Revista Colombiana de Estadística
Volumen 29 N° 1. pp. 1 a 16. Junio 2006

Una estimación del parámetro de la distribución *g* de Tukey

An Estimation of the Parameter of the *g* Tukey Distribution

José Alfredo Jiménez*, Jorge Martínez†
Universidad Nacional de Colombia, Bogotá

Resumen

Se presenta una forma explícita de la función de densidad de una variable con distribución $g-h$ de Tukey, en términos de los cuantiles de la distribución normal estándar. La expresión de la densidad de probabilidad propuesta permite establecer un estimador del parámetro g asociado a la subfamilia de distribuciones g de Tukey.

Palabras clave: Cuantiles, distribución $g-h$ de Tukey.

Abstract

This paper presents an explicit form of the density function of a random variable with the $g-h$ Tukey distribution, in terms of the quantiles of the standard normal distribution. The estimator of the parameter g is obtained based on this particular form of the probability density.

Key words: Quantiles, $g-h$ Tukey distribution.

1. Introducción

La familia de distribuciones $g-h$ de Tukey comprende una considerable variedad de distribuciones continuas con características especiales en cuanto a asimetría y elongación, por lo cual resulta de gran utilidad cuando se desea construir un modelo distribucional para un conjunto de datos o analizar la sensibilidad de un proceso de simulación frente a diversas alternativas para la forma de la distribución de las variables. A partir de esta familia de distribuciones se obtienen dos subfamilias: la g y la h. Algunas de las propiedades de esta familia de distribuciones las presentan Martínez (1981) y Martínez & Iglewicz (1984).

*Profesor asociado, Departamento de Matemáticas. E-mail: josajimenezm@unal.edu.co
†Profesor asociado, Departamento de Estadística. E-mail: jmartinezc@unal.edu.co

En este artículo se estudia *la subfamilia de distribuciones* g por su gran importancia en el estudio de distribuciones no simétricas de especial interés en campos como el análisis de sobrevivencia o el análisis de algunas variables económicas típicamente asimétricas.

2. Distribución $g-h$ de Tukey

Si Z es una variable aleatoria con distribución normal estándar y se asume que g y h son constantes arbitrarias (parámetros), la variable aleatoria Y definida como:

$$Y = T_{g,h}(Z) = \frac{1}{g}(e^{gZ} - 1)e^{\frac{1}{2}hZ^2} \qquad \text{con } g \neq 0, \quad h \in \mathbb{R} \tag{1}$$

se dice que tiene distribución $g-h$ de Tukey.

Sea $p > 0.5$ el p-valor que permite calcular los cuantiles Z_p de la distribución normal, entonces de (1) se puede obtener:

$$y_p = T_{g,h}(Z_p) = \frac{1}{g}(e^{gZ_p} - 1)e^{\frac{1}{2}hZ_p^2} \tag{2}$$

Por la simetría de la distribución normal ($Z_{(1-p)} = -Z_p$) se obtiene:

$$\begin{aligned} y_{(1-p)} = T_{g,h}(Z_{1-p}) &= \frac{1}{g}(e^{-gZ_p} - 1)e^{\frac{1}{2}hZ_p^2} \\ &= T_{g,h}(-Z_p) = -e^{-gZ_p}\left[\frac{1}{g}(e^{gZ_p} - 1)e^{\frac{1}{2}hZ_p^2}\right] \\ &= -e^{-gZ_p}y_p \end{aligned} \tag{3}$$

Por lo tanto, conjugando estos dos resultados y el hecho de que la constante $g \neq 0$, se obtiene además:

$$T_{-g,h}(Z_p) = -T_{g,h}(Z_{1-p}) \tag{4}$$

Por otra parte, si se asume $g = 0$, la distribución $g-h$ de Tukey resulta ser una distribución simétrica respecto al origen, ya que:

$$\lim_{g\to 0} Y = T_{0,h}(Z) = \lim_{g\to 0}\left(\frac{e^{gZ} - 1}{g}\right)e^{\frac{1}{2}hZ^2} = Ze^{\frac{1}{2}hZ^2} \tag{5}$$

Esta distribución pertenece a la llamada *subfamilia de distribuciones* h, conformada por distribuciones simétricas con la propiedad de que sus colas se alargan con el crecimiento de h. Además, como

$$\begin{aligned} T_{0,h}(Z_{1-p}) &= Z_{1-p}\, e^{\frac{1}{2}hZ_{1-p}^2} \\ T_{0,h}(-Z_p) &= -Z_p\, e^{\frac{1}{2}hZ_p^2} = -T_{0,h}(Z_p) \end{aligned}$$

es decir que $T_{0,h}(Z)$ es una función impar, $T_{0,h}(Z)$ es simétrica con respecto al origen.

