^{n} \overline{Y}(x_i)$$

En esta ecuación no es posible despejar β*. Para resolverla pueden utilizarse los denominados métodos numéricos, por ejemplo el de la bisección, Newton-Raphson, etc.*

3.2. Estimación por mínimos cuadrados

Consiste en asignar a los parámetros aquel valor que minimice la diferencia al cuadrado entre los datos observados y los predichos por el modelo estadístico.

El método puede entenderse con un ejemplo. Supongamos que Y es una variable aleatoria. Se desea estimar la regresión lineal de Y sobre un predictor X, es decir para un sujeto i: $Y_i = \alpha + \beta X_i + E_i$, donde E_i es un error aleatorio con media 0 y varianza σ^2 para todos los valores de X.

Los parámetros del modelo son α y β. Para cada valor de X el valor predicho de Y es: $E(Y) = \alpha + \beta X$, por tanto para estimar los parámetros se minimiza la diferencia cuadrática entre los datos observados y las predicciones del modelo. Si el tamaño muestral es n la diferencia es:

$$d(\alpha, \beta) = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2$$

Los estimadores son aquellos valores que minimizan la función $d(\alpha,\beta)$, es decir aquellos valores que hagan que su primera derivada sea 0:

$$\frac{\partial}{\partial\alpha} d(\alpha,\beta) = -2\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i) = 0$$

$$\frac{\partial}{\partial\beta} f(\alpha,\beta) = -2\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)X_i = 0$$

Desarrollando estas ecuaciones se encuentran los estimadores descritos en cursos anteriores (Pardo y San Martin, 1999): $\alpha = \overline{Y} - \hat{\beta}\overline{X}$ y $\hat{\beta} = Cov(X, Y)/Std(X)Std(Y)$.

3.3. Estimación por máxima verosimilitud

El método de máxima verosimilitud consiste en asignar a los parámetros aquel valor que haga máxima la probabilidad de las datos observados. A diferencia de los anteriores utiliza la distribución completa de probabilidad.

En caso de que la distribución de la variable sea normal los estimadores de mínimos cuadrados y máxima verosimilitud coinciden. Continuando con el ejemplo anterior, la distribución de Y es:

$$f(Y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{Y_i - \alpha - \beta X_i}{\sigma^2}\right)^2\right)$$

Por tanto la función de densidad de la muestra compuesta por n observaciones es:

$$L(\alpha,\beta) = \prod_{i=1}^{n} f(Y_i) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(Y_i - \alpha - \beta X_i)^2}{\sigma^2}\right)$$

Los estimadores máxima verosímiles son los valores que maximizan la función de verosimilitud $L(\alpha, \beta)$, es decir la función de densidad de la muestra. Analizando la forma de esta función

puede verse que estos valores son aquellos que minimizan el término en el exponente:

$$\sum_{i=1}^{n} \frac{(Y_i - \alpha - \beta X_i)^2}{\sigma^2}$$

La forma de este término es la misma que la de la función $d(\alpha,\beta)$ comentada en relación con la estimación por mínimos cuadrados. Por tanto, los estimadores obtenidos por ambos métodos coinciden. Esto es una consecuencia de haber asumido que la distribución de Y es normal. Con cualquier otra distribución la forma de proceder sería la misma, obtener la función de verosimilitud y maximizarla con respecto a los parámetros.

Ejemplo 3.3. (Distribución de Bernoulli)

Supongamos que a un sujeto se le presenta 25 veces una determinada tarea. El resultado de cada presentación se clasifica como éxito o fracaso y se considera que la probabilidad de éxito $\neq$ *no cambia a lo largo de las presentaciones. ¿Cual es la probabilidad de éxito asumiendo independencia entre distintas presentaciones?*

