Revista Colombiana de Estadística
Volumen 27 N° 2. Págs. 99 a 107. Diciembre 2004

Método bayesiano bootstrap y una aplicación en la estimación del percentil 85 en ingeniería de tránsito

Juan Carlos Correa M.*

Resumen

El percentil 85 juega un papel fundamental en ingeniería de tránsito. En este artículo presentamos diferentes procedimientos estadísticos, tanto paramétricos como no paramétricos, para su estimación. Mediante un ejemplo, ilustramos la diferencia entre ellos.

***Palabras Clave**: Percentil 85, estimación, ingeniería de tránsito.*

Abstract

The 85th quantile plays an important role in transportation engineering. In this paper we present different statistical procedures for its estimation, considering both, parametric and nonparametric procedures. With an example, we illustrate the difference between them.

***Key words**: 85th Quantile, Estimation, Transportation Engineering.*

*Profesor asociado. Escuela de Estadística Universidad Nacional de Colombia, Sede Medellín. E-mail: jccorrea@perseus.unalmed.edu.co

1. Introducción

La teoría clásica considera la información previa disponible básicamente para determinar los tamaños muestrales y los diseños de experimentos y, a veces, como forma de crítica de los resultados obtenidos. Una característica distintiva de la estadística bayesiana es la forma explícita como tiene en cuenta la información previa; sin embargo, uno de sus problemas se encuentra en la necesidad de asumir la forma paramétrica de la distribución que genera los datos. En este artículo vemos cómo, mediante la técnica bootstrap es posible evitar este supuesto.

Supongamos que estamos interesados en un parámetro particular de la población, digamos θ y que la información a priori sobre θ está resumida en $\xi(\theta)$. Si $x_1, x_2, \cdots, x_n$ representan la muestra obtenida de la población con densidad f desconocida, podemos aproximarla utilizando un estimador de densidades, digamos $\widehat{f}(x \mid \theta)$, y hallar un estimador de la distribución a posteriori como:

$$\xi(\theta \mid x_1, x_2, \cdots, x_n) \propto \widehat{L}(\theta \mid x_1, x_2, \cdots, x_n)\, \xi(\theta),$$

donde $\widehat{L}(\theta \mid x_1, \cdots, x_n)$ representa la función de verosimilitud estimada bootstrap, proporcional a $\widehat{f}$.

Boos & Monahan (1986) proponen la siguiente técnica bootstrap para determinar $\widehat{L}$:

1. Calcular la función de distribución empírica $\hat{F}_n$ de las x_i's.
2. Generar B muestras aleatorias de tamaño n de $\hat{F}_n$ y calcular $\hat{\theta}_j^*$ para la muestra j.
3. De las B estimadores simulados $\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_B^*$, calcular el estimador de densidades kernel,

$$\widehat{f}_{NB}(u) = \frac{1}{Bh_B} \sum_{i=1}^{B} K\left(\frac{u - (\widehat{\theta}_j^* - \widehat{\theta})}{h_B}\right),$$

como una estimación de la densidad de $\widehat{\theta} - \theta$. Si se hace $u = x - \theta$ en la ecuación anterior, $\widehat{f}_{NB}(x - \theta)$ es una estimación de la densidad muestral de $\widehat{\theta}$ dado θ. Evaluándola en $x = \widehat{\theta}$, resulta como función de θ para ser usada como verosimilitud:

$$\widehat{L}_{NB}\big(\widehat{\theta} \mid \theta\big) = \frac{1}{Bh_B} \sum_{i=1}^{B} K\left(\frac{2\widehat{\theta} - \theta - \widehat{\theta}_j^*}{h_B}\right).$$

4. La distribución posterior resultante $\xi(\widehat{\theta} \mid \theta)$ es entonces proporcional a $\xi(\theta)$
$\widehat{L}_{NB}(\widehat{\theta} \mid \theta)$, y la constante de normalización se puede hallar mediante integración numérica.

El percentil 85 es un parámetro importante en ingeniería de tránsito. En el presente artículo revisamos diferentes métodos de estimación, puntual y por intervalo de confianza, para dicho parámetro. Los métodos presentados se aplican también al percentil 15, otro parámetro importante para los ingenieros de tránsito, el cual puede considerarse como el dual del percentil 85. Al final presentamos un ejemplo con datos reales donde se aplican los diferentes métodos.

2. El procedimiento bootstrap

La técnica conocida como bootstrap fue propuesta por Efron (1979, 1982) para hallar intervalos de confianza en situaciones donde es imposible hallar analíticamente la distribución muestral del estimador. Es una técnica de remuestreo, de uso intensivo del computador, y funciona de la siguiente forma:

1. Sea $X_1, X_2, \cdots, X_n$ la muestra a nuestra disposición, y $\widehat{F}$ la función de distribución empírica.

2. Se utiliza un generador de números aleatorios para obtener n nuevos puntos $X_1^*, X_2^*, \cdots, X_n^*$ independientemente y con reemplazo de $\widehat{F}$. Estos nuevos valores son llamados una *muestra bootstrap*.