Cuando $h = 0$ en (1), se obtiene la variable aleatoria,

$$Y = T_{g,0}(Z) = \left(\frac{e^{gZ} - 1}{g}\right)$$

la cual define la *subfamilia de distribuciones g*.

2.1. Propiedades

En esta sección se examinará si la transformación $Y = T_{g,h}(Z)$ es una función creciente o decreciente de Z. Para ello se deriva la ecuación (1) respecto a Z y se obtiene:

$$\frac{dy}{dZ} = \frac{d}{dZ}T_{g,h} = \begin{cases} e^{gZ}\left[1 + hZ\dfrac{e^{-gZ} - 1}{-g}\right]e^{\frac{1}{2}hZ^2} & g \neq 0 \\ \left(1 + hZ^2\right)e^{\frac{1}{2}hZ^2} & g = 0 \end{cases} \tag{6}$$

1. Si $g \neq 0$, la función definida por (1) es creciente cuando:

$$1 + h\,Z\,T_{-g,0}(Z) > 0 \tag{7}$$

en el caso que $p > 0.5$ se tiene que $Z_p > 0$ y por lo tanto:

$$Z_p T_{-g,0}(Z_p) > 0 \qquad \forall g$$

Por la simetría de la distribución normal se tiene que:

$$\begin{aligned} \underbrace{Z_{1-p}}\, T_{-g,0}\,\underbrace{(Z_{1-p})} &= \underbrace{Z_{1-p}}\,\frac{e^{-gZ_{1-p}} - 1}{-g} \\ -Z_p\,T_{-g,0}(-Z_p) &= -Z_p\,\frac{e^{gZ_p} - 1}{-g} = Z_p\,T_{g,0}(Z_p) \end{aligned}$$

es decir que $Z\,T_{-g,0}(Z)$ es una función positiva en todo su dominio; por lo tanto, la ecuación (7) depende únicamente del valor que tome h.

- Si $h \geq 0$, la función $T_{g,h}(Z)$ es creciente ya que (7) se verifica de manera inmediata.
- Cuando $h < 0$, la derivada se anula si:

$$Z\,T_{-g,0}(Z) = \frac{1}{|h|}$$

en este caso $T_{g,h}(Z)$ es creciente cuando:

$$Z\,T_{-g,0}(Z) < \frac{1}{|h|}$$

2. Si $g = 0$, se tiene que la función $T_{0,h}(Z) = Ze^{\frac{1}{2}hZ^2}$ es creciente cuando

$$1 + hZ^2 > 0 \tag{8}$$

como Z^2 es positiva en todo su dominio, entonces la ecuación (8) depende únicamente del valor que se le asigne a h.

- Si $h \geq 0$, entonces $T_{0,h}(Z)$ es creciente.
- Cuando $h < 0$, la derivada se anula si

$$Z = \pm\frac{1}{\sqrt{|h|}}$$

luego $T_{0,h}(Z)$ tiene un mínimo local en $Z = -\frac{1}{\sqrt{|h|}}$ y un máximo local en $Z = \frac{1}{\sqrt{|h|}}$. Por lo tanto, $T_{0,h}(Z)$ es creciente siempre que

$$Z^2 < \frac{1}{|h|}$$

2.2. Función de densidad

Ahora se establecerá la función de densidad de la variable aleatoria $Y = T_{g,h}(Z)$ para cualquier valor de g y h. Para ello, se enuncia sin demostración el teorema 1, citado en Apostol (1985).

Teorema 1. Regla de la función inversa

Sea f una función estrictamente creciente y continua en un intervalo $[a, b]$, y sea g la inversa de f. Si existe la derivada $f'(x)$ y no es nula en un punto x de (a, b), entonces la derivada $g'(y)$ también existe y no es nula en el correspondiente punto y, siendo $y = f(x)$. Además, las dos derivadas son recíprocas una de otra; esto es, se cumple

$$g'(y) = \frac{1}{f'(x)}$$

En Jiménez (2004) se emplea este teorema para establecer el siguiente resultado relativo a los percentiles de una variable aleatoria continua.

Proposición 1. *Sea F la función de distribución acumulada de una variable aleatoria continua X, si F' no toma nunca el valor cero, entonces F^{-1} es diferenciable y el valor de su derivada en el punto $p = F(x_p)$ es*

$$\left(F^{-1}\right)'(p) = \frac{1}{F'(x_p)}$$

donde p es el único número que satisface $F(x_p) = p$. En otras palabras,

$$\left(F^{-1}\right)'(F(x_p)) = \frac{d}{dp}x_p = \frac{1}{F'(x_p)} = \frac{1}{f(x_p)} \tag{9}$$

siendo f la función de densidad de la variable aleatoria continua X.

