Clase 5: Métodos de estimación

Matías Carrasco

6 de octubre de 2019

Índice

1. El modelo de población

Este modelo se basa en la noción de población y muestreo. Imaginemos una población muy grande, cuyos individuos son entidades generales, como por ejemplo personas, animales, plantas, lapiceras, autos, etc. No es importante la naturaleza de los individuos, lo que importa es que estamos interesados en una o varias características de los mismos, que podemos medir experimentalmente. Llamemos X a la característica de interés de un individuo elegido al azar en la población. Entonces X es una variable aleatoria.

Una parte muy importante del modelo consiste en precisar qué tipo de distribución tiene X. Para empezar, debemos definir si X es una variable discreta o continua. Si elegimos un modelo continuo, podemos elegir distintos tipos de distribución, por ejemplo, podemos suponer que X es normal, o que es exponencial, etc. La población queda entonces definida una vez que elegimos un modelo de distribución para X.

Población
Una población queda definida por una o varias variables aleatorias y su distribución.

Discutamos un ejemplo concreto. Supongamos que trabajamos para una marca de refrescos y estamos encargados de la fabricación de las tapitas. Se nos encomienda la tarea de controlar el diámetro de las tapitas, muy chicas no entran en el pico de la botella, muy grandes no tapan.

La población consiste de las tapitas, y la característica que nos interesa es su diámetro. Sea entonces X el diámetro de una tapita elegida al azar. Para definir la población desde el punto de vista estadístico, debemos elegir un tipo de distribución para X. En este caso lo más natural es suponer que X es una variable continua, por lo que su distribución queda especificada por una densidad de probabilidad $p(x)$.

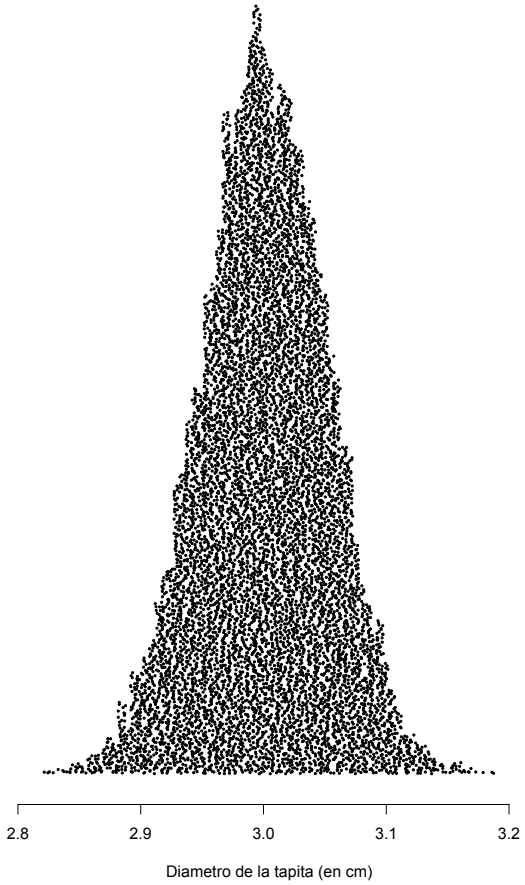

Figura 1: Población de tapitas en función de su diámetro. Cada punto representa una tapita de la población.

Podemos imaginar que ordenamos a las tapitas en una linea según su diámetro. Entonces $p(x)$ indica el amontonamiento de tapitas en la posición x. Ver la Figura 1. Por supuesto, la densidad $p(x)$ es desconocida. Un método de simplificación muy usado es imponer restricciones a $p(x)$ para que sea más fácil obtener información sobre ella.

Un *modelo paramétrico* consiste en suponer que la fórmula de $p(x)$ es conocida, excepto por algunos parámetros. Por ejemplo, parece razonable suponer que $p(x)$ es una densidad normal, de la cuál no conocemos los "verdaderos" parámetros μ (la media) y σ^2 (la varianza). Ver la Figura 2.

Parámetros

En un modelo paramétrico los parámetros determinan la densidad de la población $p(x)$. Por supuesto, son desconocidos.

Denotaremos un parámetro general por la letra θ, y la densidad correspondiente por $p(x;\theta)$. Si el modelo es discreto $p(x;\theta)$ denota la función de probabilidad puntual.