3. Se calcula el estadístico de interés para la muestra bootstrap.

4. Se repiten los pasos 1) y 2) un número muy grande de veces, digamos N, cada vez con una muestra independiente. Digamos que la secuencia de estimadores bootstrap para el estadístico de interés es $\widehat{\theta}^{*^1}, \widehat{\theta}^{*^2}, \widehat{\theta}^{*^3}, \cdots, \widehat{\theta}^{*^N}$.

5. Con estas muestras bootstrap se puede realizar todo el trabajo inferencial deseado.

Refinamientos del procedimiento anterior se encuentran en DiCiccio & Tibshirani (1987).

3. Estimación clásica del percentil 85

3.1. Métodos paramétricos

Los métodos paramétricos requieren la especificación de la distribución de la cual provienen los datos, por ejemplo, si la distribución de los datos es normal, Weibull, etc. Una vez estimados los parámetros que caracterizan la distribución, por alguno de los métodos tradicionales, —el de máxima verosimilitud es uno de ellos— se procede a estimar el percentil poblacional, digamos ζ_{85}, calculado como:

$$\int_{-\infty}^{\zeta_{85}} f(x \mid \theta)dx = 0{,}85,$$

donde $f(x \mid \theta)$ es la densidad de la población de la cual provienen los datos con función de distribución $F(x \mid \theta)$. Si $\widehat{\theta}$ es un estimador para θ, basado en la muestra $X_1, X_2, \cdots, X_n$, entonces el estimador de ζ_{85} será $\widehat{\zeta}_{85}$ y se puede calcular de la ecuación:

$$F(\widehat{\zeta}_{85}) = \int_{-\infty}^{\widehat{\zeta}_{85}} f(x \mid \widehat{\theta})dx = 0{,}85.$$

En el caso de la distribución Weibull tendremos:

$$F(\widehat{\zeta}_{85}) = 1 - \exp\Big(-\Big(\frac{\widehat{\zeta}_{85}}{\widehat{\beta}}\Big)^{\widehat{\alpha}}\Big) = 0{,}85,$$

donde $\widehat{\beta} > 0$ y $\widehat{\alpha} > 0$ son los parámetros estimados de la distribución. De la anterior expresión obtenemos:

$$\widehat{\zeta}_{85} = \widehat{\beta}\,(-\ln(0{,}15))^{\frac{1}{\widehat{\alpha}}}.$$

Un estimador sencillo que corresponde a un elemento en la muestra es:

$$\widehat{\zeta}_{85} = X_{([0{,}85n]+1)},$$

donde $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ son los llamados *estadísticos de orden* de la muestra, esto es, la muestra ordenada en forma creciente, y $[0{,}85n]$ es el menor entero más cercano a $0{,}85n$.

La densidad de $\widehat{\zeta}_{85}$, asumiendo el estimador sencillo, está dada por:

$$g_{[0{,}85n]+1}(t) = \frac{n!}{[0{,}85n]!\big(n-[0{,}85n]-1\big)!}\big\{F(t)\big\}^{[0{,}85n]}\big\{1-F(t)\big\}^{n-[0{,}85n]-1}f(t).$$

Para la distribución Weibull tratada anteriormente, la función densidad de probabilidad será:

$$\begin{aligned} g_{[0,85n]+1}(t) &= \frac{n!}{[0{,}85n]!(n-[0{,}85n]-1)!}\left\{1-\exp\Big(-\Big(\frac{t}{\beta}\Big)^{\alpha}\Big)\right\}^{[0,85n]} \\ &\quad\times\left\{\exp\Big(-\Big(\frac{t}{\beta}\Big)^{\alpha}\Big)\right\}^{n-[0,85n]-1}\left(\frac{\alpha t^{\alpha-1}}{\beta^{\alpha}}\right)\exp\Big(-\Big(\frac{t}{\beta}\Big)^{\alpha}\Big). \end{aligned}$$

Los estimadores de máxima verosimilitud para α y β son la solución del siguiente sistema de ecuaciones simultáneas (Johnson & Kotz 1970, Pág. 255):

$$\begin{aligned} \widehat{\beta} &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\widehat{\alpha}}\right)^{\frac{1}{\widehat{\alpha}}}, \\ \widehat{\alpha} &= \frac{n}{\left(\frac{1}{\widehat{\beta}}\right)^{\widehat{\alpha}}\sum_{i=1}^{n} x_i^{\widehat{\alpha}}\ln(x_i)-\sum_{i=1}^{n}\ln(x_i)}. \end{aligned}$$