Por lo tanto, si se considera la región donde la transformación $Y = T_{g,h}(Z)$ es una función creciente y diferenciable respecto a Z, y se utiliza la proposición 1, se puede determinar la función de densidad de Y. Como

$$\frac{d}{dp} y_p = \frac{1}{t_{g,h}(y_p)} \tag{10}$$

mediante el empleo de la regla de la cadena se tiene que

$$\frac{d}{dp} y_p = \frac{dy_p}{dZ_p}\frac{dZ_p}{dp}$$

usando la proposición 1 se establece que

$$\frac{d}{dp} Z_p = \frac{1}{\varphi(Z_p)}$$

donde $\varphi(\cdot)$ es la función de densidad normal estándar, entonces

$$\frac{d}{dp} y_p = \frac{1}{\varphi(Z_p)}\frac{dy_p}{dZ_p}$$

Al sustituir la ecuación (6) en esta última expresión se obtiene:

$$\frac{d}{dp} y_p = \sqrt{2\pi}\left[e^{gZ_p} + \frac{h}{g} Z_p(e^{gZ_p} - 1)\right] e^{\frac{1}{2}(h+1)Z_p^2}$$

si esta expresión se reemplaza en (10) se llega a:

$$t_{g,h}(y_p) = \frac{e^{-\frac{1}{2}(h+1)Z_p^2}}{\sqrt{2\pi}\left[e^{gZ_p} + \frac{h}{g} Z_p(e^{gZ_p} - 1)\right]}$$

Esta nueva expresión relaciona las funciones de densidad de las variables aleatorias $Y = T_{g,h}(Z)$ y Z por medio de sus cuantiles y permite una construcción de la función de densidad de Y para cada pareja de parámetros (g, h).

El anterior resultado se puede recopilar en el siguiente teorema.

Teorema 2. *Sea Y una variable aleatoria con distribución $g{-}h$ de Tukey y sea $t_{g,h}(y)$ su función de densidad, entonces*

$$t_{g,h}(y_p) = \frac{e^{-\frac{1}{2}(h+1)Z_p^2}}{\sqrt{2\pi}\left[e^{gZ_p} + \frac{h}{g} Z_p(e^{gZ_p} - 1)\right]} \tag{11}$$

donde y_p y Z_p denotan el p-ésimo cuantil de la transformación $Y = T_{g,h}(Z)$ y de la distribución normal estándar, respectivamente.

En la tabla 1 se presenta la función $t_{g,h}(y_p)$ para algunos valores específicos de g y h. Nótese que el caso V corresponde a la distribución normal estándar. En este artículo se estudiará con más detalle la distribución obtenida en el caso II.

Tabla 1: Funciones de densidad para algunos valores de g y h.

Casos	Valores g	Valores h	Función de densidad
I	**arbitrario**	**arbitrario**	$\frac{e^{-\frac{1}{2}(h+1)Z_p^2}}{\sqrt{2\pi}\left[e^{gZ_p}+\frac{h}{g}Z_p(e^{gZ_p}-1)\right]}$
II	**arbitrario**	0	$e^{\frac{1}{2}g^2}\frac{e^{-\frac{1}{2}(Z_p+g)^2}}{\sqrt{2\pi}}$
III	**positivo**	g	$\frac{e^{-\left[\frac{1}{2}(g+1)Z_p+g\right]Z_p}}{\sqrt{2\pi}\left[1-Z_p(e^{-gZ_p}-1)\right]}$
IV	0	**arbitrario**	$\frac{e^{-\frac{1}{2}(h+1)Z_p^2}}{\sqrt{2\pi}(1+hZ_p^2)}$
V	0	0	$\frac{e^{-\frac{1}{2}Z_p^2}}{\sqrt{2\pi}}$
VI	0	1	$\frac{e^{-Z_p^2}}{\sqrt{2\pi}(1+Z_p^2)}$

2.3. Estimación de los parámetros

Para la estimación de g y h se requieren la mediana y un conjunto de cuantiles simétricos alrededor de la mediana. Estos cuantiles forman las parejas x_p y x_{1-p}, para valores convenientes de p, $(0.5 < p < 1)$. Para estimar g y h, Hoaglin & Peters (1979) plantean la siguiente ecuación:

$$X = A + BY \tag{12}$$

donde A es la mediana de la variable aleatoria X, B es una constante de escala y la variable aleatoria Y se define como en (1). En Jiménez (2004), partiendo de esta ecuación y empleando los resultados obtenidos anteriormente, se establecen relaciones para A, B, g y h. Como

$$x_p = A + By_p, \qquad p > 0.5 \tag{13}$$

de (3) se tiene que:

$$x_{1-p} = A - Be^{-gZ_p}y_p, \qquad p > 0.5 \tag{14}$$

Para lograr la continuidad de (13) y (14) se toma $p = 0.5$ y así se llega a:

$$A = x_{0.5}. \tag{15}$$

Si se multiplica (14) por e^{gZ_p} se obtiene que:

$$x_{1-p}e^{gZ_p} = Ae^{gZ_p} - By_p, \qquad p > 0.5 \tag{16}$$

y al sumar esta expresión con (13) se tiene que:

$$x_p + x_{1-p}e^{gZ_p} = x_{0.5} + x_{0.5}e^{gZ_p}$$
$$(x_{0.5} - x_{1-p})e^{gZ_p} = x_p - x_{0.5}$$