¿Cuál es el verdadero diámetro de las tapitas? Según nuestro modelo normal, el verdadero

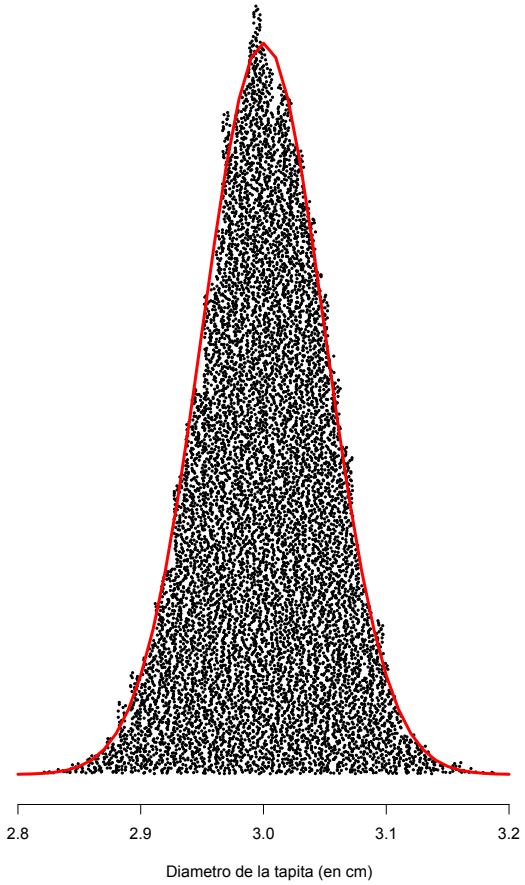

Figura 2: Población de tapitas en función de su diámetro. El modelo paramétrico consiste en suponer que la densidad $p(x)$ es normal de parámetros μ y σ^2. Una densidad normal se muestra en rojo.

diámetro medio de las tapitas es μ, pero esto no debe confundirnos. No existe un verdadero diámetro de las tapitas, μ es un parámetro inventado por nosotros, al igual que lo es la densidad $p(x)$. Son parte de nuestro modelo, y como todo modelo son simplificaciones, no son la realidad y no hay nada de verdadero en ellos. ¿Qué representa μ en la realidad? Nada, es un artilugio intelectual.

2. Muestreo aleatorio

Como no conocemos los parámetros de la población, el problema central de este capítulo es el de cómo estimarlos. Para esto se utiliza el método de *muestreo aleatorio*. Es muy importante distinguir la noción de *muestra* de la noción de *muestreo*.

Muestra observada

Una muestra es un subconjunto de observaciones seleccionadas de una población. En general, denotamos una muestra de tamño n por letras minúsculas: $x_1, \ldots, x_n$.

De este modo, no tiene sentido hablar de muestra aleatoria. Lo que es aleatorio es el muestreo, que es un procedimiento para obtener muestras.

En la Figura 3 se indica con puntos rojos una muestra de tamaño 10 de la población. La misma consiste de las siguientes mediciones

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
3.01	2.98	3.03	3.01	3.02	2.95	2.97	2.93	2.97	2.94

La muestra tiene un análogo de μ, que es el promedio $\bar{x} = 2.98$, y un análogo para σ, que es el desvío estándar $\hat{\sigma} = 0.034$[1]. Estos no son los valores verdaderos de μ y σ, pero nos gustaría usarlos como estimaciones de estos.

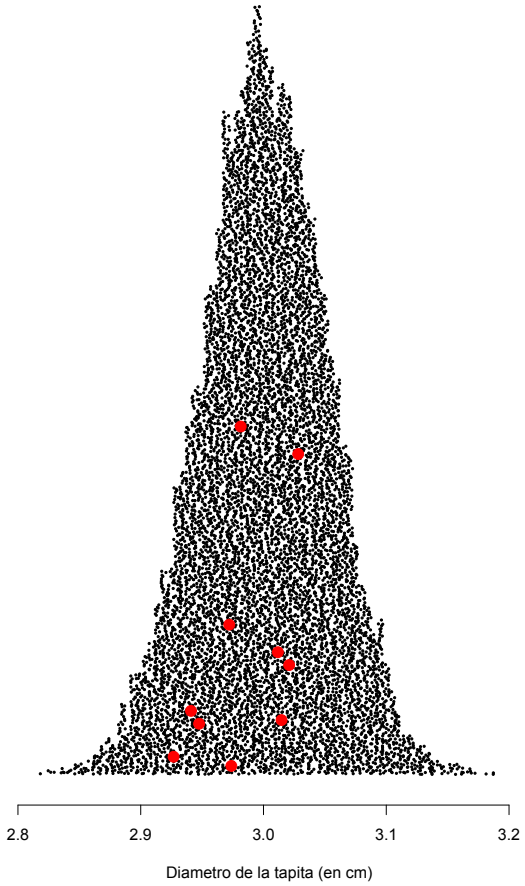

Figura 3: Una muestra de tamaño 10 de la población. El promedio observado en la muestra es $\bar{x} = 2.98$ y el desvío es $\hat{\sigma} = 0.034$ (o $s = 0.035$).