Cuando $n\to\infty$ podemos utilizar el siguiente resultado asintótico: si F posee una densidad f en una vecindad de ζ_p, donde f es positiva y constante, entonces:

$$\widehat{\zeta}_p \text{ es } AN\Big(\zeta_p, \frac{p(1-p)}{f^2(\zeta_p)\,n}\Big).$$

Por lo tanto, un intervalo de confianza asintótico de nivel $100(1-\alpha)\,\%$ para $\zeta_{0,85}$, está dado por:

$$\Big(\widehat{\zeta}_{0,85} - z_{\frac{\alpha}{2}}\sqrt{\frac{0{,}85\times 0{,}15}{n}}\frac{1}{f(\widehat{\zeta}_{0,85})},\ \widehat{\zeta}_{0,85} + z_{\frac{\alpha}{2}}\sqrt{\frac{0{,}85\times 0{,}15}{n}}\frac{1}{f(\widehat{\zeta}_{0,85})}\Big).$$

En la práctica f es desconocida; por lo tanto, se puede utilizar un estimador kernel de densidades, de la forma:

$$\widehat{f}(x) = \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right).$$

3.2. Métodos no paramétricos

El cuantil muestral de orden p $(0<p<1)$ es:

$$\widehat{\zeta}_p = X_{([np]+1)},$$

donde $[np]$ denota el mayor entero menor o igual que np.

El intervalo de confianza no paramétrico para ζ_p, está dado por $\left(X_{(i)}, X_{(j)}\right)$, con nivel de confianza $Q(i,j \mid p,n)$, con $1 \leq i < j \leq n$ y $0 < p < 1$,

$$Q(i,j \mid p,n) = \sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{n-k}.$$

3.3. Bootstrap

La técnica bootstrap, ya descrita antes, funciona de la siguiente manera:

1. Sea $X_1, X_2, \cdots, X_n$ la muestra a nuestra disposición. Sea $\widehat{F}$ la función de distribución empírica.

2. Utilice un generador de números aleatorios para obtener n nuevos puntos $X_1^*, X_2^*, \cdots, X_n^*$ independientemente y con reemplazo de $\widehat{F}$. Estos nuevos valores son llamados una *muestra bootstrap*.

3. Calcule el percentil 85 para la muestra bootstrap.

4. Repita los pasos 1) y 2) un número muy grande, digamos N, cada vez con una muestra independiente. Digamos que la secuencia de estimadores bootstrap para el percentil 85 es $\widehat{\zeta}_{0,85}^{*\,1}, \widehat{\zeta}_{0,85}^{*\,2}, \widehat{\zeta}_{0,85}^{*\,3}, \cdots, \widehat{\zeta}_{0,85}^{*\,N}$.

5. Denotemos por $[a^*, b^*]$ el intervalo central con 95 % de los valores $\widehat{\zeta}_{0,85}^*$, o sea,
$$\frac{\#\left\{\widehat{\zeta}_{0,85}^{*\,j} < a^*\right\}}{N} = 0{,}025, \quad \text{y} \quad \frac{\#\left\{\widehat{\zeta}_{0,85}^{*\,j} < b^*\right\}}{N} = 0{,}975.$$

Refinamientos del intervalo anterior se encuentran en DiCiccio & Tibshirani (1987).

Los métodos bayesianos consideran los parámetros como variables aleatorias, no fijos como en la escuela clásica; por lo tanto, el concepto de distribución de los parámetros es fundamental. También se considera posible el uso de información a priori, no obtenida por la observación de una muestra de la distribución de los datos. Esta parte ha sido controversial, y el carácter multivariado de los parámetros dificulta en grado sumo la aplicación de estas técnicas.

En general, la técnica se resume así: Sea $\xi(\theta)$, la distribución a priori, y $f(x_1, x_2, \cdots, x_n \mid \theta)$, la distribución de la muestra aleatoria observable. La

unión de la informaciones a priori y muestral genera una distribución conocida como la distribución a posteriori, denotada por $\xi\left(\theta \mid x_1, x_2, \cdots, x_n\right)$, calculada esta última como:

$$\xi\left(\theta \mid x_1, x_2, \cdots, x_n\right) \propto \xi(\theta) \times f(x_1, x_2, \cdots, x_n \mid \theta),$$

donde $\propto$ es el símbolo de proporcionalidad.