Por lo tanto

$$e^{gZ_p} = \frac{x_p - x_{0.5}}{x_{0.5} - x_{1-p}} = \frac{UHS_p}{LHS_p}, \qquad \forall p > 0.5 \tag{17}$$

donde UHS_p y LHS_p denotan los p-ésimos *upper half-spread* y *lower half-spread*, definidos en Hoaglin et al. (1985).

Por otra parte, si se resta (14) de (13) y se reemplaza y_p, se obtiene:

$$\begin{aligned} x_p - x_{1-p} &= B\big(1 + e^{-gZ_p}\big)y_p \\ &= \frac{B}{g}\big(e^{-gZ_p} + 1\big)\big(e^{gZ_p} - 1\big)e^{\frac{1}{2}hZ_p^2} \\ &= \frac{B}{g}\big(e^{gZ_p} - e^{-gZ_p}\big)e^{\frac{1}{2}hZ_p^2} \end{aligned} \tag{18}$$

Al reemplazar (17) en (18) se tiene que:

$$\begin{aligned} g(x_p - x_{1-p}) &= B\left[\frac{x_p - x_{0.5}}{x_{0.5} - x_{1-p}} - \frac{x_{0.5} - x_{1-p}}{x_p - x_{0.5}}\right] e^{\frac{1}{2}hZ_p^2} \\ &= B\frac{(x_p - x_{1-p})(x_p + x_{1-p} - 2x_{0.5})}{(x_{0.5} - x_{1-p})(x_p - x_{0.5})}\, e^{\frac{1}{2}hZ_p^2} \end{aligned}$$

luego,

$$\begin{aligned} Be^{\frac{1}{2}hZ_p^2} &= g\frac{(x_p - x_{0.5})(x_{0.5} - x_{1-p})}{(x_p - x_{0.5}) - (x_{0.5} - x_{1-p})}, \qquad \forall p > 0.5 \\ &= g\frac{(UHS_p)(LHS_p)}{UHS_p - LHS_p} \end{aligned} \tag{19}$$

Cuando el denominador del término de la derecha de (19) es cero, es decir

$$UHS_p = LHS_p, \qquad \forall p > 0.5$$

en la ecuación (17) se llega a:

$$e^{gZ_p} = 1 \qquad \Longleftrightarrow \qquad g = 0$$

Por otra parte, si se multiplica la ecuación (19) por Z_p se tiene que:

$$\begin{aligned} BZ_pe^{\frac{1}{2}hZ_p^2} &= gZ_p\frac{(UHS_p)(LHS_p)}{UHS_p - LHS_p}, \qquad \forall p > 0.5 \\ &= \frac{(UHS_p)(LHS_p)}{UHS_p - LHS_p}\ln\left(\frac{UHS_p}{LHS_p}\right) \end{aligned}$$

en consecuencia,

$$\lim_{LHS_p \to UHS_p} \frac{(UHS_p)(LHS_p)}{UHS_p - LHS_p}\ln\left(\frac{UHS_p}{LHS_p}\right) = UHS_p \tag{20}$$

Al usar estos resultados se puede reescribir la expresión (19) como:

$$Be^{\frac{1}{2}hZ_p^2} = \begin{cases} g\left[x_{0.5} + \dfrac{x_{0.5}^2 - x_p x_{1-p}}{(x_p - x_{0.5}) - (x_{0.5} - x_{1-p})}\right], & g \neq 0, \\ \dfrac{x_p - x_{0.5}}{Z_p}, & g = 0. \end{cases} \tag{21}$$

En otras palabras, la estimación del parámetro h depende del valor que asume el parámetro g.

Teorema 3. *Sea Y una variable aleatoria con distribución $g{-}h$ de Tukey con $g \neq 0$. Si una variable aleatoria X definida para enteros positivos es expresada como en* (12)*, entonces*

$$x_p \cdot x_{1-p} = x_{0.5}^2 \quad \Longleftrightarrow \quad h = 0 \ \ y \ \ B = Ag \tag{22}$$

Demostración

($\Rightarrow$) Asumamos que $x_p \cdot x_{1-p} = x_{0.5}^2$ (x_p y x_{1-p} son ambos positivos); al multiplicar la expresión (13) por (16) se obtiene:

$$\begin{aligned} x_p \cdot x_{1-p} e^{gZ_p} &= A^2 e^{gZ_p} + ABe^{gZ_p} y_p - ABy_p - B^2 y_p^2 \\ &= A^2 e^{gZ_p} + A(e^{gZ_p} - 1)By_p - B^2 y_p^2 \\ &= A^2 e^{gZ_p} + AgT_{g,0}(Z_p)By_p - B^2 y_p^2 \end{aligned}$$

Al reescribirla se obtiene:

$$(By_p)^2 - AgT_{g,0}(Z_p)By_p + (x_p \cdot x_{1-p} - A^2)e^{gZ_p} = 0, \qquad \forall p > 0.5$$

Si se reemplaza y_p por la expresión (2), y dado que $A^2 = x_{0.5}^2$, se tiene que:

$$\begin{aligned} By_p\big[By_p - AgT_{g,0}(Z_p)\big] &= 0, \qquad \forall p > 0.5 \\ Be^{\frac{1}{2}hZ_p^2}\big[Be^{\frac{1}{2}hZ_p^2} - Ag\big]\big[T_{g,0}(Z_p)\big]^2 &= 0 \end{aligned}$$

Si $p > 0.5$, el último término nunca es cero y además, por la expresión dada en (21), se tiene que:

$$Be^{\frac{1}{2}hZ_p^2} \neq 0, \qquad \forall p > 0.5$$

por lo tanto, la única forma de que dicha expresión sea cero es cuando

$$Be^{\frac{1}{2}hZ_p^2} = Ag \iff e^{\frac{1}{2}hZ_p^2} = \frac{A}{B}g, \qquad \forall p > 0.5$$

Como A, B y g son constantes, entonces la función $e^{\frac{1}{2}hZ_p^2}$ es constante para todo $p > 0.5$, pero esto únicamente sucede si $h = 0$ y por lo tanto

$$B = Ag \tag{23}$$

($\Leftarrow$) Supongamos que $h = 0$ y $B = Ag$; al sustituir en (12) se obtiene

$$X = A + BY = A + AgT_{g,0}(Z) = A\left[1 + g\left(\frac{e^{gZ} - 1}{g}\right)\right] = Ae^{gZ}$$

luego,

$$x_p = Ae^{gZ_p} \qquad \text{y} \qquad x_{1-p} = Ae^{-gZ_p}$$

por lo tanto,

$$x_p \cdot x_{1-p} = A^2 = x_{0.5}^2$$

3. La distribución g de Tukey

En esta sección se deduce la función de densidad para una variable aleatoria Y con distribución g. Esta distribución se obtiene asumiendo que el parámetro $h = 0$ en la expresión (1).

Definición 1. *Sea Z una variable aleatoria con distribución normal estándar y sea g un real (parámetro). La variable aleatoria Y dada por*

$$Y = T_{g,0}(Z) = \begin{cases} \dfrac{e^{gZ} - 1}{g}, & g \neq 0, \\ Z, & g = 0. \end{cases} \tag{24}$$

se dice que tiene distribución g de Tukey.

Para $g = 0$ la función de densidad es muy conocida *(distribución normal Estándar)*. Se establecerá la función de densidad sólo para $g \neq 0$. Si se despeja Z_p de (24) se llega a

$$Z_p = \frac{\ln(1 + gy_p)}{g}, \qquad y_p > -\frac{1}{g} \tag{25}$$

Por otra parte, en la sección 2.2. se obtuvo que la función de densidad para las variables aleatorias con distribución $g{-}h$ de Tukey cuando el parámetro g es arbitrario y $h = 0$, viene dada por:

$$t_{g,0}(y_p) = e^{\frac{1}{2}g^2}\left[\frac{e^{-\frac{1}{2}(Z_p+g)^2}}{\sqrt{2\pi}}\right] = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(Z_p^2+2gZ_p)}$$

Si en esta última ecuación se reemplaza la expresión (25), se obtiene:

$$\begin{aligned} t_{g,0}(y_p) &= \frac{\exp\left\{-\frac{1}{2}\left[\left(\frac{\ln(1+gy_p)}{g}\right)^2 + 2\ln(1 + gy_p)\right]\right\}}{\sqrt{2\pi}} \\ &= \frac{\exp\left\{-\frac{1}{2}\left(\frac{\ln(1+gy_p)}{g}\right)^2 - \ln(1 + gy_p)\right\}}{\sqrt{2\pi}} \\ &= \frac{1}{\sqrt{2\pi}(1 + gy_p)} \exp\left\{-\frac{1}{2}\left(\frac{\ln(1+gy_p)}{g}\right)^2\right\} \end{aligned}$$

luego,

$$t_{g,0}(y) = \frac{\exp\left\{ -\frac{1}{2}\left[\frac{\ln(1+gy)}{g}\right]^2\right\}}{\sqrt{2\pi}(1+gy)}, \qquad y > -\frac{1}{g}$$

Esta función coincide con la obtenida por Caballero (1986), la cual fue establecida empleando la técnica de la función de distribución acumulativa.

La anterior discusión se puede resumir en el teorema que se anuncia a continuación.

Teorema 4. *Sea Y una variable aleatoria con distribución g de Tukey, entonces su función de densidad está dada por*

$$t_{g,0}(y) = \frac{\exp\left\{ -\frac{1}{2}\left[\frac{\ln(1+gy)}{g}\right]^2\right\}}{\sqrt{2\pi}\,(1+gy)} \qquad y > -\frac{1}{g} \tag{26}$$

donde el parámetro g es un real positivo.

Nótese que la función $t_{g,0}(y)$ es una densidad de probabilidad ya que

$$\int_{-\frac{1}{g}}^{\infty} t_{g,0}(y)\,dy = 1$$

Por otra parte, si se despeja de la ecuación (12) la variable y y se reemplaza en la expresión dada en (26), se obtiene:

$$\begin{aligned} t_{g,0}\left(\frac{x-A}{B}\right) &= \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2g^2}\left[\ln\left(1+g\frac{x-A}{B}\right)\right]^2\right\}\left(1+g\frac{x-A}{B}\right)^{-1} \\ &= \frac{1}{\sqrt{2\pi}}\,\frac{\exp\left\{-\frac{1}{2g^2}\left[\ln\left(\frac{g}{B}\left(x+\frac{B}{g}-A\right)\right)\right]^2\right\}}{\frac{g}{B}\left(x+\frac{B}{g}-A\right)} \\ &= \frac{1}{\sqrt{2\pi}}\,\frac{B}{g}\,\frac{\exp\left\{-\frac{1}{2g^2}\left[\ln\left(x+\frac{B}{g}-A\right)-\ln\left(\frac{B}{g}\right)\right]^2\right\}}{x+\frac{B}{g}-A} \end{aligned}$$

Al reemplazar B por Ag se llega a:

$$t_{g,0}\left(\frac{x-A}{Ag}\right) = A\,\frac{1}{x\sqrt{2\pi}}\,\exp\left\{-\frac{1}{2}\left(\frac{\ln x - \ln A}{g}\right)^2\right\} \tag{27}$$

y usando el resultado obtenido por Jiménez (2004), en el cual se relacionan las funciones de densidad de una variable aleatoria X expresada como en (12) y $Y = T_{g,h}(Z)$ por medio de

$$f_X(A+By_p) = \frac{1}{B}t_{g,h}(y_p)$$

Al sustituir (27) en esta expresión, la función de densidad de X queda:

$$f_X(A+By) = \frac{1}{x\sqrt{2\pi}g} \exp\left\{-\frac{1}{2}\left(\frac{\ln x - \ln A}{g}\right)^2\right\} \tag{28}$$

Nótese que esta expresión es semejante a la función de densidad Log-normal con parámetros μ, σ^2, dada por (*véase* el apéndice)

$$f_X(x) = \frac{1}{x\sqrt{2\pi}\sigma \ln C} \exp\left\{-\frac{1}{2}\left(\frac{\log_C x - \mu}{\sigma}\right)^2\right\} \qquad x > 0,\ \sigma > 0 \tag{29}$$

donde $\log_C x$ es el logaritmo de x en base C ($C > 1$). La mayoría de textos que definen la distribución Log-normal lo hacen con $C = e$.

Si en la expresión (28), además de asumir $B = Ag$, se considera que:

$$A = C^{\mu} \qquad \text{y} \qquad g = \sigma \ln C \tag{30}$$

las dos funciones dadas en (28) y (29) resultan idénticas. En otras palabras, si se emplean las constantes dadas en (30), la variable resultante de la transformación $T_{g,0}(Z)$ tiene distribución Log-normal.

Estudios empíricos permiten establecer que un conjunto de datos se puede aproximar a una distribución g de manera precisa cuando se satisface que:

$$\begin{matrix}\text{Error}\\ \text{relativo}\end{matrix} = \left|\frac{C^{\widehat{\mu}} - x_{0.5}}{x_{0.5}}\right| \times 100\,\% < 5\,\% \tag{31}$$

4. Ejemplos

En esta sección se ilustra el procedimiento de estimación de los parámetros A, B y g.

Ejemplo 1. Distribuciones teóricas Log-normales

Se consideran inicialmente los tres casos citados en Herazo (1984, sección 3.4., pp. 45–50) para la distribución Log-normal (con $C = e$).

Tabla 2: Distribuciones Log-normales consideradas por Herazo (1984).

Caso	Función de densidad	Parámetros μ	σ^2
I	Log-normal	0	0.1
II	Log-normal	1	0.5
III	Log-normal	0	1.0

En su trabajo, Herazo (1984) emplea el *programa TEXPER* (construido por el autor) y obtiene los resultados de la tabla 3. Si se asumen las expresiones dadas en (30) y se toma $C = e = 2.718281828$, se obtienen las estimaciones de la tabla 4.

Tabla 3: Valores estimados de A, B y g obtenidos por Herazo (1984).

	Parámetros		**Valores estimados**			
Caso	μ	σ^2	A	B	g	h
I	0	0.1	–	0.3162	0.3162	0
II	1	0.5	–	1.9220	0.7070	0
III	0	1.0	–	1.0000	1.0000	0

Tabla 4: Valores estimados de A, B y g mediante la expresión (30).

	Parámetros		**Valores estimados**			
Caso	μ	σ^2	A	B	g	h
I	0	0.1	1	$\sqrt{0.1}$	$\sqrt{0.1}$	0
II	1	0.5	e	$e\sqrt{0.5}$	$\sqrt{0.5}$	0
III	0	1.0	1	1	1	0

Nótese que las estimaciones presentadas en la tabla 4 coinciden numéricamente con las obtenidas por Herazo (1984).

Ejemplo 2.

Los datos citados por Lee (1992, p. 168), tabla 5, se refieren a un insecticida al cual fueron expuestos 20 insectos hasta su muerte. Los tiempos de sobrevivencia están dados en segundos.

Tabla 5: Tiempo de sobrevivencia de 20 insectos.

3	8	12	19	28
5	9	15	20	30
6	10	15	22	40
7	10	18	25	60

Se modelará la variable aleatoria X, tiempo de sobrevivencia, empleando la distribución g de Tukey.

A partir de estos datos, empleando la metodología expuesta por Hoaglin et al. (1985), se obtiene la tabla 6 de valores literales[1] correspondiente a los cuantiles muestrales de la forma $p = 2^{-k}$, $k = 1, 2, \ldots, 6$.

Este problema fue resuelto por Caballero (1986) quien utilizó el método propuesto por Hoaglin & Peters (1979) y obtuvo la siguiente expresión para la ecuación (12):

$$X = 15 + 9.074\left[\frac{e^{0.597Z} - 1}{0.597}\right] \tag{32}$$

[1]Traducción del término *letter values* definido por Tukey (1977).

Tabla 6: Valores literales correspondientes a los datos de la tabla 5.

k	p	x_p	x_{1-p}
1	$\frac{1}{2}$	15.0	15.0
2	$\frac{1}{4}$	8.5	23.5
3	$\frac{1}{8}$	6.0	30.0
4	$\frac{1}{16}$	5.0	40.0
5	$\frac{1}{32}$	4.0	50.0
6	$\frac{1}{64}$	3.0	60.0

Como la variable aleatoria X sólo asume valores enteros positivos, entonces para usar las estimaciones propuestas en este artículo, primero se verifica si se cumple la condición dada en (22). La tabla 7 muestra que $x_p+x_{1-p} \neq 2x_{0.5}$ y $\sqrt{x_p.x_{1-p}} \simeq x_{0.5}$, entonces se puede asumir que el parámetro $g \neq 0$ y $h = 0$.

Tabla 7: Exploración de los cuantiles para elegir los valores de g y h.

k	1	2	3	4	5	6
p	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
$\frac{x_p+x_{1-p}}{2}$	15.0	16.0	18.0	22.5	27.0	31.5
$\sqrt{x_p \cdot x_{1-p}}$	15.0	14.1	13.4	14.1	14.1	13.4

Para emplear las estimaciones dadas en (30), en el exponente de la expresión (29) se hace el cambio de variable:

$$\frac{\log_C x - \mu}{\sigma} = Z \tag{33}$$

de este modo se tiene que:

$$\begin{aligned} \log_C x &= \mu + \sigma Z \\ \ln x &= (\mu \ln C) + (\sigma \ln C)Z \end{aligned}$$

Dado que $Z \sim N(0,1)$, se tiene además:

$$E[\ln x] = \mu \ln C \quad \text{y} \quad Var[\ln x] = (\sigma \ln C)^2$$

Si se define una nueva variable aleatoria $U = \ln X$ y se calcula la media y la desviación estándar de la variable U, obteniéndose:

$$\overline{U} = \widehat{\mu} \ln C = 2.64428252 \quad \text{y} \quad S_U = \widehat{\sigma} \ln C = 0.74354822$$

se tiene que:

$$A = C^{\widehat{\mu}} = e^{\widehat{\mu} \ln C} = 14.0733$$

valor que coincide con el estimado por Lee (1992). Las otras constantes son:

$$g = \widehat{\sigma} \ln C = 0.74354822 \quad \text{y} \quad B = Ag = 10.46421 \tag{34}$$

Al considerar estas estimaciones, la ecuación (12) queda como:

$$\begin{aligned} X &= 14.0733 + 10.46421 \left[\frac{e^{0.743548Z} - 1}{0.743548} \right] \\ &= 14.0733\, e^{0.743548Z} \end{aligned} \tag{35}$$

En la tabla 8 se dan los valores observados de la variable aleatoria X y los estimados mediante (32) y (35), para algunos valores normales estándar.