La variable X_i *describe el resultado de la presentación i y sigue la distribución de Bernoulli:*

$$f(X_i;\pi) = \pi^{X_i}(1 \pm \pi)^{(1 \pm X_i)}$$

La función de probabilidad del vector de resultados de las 25 presentaciones tiene la forma:

$$f(X;\pi) = \prod_{i=1}^{25} \pi^{X_i}(1 \pm \pi)^{(1 \pm X_i)}$$

Supongamos que el numero de éxitos se indica por z, siendo $z = \Sigma_{i=1}^{25} X_i$. *Entonces la función de verosimilitud es:*

$$L(\pi) = \pi^{z}(1 \pm \pi)^{(25 \pm z)}$$

En la práctica se trabaja con el logaritmo de la función de verosimilitud por mayor sencillez matemática:

$$\log L(\pi) = z \log \pi + (25 - z) \log(1 - \pi)$$

El estimador máximo verosímil es el valor que maximiza log $L(\pi)$*. Un criterio necesario, pero no suficiente, para que a un valor concreto de* π *le corresponda un máximo de log* $L(\pi)$ *es que anule su primera derivada. En el ejemplo:*

$$\frac{\partial}{\partial \pi} \log L(\pi) = \frac{z}{\pi} - \frac{25 - z}{1 - \pi}$$
$$= 0$$

La ecuación anterior se denomina ecuación de estimación. Su solución es el estimador máximo verosímil de π*:*

$$\frac{z}{n} = \frac{25 - z}{1 - \pi}$$
$$z - z\pi = 25\pi - z\pi$$
$$\hat{\pi} = \frac{z}{25}$$

En realidad el criterio de la primera derivada no es suficiente para determinar que $\hat{\pi}$ *es un estimador máximo verosímil porque dicha derivada se anula tanto si la función tiene un máximo en* $\hat{\pi}$ *cómo si tiene un mínimo. En caso de que la función tenga un máximo se cumple que su segunda derivada es negativa. En el ejemplo:*

$$\frac{\partial}{\partial \pi^2} \log L(\pi) = -\frac{z}{\pi^2} - \frac{25 - z}{(1 - \pi)^2}$$

La cual es necesariamente menor que 0, por lo que efectivamente $\hat{\pi}$ *es un máximo de log* $L(\pi)$*.*

Ejemplo 3.4.

Continuando con el mismo ejemplo, supongamos que se relaja el supuesto de independencia y se asume únicamente independencia condicional. Supongamos que la probabilidad de éxito en la tarea i depende del número de aciertos y errores cometidos en las i - 1 tareas anteriores. El número de aciertos anteriores es $C_i=\Sigma_{j=1}^{i-1}X_j$ *y el de errores* $D_i=\Sigma_{j=1}^{i-1}(1-X_j)$*. Podría plantearse el modelo:*

$$\pi_i = \frac{\exp(\alpha + \beta C_i + \delta D_i)}{1 + \exp(\alpha + \beta C_i + \delta D_i)}$$

La cantidad π_i *debe interpretarse como la probabilidad del ensayo i condicionada en los ensayos anteriores. La función de verosimilitud se obtiene a partir de la regla del producto y tiene la forma:*

$$L(\alpha, \beta, \delta) = \prod_{i=1}^{25} \pi_i^{X_i} (1 - \pi_i)^{(1-X_i)}$$

3.4. Estimación bayesiana

La estimación bayesiana se caracteriza porque permite incorporar las expectativas del investigador acerca de los valores de los parámetros. La estadística bayesiana se diferencia la denominada estadística frecuentista en dos aspectos importantes:

- En la estadística bayesiana tanto los datos como los parámetros son cantidades aleatorias. En los métodos anteriores los parámetros se consideran cantidades fijas.
- Al ser los parámetros cantidades aleatorias siguen una distribución, denominada *distribución previa,* que expresa las expectativas del investigador en ausencia de ningún dato.

En los métodos frecuentistas el valor que toman las estimadores depende únicamente de cuales hayan sido los datos ob-

servados. Por el contrario, en los métodos bayesianos dichas valores dependen de los datos y también de las expectativas previas. Por esta razón existen distintas escuelas en el seno de la estadística, dependiendo de sí se considera legítimo o no utilizar distribuciones previas. Lo cierto es que desde un punto de vista aplicado los métodos bayesianos pueden ser de gran utilidad para determinados problemas. Además, a medida que el tamaño muestral aumenta los estimadores dependen mas de los datos y menos de las distribuciones previas, por lo que en el límite $n \to \infty$ el estimador máximo verosímil y el bayesiano coinciden.

A modo de ejemplo, consideremos el problema de estimar la media de una distribución normal $f(Y|\mu)$ con varianza conocida a partir de n réplicas de un experimento aleatorio. El estimador máximo verosímil es aquel que maximiza la función de verosimilitud:

$$f(Y \mid \mu) = \prod_{i=1}^{n} f(Y_i \mid \mu)$$

Supongamos que la variable en cuestión mide el cociente intelectual de los sujetas. Se sabe que en la población general su distribución es normal (100, 16). Por tanto, el investigador puede suponer que en la subpoblación con la que esta trabajando el valor de la media oscilara en torno a 100, siendo muy inesperado encontrar medias, digamos, superiores a 150 o inferiores a 50. Esta expectativa se formaliza mediante la definición de una distribución previa. Por ejemplo, la distribución previa de μ puede ser la normal (100, 20), denominada $f(\mu)$. Es importante advertir que esta distribución previa no depende de ningún dato observado en el experimento actual sino que se define arbitrariamente.

En la estadística bayesiana un objetivo es modificar las expectativas previas del investigador de acuerdo con la evidencia encontrada en la muestra. Según el teorema de Bayes la distribución posterior de μ puede obtenerse del siguiente

modo, a partir de la función de verosimilitud y la distribución previa:

$$f(\mu \mid Y) = \frac{f(Y \mid \mu) f(\mu)}{f(Y)}$$

Siendo $f(\mathbf{Y})$ la distribución marginal de **Y,** es decir: $f(\mathbf{Y}) = \int f(\mathbf{Y} \mid \mu)\, f(\mu) d\mu$. El estimador bayesiano de μ es la distribución posterior $f(\mu \mid \mathbf{Y})$. Una descripción mas completa de la estadística bayesiana aparece en el capítulo 7.

3.5. Propiedades de los estimadores puntuales

Hasta el momento se han descrito varios métodos mas o menos intuitivos para obtener estimadores. Puede plantearse la cuestión de en qué sentido son unos mejores que otros. Para verificarlo se han descrito una serie de propiedades que deberían cumplir los estimadores. Para un valor fijo de θ se define la distribución muestral del estimador $\hat{\theta}$, de la cual dependen las propiedades de dicho estimador:

1. *Insesgado.* Un estimador es insesgado cuando cumple:

$$E(\hat{\theta}) = \theta$$

para todo valor de θ en el espacio paramétrico. A la cantidad $b(\hat{\theta}) = E(\hat{\theta}) - \theta$ se le denomina sesgo del estimador.

Ejemplo 3.5.

Según se ha estudiado en cursos anteriores, para estimar la media poblacional μ se utiliza la media muestral: $\overline{X} = \Sigma_{i=1}^{n} X_i/n$. Se trata de un estimador insesgado dado que $E(\overline{X}) = \mu$.

2. *Consistente.* Un estimador es consistente si para cualquier $\in > 0$:

$$\lim_{n\to\infty} P(|\theta - \hat{\theta}| > \in) = 0$$

Un estimador insesgado y que cumpla $Var(\hat{\theta}) \to 0$ cuando $n \to \infty$ se dice que es consistente.

Ejemplo 3.6.

La varianza de la media muestral es σ^2/n. *Esta cantidad tiende a 0 por lo que la media muestral es un estimador consistente de* μ.

3. *Eficiente.* Un estimador es eficiente en caso de que sea insesgado y su varianza mínima. Para estimar un parámetro pueden utilizarse distintos estimadores con diferente varianza. Par tanto es deseable utilizar aquel que tenga menos dispersión.

Las diferencias entre el parámetro y el estimador obtenido en las posibles muestras pueden cuantificarse mediante el error cuadrático medio, cuyo valor depende de las tres propiedades anteriores. Se conoce como $MSE(\hat{\theta})$ *(Mean squared error o error cuadrático medio)* a la cantidad $MSE(\hat{\theta}) = E((\hat{\theta}-\theta)^2)$. Desarrollando esta expresión se llega a:

$$\begin{aligned} MSE(\hat{\theta}) &= E((\hat{\theta} - \theta)^2) \\ &= E(([\hat{\theta} - E(\hat{\theta})] - [E(\hat{\theta}) - \theta])^2) \\ &= E((\hat{\theta} - E(\hat{\theta}))^2) + E((E(\hat{\theta}) - \theta)^2) - 2E((\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)) \\ &= Var(\hat{\theta}) + b(\hat{\theta})^2 \end{aligned}$$

El *MSE* es la suma de la varianza del estimador mas su sesgo al cuadrada y muestra la relación entre estas tres cantidades.