[1]Recordar que hay dos formas de estimar el desvío: una dividiendo entre n que denotamos por $\hat{\sigma}$, y otra dividiendo entre $n-1$ que denotamos s. En este ejemplo $s = 0.035$.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	$\bar{x}$	$\hat{\sigma}$
3.01	2.98	3.03	3.01	3.02	2.95	2.97	2.93	2.97	2.94	2.98	0.034
2.98	2.99	2.89	3.10	3.04	2.92	2.92	2.99	3.03	2.97	2.98	0.059
2.97	2.92	2.98	2.93	3.05	2.94	3.08	2.92	3.01	3.05	2.99	0.057
3.01	2.96	3.04	3.08	2.95	3.00	2.94	2.95	3.03	3.05	3.00	0.048
2.98	3.05	2.98	3.00	2.95	3.02	2.99	2.88	3.00	2.98	2.98	0.041
3.01	3.05	3.03	3.03	2.95	3.03	2.99	2.94	3.07	3.04	3.01	0.040
3.03	2.98	2.94	2.93	2.99	3.03	3.02	3.06	2.94	3.00	2.99	0.041
3.05	2.97	3.02	2.97	2.99	2.93	3.03	3.02	3.05	3.07	3.01	0.043
2.95	3.00	3.00	2.89	3.02	2.99	2.94	2.99	2.93	2.93	2.96	0.039
3.01	3.05	3.06	3.01	3.05	3.05	2.98	3.04	2.92	2.91	3.01	0.052
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	$\overline{X}_{10}$	S_{10}

Tabla 1: Varias muestras diferentes, todas de tamaño 10, de la población de tapitas. Cada muestra tiene su propio promedio y desvío. La variable aleatoria X_i representa el i-ésimo valor en el muestreo aleatorio.

El promedio $\bar{x}$ y el desvío $\hat{\sigma}$ cambian si cambiamos la muestra. La Tabla 1 representa varias muestras, junto con el correspondiente promedio $\bar{x}$ y desvío $\hat{\sigma}$.

Una muestra es una realización concreta de un muestreo aleatorio. Si elegimos al azar n individuos de la población, con reposición, obtenemos así n variables aleatorias

$$X_1, X_2, \ldots, X_n$$

que corresponden al valor de la variable de interés para el primer elemento de la muestra, para el segundo, y así sucesivamente hasta el n-ésimo.

¿Qué restricciones debemos imponer sobre las X_i? Para empezar, si el muestreo es aleatorio, al observar solamente los valores del i-ésimo elemento de la muestra es como observar a la propia variable X. Esto quiere decir que las variables X_i tienen todas la misma distribución que X.

Por otro lado, como el muestreo es con reposición[2], las variables X_i son independientes.

Muestreo aleatorio

Un muestreo aleatorio de X de tamaño n consiste de n variables

$$X_1, X_2, \ldots, X_n$$

independientes y con la misma distribución que X.

Esto lo resumimos escribiendo

$$X_1, X_2, \ldots, X_n \text{ i.i.d.} \sim p_\theta(x),$$

en donde i.i.d. significa independientes e idénticamente distribuidas.

[2]Esto no es importante cuando la población es grande respecto al tamaño de la muestra.

De este modo, el promedio muestral

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

y los desvíos muestrales

$$\Sigma_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2}, \quad S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2}$$

son variables aleatorias. Por eso las escribimos en mayúscula.

La situación acá es muy parecida a la que teníamos en los tests de permutaciones. La realidad observada corresponde a la muestra observada $x_1, \ldots, x_n$. Estos son los valores observados de $X_1, \ldots, X_n$:

$$(X_1)_{\text{obs}} = x_1,\ (X_2)_{\text{obs}} = x_2,\ \ldots,\ (X_n)_{\text{obs}} = x_n.$$

Pero estos últimos corresponden a todas las realidades hipotéticas que no observamos, pero hubiéramos podido observar. De este modo, las variables $X_1, \ldots, X_n$ corresponden a todas las muestras posibles.