4. Ejemplo

Con el propósito de ilustrar los métodos presentados anteriormente utilizaremos una información sobre velocidades recogida por estudiantes del posgrado de vías de la Universidad Nacional de Colombia, Sede Medellín, en la carretera El Volador. Se tomó un tramo de 25,75 mts y con el uso de un cronómetro y un enoscopio se calcula la velocidad de un carro. Las velocidades registradas para automóviles fueron, en km/h,

60,2	43,3	51,2	46,6	32,5	41,8	45,9	60,6	32,3	31,7
39,4	41,2	60,2	49,0	40,5	58,3	42,7	61,4	26,0	53,3
58,7	46,4	39,1	63,9	51,5	53,3	41,6	54,9	55,2	60,2
47,3	39,8	46,8	64,4	57,9	39,1	44,8	65,3	69,7	50,4
54,2	39,4	46,6	55,8	53,6	61,8	44,3	48,5	53,9	61,4
38,1	47,8								

La media de los datos es 49,49615 y la desviación estándar es 9,87119.

Si asumimos la distribución de Weibull como la que origina los datos, tenemos como parámetros estimados por el método de máxima verosimilitud $\widehat{\alpha} = 5{,}791988$ y $\widehat{\beta} = 53{,}48502$. Con esta distribución obtenemos un percentil 85 estimado igual a 59,73958.

El estimador sencillo del percentil 85 es 60,6. El intervalo de confianza obtenido utilizando la f.d.p. $g_{[0,85n]+1}(t)$, utilizando $\widehat{\alpha}$ y $\widehat{\beta}$, es $(55{,}85; 62{,}80)$. Se calcula resolviendo la ecuación:

$$\int_A^B g_{[0,85n]+1}(t)dt = 0{,}95,$$

donde el intervalo de confianza es (A, B). El intervalo de confianza asintótico del 95 % para el percentil 85, asumiendo que la distribución que genera los datos es Weibull es $(56{,}7297; 64{,}4703)$.

El intervalo de confianza del 95 % bootstrap es (57,9; 63,9). El intervalo de confianza no paramétrico presentado en la sección 3 es (58,3; 64,4) que corresponde a las observaciones ordenadas 40 y 50. El nivel de significancia es 0,948567, que es el más cercano al nivel deseado 0,95.

Tabla 1: Resumen de los intervalos clásicos.

Método	Límite Inferior	Límite Superior
Exacto	55.85	62.80
Asintótico	56.7297	64.4703
No paramétrico	58.3	64.4
Bootstrap	57.9	63.9

4.1. Método bayesiano bootstrap

Si asumimos que podemos resumir nuestro conocimiento a priori sobre el percentil 85, con una distribución normal con media μ_0 y desviación típica σ_0, entonces los intervalos de probabilidad, para diferentes valores, están dados a continuación:

Tabla 2: Intervalos de probabilidad para diferentes a priori.

A priori	Media	Moda	Límite inferior	Límite superior
$N(70, 20^2)$	60,37954	60,60606	56,36364	63,63636
$N(70, 10^2)$	60,69627	60,60606	56,96970	64,24242
$N(70, 5^2)$	61,53235	61,21212	58,18182	64,84848
$N(60, 10^2)$	60,30264	60,60606	56,36364	63,63636
$N(60, 3^2)$	60,29840	60,60606	56,96970	63,03030
$N(60, 1^2)$	60,14754	60,00000	58,18182	61,81818

Hemos seleccionado distribuciones a priori que reflejan desde muy poco conocimiento, llamadas poco informativas, pero en términos de una distribución

normal, que se muestran en términos de una gran varianza, hasta distribuciones a priori con varianzas muy pequeñas, lo que indica buena información previa. Sin embargo, el intervalo de probabilidad a posteriori es relativamente estable, lo cual indica un gran dominio de la información muestral.

5. Conclusiones y recomendaciones

El ingeniero de tránsito puede seleccionar el método de estimación de los percentiles según las condiciones que se presenten en su caso particular. Si no tiene una idea clara y justificable de la distribución teórica, es preferible seleccionar uno de los métodos no paramétricos.

El método bayesiano permite la incorporación explícita de información previa disponible, lo cual es muy atractivo para el ingeniero de tránsito, ya que usualmente esta información es abundante. Cómo resumir esta información en forma de distribución de probabilidad, es un problema que no tiene una solución única y clara. Además, el método bayesiano permite realizar inferencias aún sin haber obtenido una muestra, lo cual no es suficientemente resaltado.

Bibliografía

Boos, D. D. & Monahan, J. F. (1986), 'Bootstrap methods using prior information', *Biometrika* **73**(1), 77–83.

DiCiccio, T. & Tibshirani, R. (1987), 'Bootstrap confidence intervals and bootstrap approximations', *Journal of American Statistical Association* **82**(397), 163–170.

Dudewicz, E. J. (1976), *Introduction to Statistics and Probability*, Holt, Rinehart and Winston.

Efron, B. (1979), 'Computers and the theory of statistics: Thinking the unthinkable', *SIAM Review* **21**(4), 460–480.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM.

Johnson, N. L. & Kotz, S. (1970), *Continuous Univariate Distributions-1*, John Wiley & Sons.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons.