SOBRE LA ESTIMACIÓN DEL PERCENTIL 85 EN INGENIERÍA DE TRÁNSITO

JUAN CARLOS CORREA M. y FRANCISCO CASTRILLÓN
Departamento de Matemáticas Universidad Nacional-Sede Medellín
A.A. 3840 e-mail: jccorrea@perseus.unalmed.edu.co

RESUMEN

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

PALABRAS CLAVES:
Percentil 85, Estimación, Ingeniería de Tránsito

SUMMARY

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

KEYWORDS:
Quantil 85, Estimation, Transportation Engineering.

1. INTRODUCCIÓN

El percentil 85 es un parámetro importante en ingeniería de tránsito. En el presente artículo revisamos diferentes métodos de estimación para dicho parámetro y diferentes intervalos de confianza. Los métodos requieren diversas condiciones. Otro parámetro importante para los ingenieros de tránsito es el percentil 15, el cual puede considerarse como el dual del percentil 85; los métodos presentados se aplican también a él. Al final presentamos un ejemplo con datos reales donde se aplican los diferentes métodos.

2. 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 los parámetros que caracterizan la distribución han sido estimados, por alguno de los métodos tradicionales, -el de máxima verosimilitud es uno de ellos-, entonces se procede a estimar el percentil poblacional, digamos ζ_{85}, calculado como

$$\int_{-\infty}^{\zeta_{85}} f(x|\theta)dx = 0.85$$

donde $f(x|\theta)$ es la densidad de la población de la cual provienen los datos con función de distribución $F(x|\theta)$. Si $\hat{\Theta}$ es un estimador para Θ, basado en la muestra $X_1, X_2, \ldots, X_n$, entonces el estimador de ζ_{85} será $\hat{\zeta}_{85}$ y se puede calcular de la ecuacion

$$F(\hat{\zeta}_{85}) = \int_{-\infty}^{\hat{\zeta}_{85}} f(x|\hat{\theta})dx = 0.85$$

En el caso de la distribucion Weibull tendremos

$$F(\hat{\zeta}_{85}) = 1 - \exp\left(-\left(\frac{\hat{\zeta}_{85}}{\hat{\beta}}\right)^{\hat{\alpha}}\right) = 0.85$$

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

$$\hat{\zeta}_{85} = \hat{\beta}\left(-\ln\left(0.15\right)\right)^{\frac{1}{\hat{\alpha}}}$$

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

$$\hat{\zeta}_{85} = X_{([n0.85]+1)}$$

donde $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ son los llamados *estadisticos de orden* de la muestra, esto es, la muestra ordenada en forma creciente, y $[n0.85]$ es el menor entero más cercano a $n0.85$

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

$$g_{[n0.85]+1}(t) = \frac{n!}{[n0.85]!(n-[n0.85]-1)!}\{F(t)\}^{[n0.85]}\{1-F(t)\}^{n-[n0.85]-1}f(t)$$

Para el caso de la Weibull tratada anteriormente la función densidad de probabilidad será

$$g_{[n0.85]+1}(t) = \frac{n!}{[n0.85]!(n-[n0.85]-1)!}\left\{1-\exp\left(-\left(\frac{t}{\beta}\right)^{\alpha}\right)\right\}^{[n0.85]} \times \left\{\exp\left(-\left(\frac{t}{\beta}\right)^{\alpha}\right)\right\}^{n-[n0.85]-1}\left(\frac{\alpha t^{\alpha-1}}{\beta^{\alpha}}\right)\exp\left(-\left(\frac{t}{\beta}\right)^{\alpha}\right)$$

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

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^{\hat{\alpha}}\right)^{\frac{1}{\hat{\alpha}}}$$

$$\hat{\alpha} = \frac{n}{\left(\left(\frac{1}{\hat{\beta}}\right)^{\hat{\alpha}}\sum_{i=1}^{n}x_i^{\hat{\alpha}}\ln(x_i) - \sum_{i=1}^{n}\ln(x_i)\right)}$$

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

$$\hat{\zeta}_p es AN\left(\zeta_p, \frac{p(1-p)}{f^2(\zeta_p)\,n}\right)$$

Por lo tanto un intervalo de confianza asintótico de nivel $100(1-\alpha)\%$ para $\zeta_{0.85}$ está dado por:

$$\left(\hat{\zeta}_{0.85} - z_{\frac{\alpha}{2}}\sqrt{\frac{0.85\times 0.15}{n}}\frac{1}{f\left(\hat{\zeta}_{0.85}\right)},\right.$$

$$\left.\hat{\zeta}_{0.85} + z_{\frac{\alpha}{2}}\sqrt{\frac{0.85\times 0.15}{n}}\frac{1}{f\left(\hat{\zeta}_{0.85}\right)}\right)$$

En el ejemplo que nos ocupará, asumiremos una población Weibull de datos con parámetros *alpha* y *beta* conocidos.

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

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

3. MÉTODOS NO PARAMÉTRICOS

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

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

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

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

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

3.1 BOOTSTRAP

La técnica conocida como Bootstrap fue propuesta por Efron (1979, 1982) para hallar intervalos de confianza en situaciones donde la distribución muestral del estimador es imposible de hallar analíticamente. Es una técnica de re-muestreo que hace uso intensivo del computador. El método funciona de la siguiente forma:

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

2) Utilice un generador de números aleatorios para obtener n nuevos puntos $X_1^*, X_2^*, \ldots, X_n^*$ independientemente y con reemplazo de $\hat{F}$. 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 $\hat{\zeta}^{*1}_{0.85}, \hat{\zeta}^{*2}_{0.85}, \hat{\zeta}^{*3}_{0.85}, \ldots, \hat{\zeta}^{*N}_{0.85}$.

5) Denotemos por $[a^*, b^*]$ el intervalo central con 95 de los valores $\hat{\zeta}^*_{0.85}$, o sea,

$$\frac{\#\left\{\hat{\zeta}^{*j}_{0.85} < a^*\right\}}{N} = 0.025, y \frac{\#\left\{\hat{\zeta}^{*j}_{0.85} < b^*\right\}}{N} = 0.975$$

Refinamientos del intervalo anterior se encuentran en DiCiccio y 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 apriori 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: Denotemos la distribución apriori como $\xi(\theta)$. La distribución de la muestra aleatoria observable como $f(x_1, x_2, \ldots, x_n|\theta)$. La unión de la información apriori y la muestral genera una distribución conocida como la distribución aposteriori, denotada por $\xi(\theta|x_1, x_2, \ldots, x_n)$. Esta última se calcula como

$$\xi(\theta|x_1, x_2, \cdots, x_n) \propto \xi(\theta) \times f(x_1, x_2, \cdots, x_n|\theta)$$

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 ví as de la Universidad Nacional-Sede Medellín en la carretera El Volador, Abril de 1996. 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. El gráfico 1 nos presenta el histograma de de los datos. El gráfico 2 es un gráfico de caja para los mismos. La caja contiene el 50% de las observaciones, mientras que el 25% de ellas están por debajo y el restante 25% por encima. La media de los datos es 49.49615 y la desviación estándar es 9.87119.

Si asumimos la distribución de Weibull como la población que origina los datos, tenemos como parámetros estimados por el método de máxima verosimilitud $\hat{\alpha}$ =5.791988 y $\hat{\beta}$ =53.48502. El gráfico

3 presenta la distribución de Weibull estimada y el gráfico 4 nos presenta un gráfico *Q-Q* (Cuantil vs. Cuantil) de los datos contra esta distribución asumida. Este último gráfico nos permite visualizar si el modelo asumido se ajusta o no se ajusta bien a los datos obtenidos. Como puede verse, pareciera éste indicar que no existe un buen ajuste de los datos al modelo planteado, lo que nos conduce a pensar que, o bien, los datos son insuficientes o el modelo no se ajusta, esto último implicaría la búsqueda de otro modelo. Sin embargo, 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. $g_{[n0.85]+1}(t)$ utilizando $\hat{\alpha}$ y $\hat{\beta}$, es (55.85, 62.80). Se calcula resolviendo la siguiente ecuación

$$\int_A^B g_{[n0.85]+1}(t)dt = 0.95$$

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 corresponden a las observaciones ordenadas 40 y 50. El nivel de significancia es 0.948567, que es el más cercano al nivel deseado 0.95.

4. PROGRAMA EN S-plus

A continuación se presenta el programa en S-plus con el cual se pueden calcular los resultados previos:

```
# Programa para hallar el estimador sencillo del percentil 85
# Argumentos de entrada: DATA: vector de datos
percentil.85<-function(DATA){
x<-sort(as.vector(DATA))
percentil<-x[trunc(0.85*length(x)+1)]
percentil
}
```

```
# Programa para hallar los estimadores de maxima verosimilitud de una #Weibull(alfa,beta)
# Argumentos de entrada: DATA: vector de datos
# valor.inicial: un valor inicial positivo para alfa

estimar.weibull<-function(DATA,valor.inicial){
x<-as.vector(DATA)
func<-function(a){abs(a-1/(x^a%*%log(x)/sum(x^a)
- mean(log(x)))) }
minimizacion<-nlmin(func,valor.inicial)
alfa<-minimizacion$x
beta<-mean(x^alfa)^(1/alfa)
cat('alfa= ',alfa,' beta= ',beta,'\n')
}

# Programa para calcular el intervalo de confianza asintotico
# para el percentil 85 cuando la poblaci\'on es Weibull
# Argumentos de entrada: estimador sencillo, calculado con percentil.85()
# alfa y beta: calcualdos con la funcion estimar.weibull()
# n:numero de observaciones
# nivel: el nivel de confianza deseado. 0<nivel<1

IC.percentil.85<-function(estimador,n,alfa,beta,nivel){
a<-qnorm(1-(1-nivel)/2)*sqrt(0.85*0.15/n)/
dweibull(estimador,alfa,beta)
LI<-estimador-a
LS<-estimador+a
cat('Limite inferior = ',LI, ' Limite superior= ', LS, 'Nivel de
confianza= ', nivel, '\n')
}

# FD de X[0.85n+1] de una Weibull
# Parametros de entrada
# alfa, beta: parametros de la distribucion
# n: tamano muestral

fd.weibull85<-function(t,alfa,beta,n){
F1<-pweibull(t,alfa,beta)
k<-trunc(n*0.85+1)
f<-1-pbinom(k-1,n,F1)
}
```

```
# Funcion para calcular el intervalo de confianza
# bootstrap para el percentil 85
# Parametros de entrada: DATA Vector de datos
# n: numero de muestras bootstrap
# nivel: nivel de confianza deseado

IC.bootstrap<-function(DATA,n,nivel){
x <-
apply(matrix(rep(DATA,n),nrow=length(DATA),byrow=F),
2,sample,replace=T)
y<-sort(apply(x,2,percentil.85))
alfa<-1-nivel
k1<-trunc(n*alfa/2)
k2<-trunc(n*(1-alfa/2))
LI<-y[k1]
LS<-y[k2]
cat(''Limite Inferior= '',LI,'' Limite Superior=
'',LS,''\n'')
}
```

5. CONCLUSIONES

El ingeniero de tránsito puede selecccionar el método de estimación de los percentiles de acuerdo a 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. En un artículo que estamos preparando realizamos un extenso estudio de simulación que permite comparar la eficiencia relativa de estos procedimientos.

6. REFERENCIAS

DiCiccio, T. y Tibshirani, R. (1987) Bootstrap Confidence Intervals and Bootstrap Approximations. *Journal of American Statistical Association*. Vol. 82, No. 397, pp.163-170

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

Efron, B. (1979) Computers and the Theory of Stattitics: Thinking the Unthinkable. *SIAM Review*. Vol. 21, No. 4, pp. 460-480

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

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

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

grafico 1

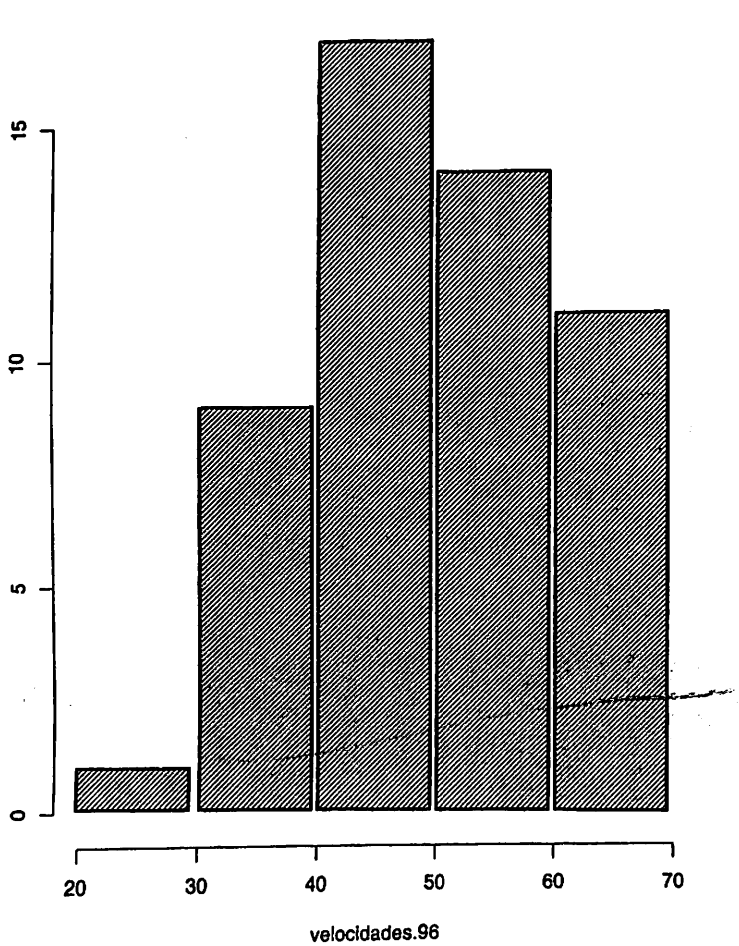

grafico 2

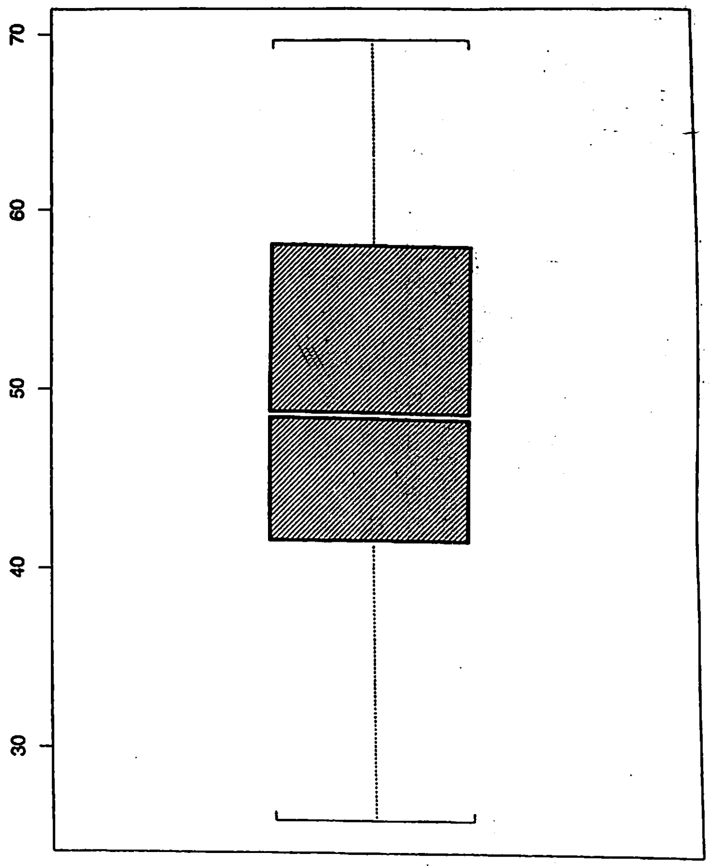

grafico 3

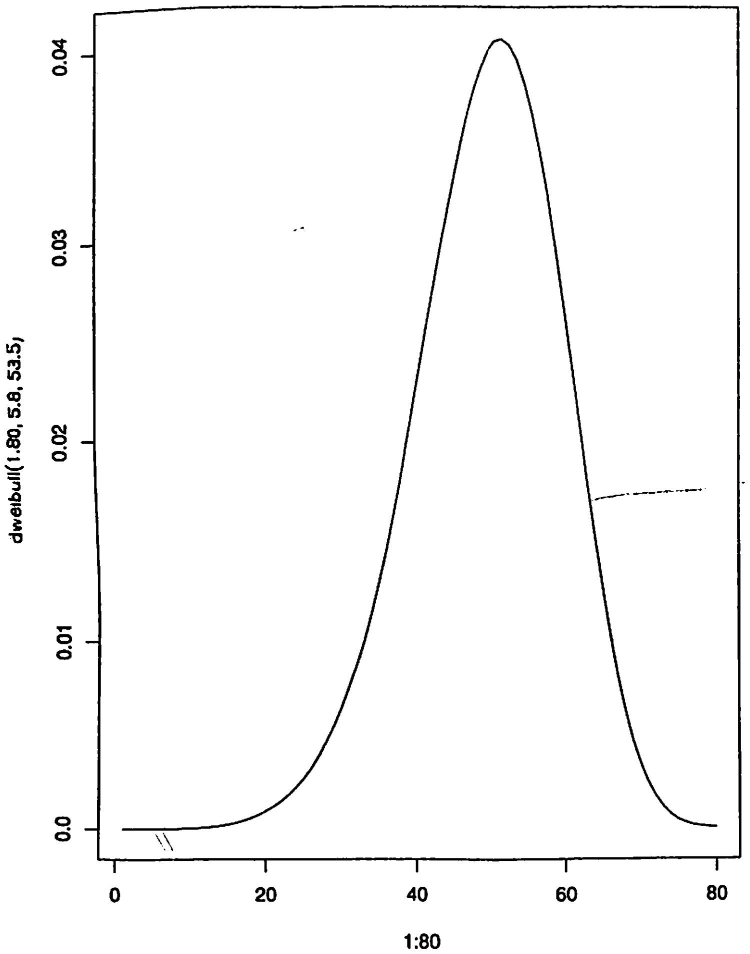

grafico 4