Una vez observada una muestra, obtenemos valores observados del promedio y del desvío.

	Hipotético	Observado
Muestra	$X_1, \ldots, X_n$	$(X_1)_{\text{obs}} = x_1,\ \ldots,\ (X_n)_{\text{obs}} = x_n$
Promedio	$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$	$(\overline{X}_n)_{\text{obs}} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Desvío (n)	$\Sigma_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2}$	$(\Sigma_n)_{\text{obs}} = \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$
Desvío $(n-1)$	$S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2}$	$(S_n)_{\text{obs}} = s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$
Tipo	Variable aleatoria	Número real

Podríamos usar otras fórmulas para estimar los parámetros de una distribución. En el ejemplo de las tapitas, podemos usar la mediana y el rango intercuartílico para estimar el centro y la dispersión de la distribución. Cuando se trata de un parámetro general θ, otras fórmulas serán las apropiadas.

Estadístico

Un estadístico T es una variable aleatoria que se calcula en función de las variables $X_1, \ldots, X_n$ que definen el muestreo aleatorio. En general, lo escribimos

$$T = T(X_1, \ldots, X_n)$$

para enfatizar que T es función de las $X_i's$.

La distribución de T se llama distribución muestral (o de muestreo) de T.

Estimador

Un estimador es un estadístico que creemos contiene información relevante sobre algún parámetro θ de la distribución de X.

El valor observado de un estadístico lo escribimos

$$T_{\text{obs}} = T(x_1, \ldots, x_n)$$

y corresponde a la muestra observada.

Estimador vs Estimación

Es importante distinguir entre el valor observado de un estimador y el estimador en sí. Para dejar en claro la diferencia, el valor observado de un estimador lo llamaremos *estimación*. Por ejemplo, en el caso de las tapitas, una estimación de μ es $\bar{x} = 2.98$, pero el estimador es $\overline{X}_{10}$. Este segundo es una variable aleatoria, mientras que el primero es un número.

3. Métodos de estimación

Existen muchísimos métodos diferentes que sistematizan la búsqueda de estimadores. Nosotros veremos solo algunos de ellos. Comenzaremos por el método de los momentos y el método de mínimos cuadrados. Luego terminaremos con uno de los más importantes, el método de máxima verosimilitud.

El método de los momentos

Dicho mal y pronto, este método consiste en despejar el parámetro de la fórmula de la esperanza, cambiando ésta por la media muestral.

Por ejemplo, supongamos que $X_1, \ldots, X_n$ es un muestreo aleatorio de X que tiene densidad

$$p(x;\theta) = (\theta+1)x^{\theta} \quad x \in [0,1].$$

La esperanza de X es entonces

$$\mathbf{E}(X) = (\theta+1)\int_0^1 x^{\theta+1}dx = \frac{\theta+1}{\theta+2}.$$

Esto quiere decir que la esperanza de X es una función del parámetro $g(\theta)$; en este caso $g(\theta) = \frac{\theta+1}{\theta+2}$.

En general uno espera que la media muestral esté cerca de la esperanza. Si en la igualdad $\mathbf{E}(X) = g(\theta)$ cambiamos el lado izquierdo por la media muestral $\overline{X}_n$, la igualdad será cierta si cambiamos el lado derecho por $g(\hat{\theta})$, para un $\hat{\theta}$ que esperamos esté cerca de θ.

En este caso la ecuación se transforma en

$$\overline{X}_n = \frac{\hat{\theta}+1}{\hat{\theta}+2} \Rightarrow \hat{\theta} = \frac{1-2\overline{X}_n}{\overline{X}_n - 1}.$$

Observar que despejar $\hat{\theta}$ de la igualdad anterior equivale a tomar $\hat{\theta} = g^{-1}(\overline{X}_n)$ en donde g^{-1} indica la inversa de g.

En general el método funciona de forma similar, salvo que cuando hay más de un parámetro es necesario calcular otros momentos además de la esperanza. Los momentos de X son por definición

$$M_1 = \mathbf{E}(X), \ldots, M_k = \mathbf{E}\left(X^k\right), \ldots$$

y supondremos que existen y son finitos.

Sea $X_1, \ldots, X_n$ un muestreo aleatorio de X, con distribución

$$p(x; \theta_1, \ldots, \theta_k) \text{ con } (\theta_1, \ldots, \theta_k) \in \Theta \subset \mathbb{R}^k.$$

Aquí p denota la densidad de X en el caso continuo o la f.p.p. en el caso discreto.

Entonces el vector de momentos es una función de los θ_i:

$$(M_1, \ldots, M_k) = \Phi(\theta_1, \ldots, \theta_k).$$

Denotamos por

$$\overline{M}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j$$

los *momentos muestrales*.

El método de los momentos consiste en resolver el sistema de ecuaciones

$$(\overline{M}_1, \ldots, \overline{M}_k) = \Phi\left(\hat{\theta}_1, \ldots, \hat{\theta}_k\right).$$

Es decir, los estimadores $\hat{\theta}_i,\ i = 1, \ldots, k$ están dados por

$$\left(\hat{\theta}_1, \ldots, \hat{\theta}_k\right) = \Phi^{-1}\left(\overline{M}_1, \ldots, \overline{M}_k\right).$$

Este sistema no tiene porque tener solución, ni ser única, pero sí en todas las aplicaciones que veremos en el curso.

Supongamos por ejemplo que la distribución p es $N(\mu, \sigma^2)$ y ambos parámetros son desconocidos. En este caso $p(x; \mu, \sigma^2)$ depende de dos parámetros, y podemos tomar $\theta_1 = \mu$ y $\theta_2 = \sigma^2$. Para aplicar el método de los momentos debemos calcular los primeros dos momentos de X. Estos son

$$M_1 = \mu,\ M_2 = \sigma^2 + \mu^2.$$

O sea que $\Phi : \mathbb{R} \times (0, +\infty) \to \{(x, y) : y > x^2\}$ es la función dada por

$$\Phi(\mu, \sigma^2) = (\mu, \mu^2 + \sigma^2).$$

La inversa es $\Phi^{-1}(x, y) = (x, y - x^2)$. Reemplazando x por $\overline{M}_1 = \overline{X}_n$ e y por $\overline{M}_2$ obtenemos los estimadores

$$\hat{\mu} = \overline{X}_n, \text{ y } \hat{\sigma}^2 = \overline{M}_2 - \overline{M}_1^2.$$

Notar que el estimador de σ^2 es la varianza muestral.

El método de mínimos cuadrados

El método de mínimos cuadrados generaliza a otros modelos de inferencia el método de los momentos. En nuestro caso, solo veremos uno de éstos modelos al final del curso, pero el método sirve para dar una base teórica más sólida al método de los momentos.

Supongamos por ejemplo que $X_1, \ldots, X_n$ es un muestro aleatorio de X, con distribución $p(x;\theta)$ que depende de un parámetro θ. El estimador de mínimos cuadrados $\hat{\theta}$ de θ corresponde al valor del parámetro que minimiza la suma

$$S(\theta) = \sum_{i=1}^{n} (X_i - \mathbf{E}(X_i))^2.$$

En nuestro caso, como todas las X_i tienen la misma distribución y $\mathbf{E}(X_i) = g(\theta)$ es una función del parámetro, vemos que

$$S(\theta) = \sum_{i=1}^{n} (X_i - g(\theta))^2.$$

En este caso, vemos que el mínimo debe satisfacer la ecuación

$$0 = S'(\hat{\theta}) = 2\sum_{i=1}^{n} (X_i - g(\hat{\theta}))g'(\hat{\theta}).$$

Si $g'(\theta)$ no se anula nunca, esta ecuación es equivalente a

$$\sum_{i=1}^{n} X_i - ng(\hat{\theta}) = 0,$$

o lo que es lo mismo $\hat{\theta} = g^{-1}(\overline{X}_n)$. Es decir, $\hat{\theta}$ coincide con el estimador de momentos.

El método de máxima verosimilitud

La estimación por máxima verosimilitud es un método para encontrar estimadores con buenas propiedades asintóticas. Motivaremos el método a través de un ejemplo que se conoce como captura/recaptura.

Imaginemos un biólogo que quiere estimar la cantidad de peces de una determinada especie en un gran lago. El problema es equivalente al siguiente: disponemos de una caja con una gran cantidad de bolitas y queremos estimar cuántas hay. Supongamos que las bolitas son todas rojas. ¿Cómo podemos hacer?

Una idea brillante es la de introducir una cantidad conocida de bolitas blancas en la caja, digamos m. Luego mezclamos bien y extraemos una cantidad n con reposición[3]. Llamemos N a la cantidad desconocida de bolitas en la caja. Digamos que entre las que hemos sacado de la caja, k son blancas. ¿Cómo podemos estimar N a partir de esta información?

Empecemos por una pregunta diferente pero relevante: ¿cuál es la probabilidad de sacar k bolas blancas de la caja? Si llamamos X a la cantidad de bolas blancas entre las extraídas, sabemos que X tiene distribución binomial de parámetros n y $p = m/N$. Entonces

$$\mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

[3] Si la cantidad de bolitas es grande, no hay diferencia prácitca entre sacar con o sin reposición.

Por supuesto, en este cálculo no conocemos p pues no conocemos N. ¿Cuál es el valor de p que hace que esta probabilidad sea los más grande posible? Dicho de forma ligeramente diferente, ¿cuál es el valor de p que maximiza la probabilidad de lo que hemos observado?

Consideremos la función

$$L(p) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Tomemos el logaritmo para que sea más fácil de derivar

$$\ell(p) = \ln L(p) = \ln \binom{n}{k} + k \ln(p) + (n-k)\ln(1-p).$$

Si derivamos respecto de p obtenemos

$$\ell'(p) = \frac{k}{p} - \frac{n-k}{1-p}$$

en donde se ve fácilmente que se anula para $\hat{p} = k/n$. Recordando que $p = m/N$, obtenemos la estimación $\hat{N} = m/\hat{p} = mn/k$. En esto consiste el método de máxima verosimilitud.

La función de verosimilitud

Consideremos una muestra $x_1, \ldots, x_n$ obtenida por un muestreo aleatorio $X_1, \ldots, X_n$ de una variable X cuya distribución $p(x;\theta)$ depende de un parámetro θ. La probabilidad de observar la muestra que observamos es

$$p(x_1;\theta)p(x_2;\theta)\cdots p(x_n;\theta)$$

interpetando este producto como una densidad en el caso continuo.

El *principio de máxima verosimilitud* consiste en estimar θ con el valor $\hat{\theta}$ que máximiza esta probabilidad. Se obtiene así una estimación $\hat{\theta}$ de θ basada en los datos de la muestra obtenida $x_1, \ldots, x_n$.

Para definir un estimador (una variable aleatoria) de θ basta reemplazar la muestra obtenida por el muestreo $X_1, \ldots, X_n$. Es así que definimos la función de verosimilitud.

Verosimilitud

Sea $X_1, \ldots, X_n$ i.i.d. $\sim p(x;\theta)$. La función de verosimilitud se define como

$$L_n(\theta) = \prod_{i=1}^{n} p(X_i;\theta).$$

Hemos indicado con el sub-índice n que la función depende del muestreo, y es por lo tanto aleatoria. La variable de la función es el parámetro θ.

Muchas veces es más sencillo hacer las cuentas con el logaritmo de la función de verosimilitud

$$\ell_n(\theta) = \sum_{i=1}^{n} \ln p(X_i, \theta).$$

El estimador de máxima verosimilitud T_n de θ se define como el valor de θ que maximiza la función de verosimilitud, o de forma equivalente su logaritmo.

Estimador de máxima verosimilitud

Sea $X_1, \ldots, X_n$ i.i.d. $\sim p(x;\theta)$. El estimador de máxima verosimilitud T_n de θ es el valor que maximiza la función de máxima verosimilitud, o de forma equivalente su logaritmo:

$$T_n = \text{argmax}_\theta \, L_n(\theta) = \text{argmax}_\theta \, \ell_n(\theta).$$

Notar que T_n depende de $X_1, \ldots, X_n$ pues $L_n(\theta)$ depende de ellas.

Ejemplos

En general, cuando la función de verosimilitud es diferenciable en el parámetro θ, para calcular T_n derivamos $\ell_n(\theta)$ respecto de θ e igualamos a cero. En general se trata siempre de un máximo, aunque para verificarlo deberíamos calcular el signo de la derivada segunda $\ell_n''(\theta)$ y ver que es negativo.

Cuando $L_n(\theta)$ no es diferenciable, cosa que sucede típicamente cuando el dominio de $L_n(\theta)$ depende de θ, hallar en dónde se da el máximo es más dificil y no se puede hacer derivando.

Veamos algunos ejemplos concretos.

Caso Bernoulli

Supongamos que $X_1, \ldots, X_n \sim \text{Ber}(p)$. El logaritmo de la función de verosimilitud es

$$\ell_n(p) = \sum_{i=1}^{n} X_i \ln(p) + (1 - X_i) \ln(1-p),$$

de donde vemos que

$$\begin{cases} \ell_n'(p) = n \left[\frac{\overline{X}_n}{p} - \frac{1-\overline{X}_n}{1-p} \right] \\ \ell_n''(p) = -n \left[\frac{\overline{X}_n}{p^2} + \frac{1-\overline{X}_n}{(1-p)^2} \right] \end{cases}$$

Notar que $\ell_n''(p) < 0$ para todo $p \in (0,1)$, por lo que el punto crítico es un máximo. Se puede deducir entonces que $T_n = \overline{X}_n$. Es decir, el estimador de máxima verosimilitud es el promedio.

Caso uniforme

Supongamos que $X_1, \ldots, X_n \sim U(0,\theta)$. La función de verosimilitud no es continua, y está dada por

$$L_n(\theta) = \begin{cases} \frac{1}{\theta^n} & \text{si } X_i < \theta \, \forall i \\ 0 & \text{si no.} \end{cases}$$

Decir que $X_i < \theta$ para todo i es equivalente a decir que $\text{máx}_i X_i < \theta$. Entonces podemos reescribir la función de verosimilitud como

$$L_n(\theta) = \begin{cases} \frac{1}{\theta^n} & \text{si } \text{máx}_i X_i < \theta \\ 0 & \text{si no.} \end{cases}$$

Ver la Figura 4. Es claro que el máximo se da en $T_n = \text{máx}_i X_i$, que es por lo tanto el estimador de máxima verosimilitud de θ.

Caso normal

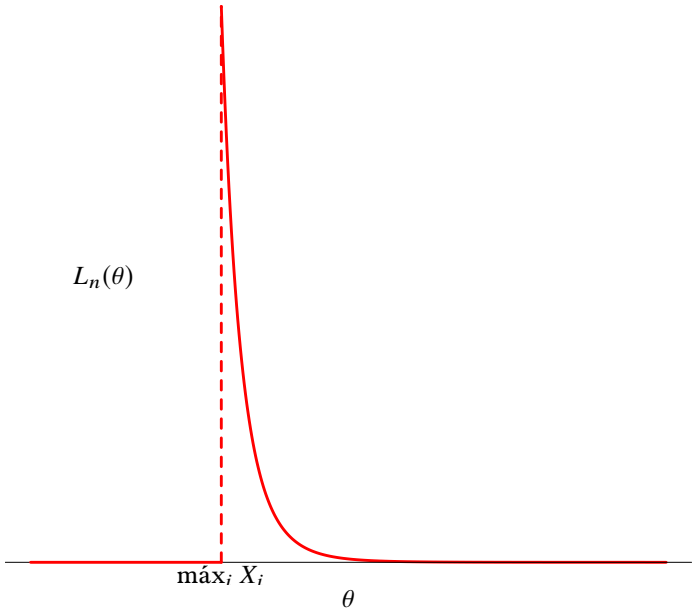

Figura 4: Función de verosimilitud $L_n(\theta)$ para variables uniformes.

El método de máxima verosimilitud puede emplearse en situaciones donde existen varios parámetros desconocidos, por ejemplo $\theta_1, \theta_2, \ldots, \theta_k$, que es necesario estimar. En tales casos la función de verosimilitud es una función de los k parámetros desconocidos $\theta_1, \ldots, \theta_k$, y los estimadores de máxima verosimilitud $T_1, \ldots, T_k$ se obtienen al igualar a cero las k derivadas parciales $\partial_i L(\theta_1, \ldots, \theta_k)$, y resolver el sistema de ecuaciones resultante.

Un ejemplo de está situación es la distribución normal de parámetros μ y σ^2, en donde $k = 2$.

Supongamos entonces que $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, con μ y σ^2 desconocidos. Entonces

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(X_i-\mu)^2/2\sigma^2} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma^2)\sum_{i=1}^{n}(X_i-\mu)^2}$$

Luego

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Ahora bien,

$$\begin{cases} \frac{\partial \ell}{\partial \mu}(\mu, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = 0 \\ \frac{\partial \ell}{\partial (\sigma^2)}(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2 \end{cases}$$

Las soluciones de este sistema son

$$\begin{cases} T_1 = \overline{X}_n \\ T_2 = S_n^2 \end{cases}$$

que son por ende los estimadores de máxima verosimilitud de μ y σ^2. Dejamos como ejercicio la verificación de que la Hessiana de ℓ es definida negativa. Esto asegura que se trata de un máximo.

Principio de invarianza

Una propiedad útil de los estimadores de máxima verosimilitud es la propiedad de invarianza. Veamos primero un ejemplo. Supongamos que queremos estimar la varianza de una distribución Bernoulli

$$p(x;p) = p^x(1-p)^{1-x},\ x \in \{0,1\}$$

Recordar que σ^2 es una función de p, dada en este caso por $\sigma^2 = g(p) = p(1-p)$. El problema es que no podemos usar σ^2 como parámetro, pues para cada valor de σ^2 existen dos posibles valores de p. Sin embargo, como el estimador de máxima verosimilitud de p es $\hat{p} = \overline{X}_n$, nos gustaría usar como estimador de máxima verosimilitud de σ^2 el estimador $\hat{\sigma}^2 = \overline{X}_n(1-\overline{X}_n)$. Esta es la idea detrás del principio de invarianza.

Supongamos que una distribución está indexada por un parámetro θ, pero el interés está en encontrar un estimador para alguna función de θ, digamos $g(\theta)$. Hablando informalmente, la propiedad de invarianza de estimador de máxima verosimilitud dice que si $\hat{\theta}$ es el estimador de máxima verosimilitud de θ, entonces $g(\hat{\theta})$ es el estimador de máxima verosimilitud de $g(\theta)$.

Por supuesto, hay algunos problemas técnicos que deben arreglarse antes de que podamos formalizar esta noción de invarianza, y se centran principalmente en la función $g(\theta)$ que estamos tratando de estimar. Si la función $g(\theta)$ es biyectiva (es decir, para cada valor de θ hay un único valor de $g(\theta)$ y viceversa), entonces no hay problema. En este caso, es fácil ver que no importa si maximizar la verosimilitud como una función de θ o como una función de $g(\theta)$. En ambos casos obtenemos la misma respuesta. Si ponemos $\eta = g(\theta)$, entonces la función inversa $\theta = g^{-1}(\eta)$ está bien definida y la función de verosimilitud de $g(\theta)$, escrita como una función de η, viene dada por

$$L^*(\eta) = \prod_{i=1}^{n} f\left(X_i; g^{-1}(\eta)\right) = L\left(g^{-1}(\eta)\right)$$

y por lo tanto

$$\max_{\eta} L^*(\eta) = \max_{\eta} L\left(g^{-1}(\eta)\right) = \max_{\theta} L(\theta).$$

Luego, el máximo de $L^*(\eta)$ se alcanza en $\eta = g(\hat{\theta})$, mostrando que el estimador de máxima verosimilitud de $g(\theta)$ es efectivamente $g(\hat{\theta})$.

Si $g(\theta)$ no es biyectiva, entonces para un valor dado de η puede haber más de un valor de θ que satisfaga $g(\theta) = \eta$. Debemos definir para $g(\theta)$ la función de verosimilitud inducida L^*, dada por

$$L^*(\eta) = \max_{\theta: g(\theta)=\eta} L(\theta).$$

El valor $\hat{\eta}$ que maximiza $L^*(\eta)$ lo llamaremos estimador de máxima verosimilitud de $\eta = g(\theta)$.

Para entender la definición, volvamos sobre el ejemplo Bernoulli. Lo anterior corresponde a definir la verosimilitud de σ^2 como la más grande de las verosimilitudes de los dos valores de p que dan el mismo σ^2.

Principio de invarianza

Si $\hat{\theta}$ es el estimador de máxima verosimilitud de θ, entonces $g(\hat{\theta})$ es el estimador de máxima verosimilitud de $g(\theta)$.

Para demostrar esta afirmación, llamemos $\hat{\eta}$ al valor de η que maximiza $L^*(\eta)$. Debemos probar que $L^*(\hat{\eta}) = L^*\left(g\left(\hat{\theta}\right)\right)$, pues esto quiere decir que el máximo también se alcanza en $g(\hat{\theta})$.

Primero, como los conjuntos $\{\theta : g(\theta) = \eta\}$ forman una partición de los valores posibles de θ, vemos que el máximo de L^* es igual al máximo de L:

$$\max_{\eta} L^*(\eta) = \max_{\eta} \max_{\theta : g(\theta)=\eta} L(\theta) = \max_{\theta} L(\theta).$$

Luego tenemos $L^*(\hat{\eta}) = L(\hat{\theta})$. Más aún, como $\hat{\theta}$ es en donde se alcanza el máximo de L, tenemos

$$L(\hat{\theta}) = \max_{\theta : g(\theta)=g(\hat{\theta})} L(\theta) = L^*(g(\hat{\theta})).$$

Juntando ambas igualdades llegamos a $L^*(\hat{\eta}) = L^*(g(\hat{\theta}))$ como queríamos demostrar.