Tabla 8: Valores observados y valores estimados mediante las expresiones (32) y (35) de tiempo de sobrevivencia.

p	Z_p	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$
$\frac{1}{64}$	-2.154	3.0	4.00	2.84
$\frac{1}{32}$	-1.863	4.0	4.80	3.52
$\frac{1}{16}$	-1.534	5.0	5.88	4.50
$\frac{1}{8}$	-1.150	6.0	7.45	5.98
$\frac{1}{4}$	-0.674	8.5	9.96	8.52
$\frac{1}{2}$	0.000	15.0	15.00	14.07
$\frac{3}{4}$	0.674	23.5	22.54	23.24
$\frac{7}{8}$	1.150	30.0	30.01	33.10
$\frac{15}{16}$	1.534	40.0	37.78	44.03
$\frac{31}{32}$	1.863	50.0	46.02	56.22
$\frac{63}{64}$	2.154	60.0	54.79	69.81

donde,

$X^{(1)}$: Cuantiles muestrales según tabla 6

$X^{(2)}$: Valores obtenidos utilizando la ecuación (32)

$X^{(3)}$: Valores obtenidos utilizando la ecuación (35)

5. Conclusiones

En este artículo se obtiene una regla de fácil manejo para determinar de manera empírica si el parámetro h puede considerarse igual a cero. Si éste es el caso, se presenta un procedimiento para estimar el valor del parámetro g ($g = \sigma \ln C, C > 1$), de una forma más práctica puesto que puede obtenerse mediante una expresión algebraica más sencilla que los métodos tradicionales.

Recibido: junio de 2004
Aceptado: abril de 2006

Referencias

Apostol, T. M. (1985), *Calculus*, Vol. I, Editorial Reverté S.A., Barcelona.

Caballero, G. (1986), Un estimador del parámetro g de la distribución g de Tukey, Tesis de maestría (estadística), Universidad Nacional de Colombia, Facultad de Ciencias. Departamento de Estadística. Bogotá.

Herazo, C. A. (1984), Distribución $g{-}h$ de Tukey y aplicaciones a las distribuciones de vida, Tesis de maestría (estadística), Universidad Nacional de Colombia, Facultad de Ciencias. Departamento de Estadística. Bogotá.

Hoaglin, D., Mosteller, F. & Tukey, J. (1985), *Exploring Data Tables, Trends, and Shapes*, John Wiley & Sons, Inc., USA.

Hoaglin, D. & Peters, S. (1979), 'Software for exploring distribution shape', *Proceedings of computer science and statistics* .

Jiménez, J. A. (2004), Aproximaciones de las funciones de riesgo del tiempo de sobrevivencia mediante la distribución $g{-}h$ de Tukey, Trabajo de grado (Especialización en Actuaría), Universidad Nacional de Colombia, Facultad de Ciencias. Departamento de Matemáticas. Bogotá.

Lee, E. T. (1992), *Statistical Methods for Survival Data Analysis*, second edn, John Wiley & Sons, Inc., New York.

Martínez, J. (1981), *Some Applications of Robust Scale Estimators*, Doctoral Thesis, Temple University, Department of Statistics. Philadelphia.

Martínez, J. & Iglewicz, B. (1984), 'Some properties of the Tukey g and h family of distributions', *Communications in Statistics: Theory and Methods* **13**(3), 353–369.

Tukey, J. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley, USA.

Apéndice

Si X tiene distribución Log-normal con parámetros μ y σ^2, los cuantiles x_p y x_{1-p} se pueden calcular mediante:

$$x_p = C^{\sigma Z_p + \mu} \quad \text{y} \quad x_{1-p} = C^{-\sigma Z_p + \mu}$$

Para expresar a X como en (12), se encuentra el valor de A mediante (15) y se tiene que:

$$A = x_{0.5} = C^{\sigma Z_{0.5}+\mu} = C^{\mu}$$

Nótese que:

$$x_p \cdot x_{1-p} = (C^{\sigma Z_p+\mu}) \cdot (C^{-\sigma Z_p+\mu}) = C^{2\mu} = x_{0.5}^2 \tag{36}$$

Por otra parte, al calcular UHS_p y LHS_p se llega a:

$$UHS_p = C^{\mu}(C^{\sigma Z_p} - 1) \quad \text{y} \quad LHS_p = C^{\mu}(1 - C^{-\sigma Z_p}) \tag{37}$$

Para obtener el valor de g se reemplazan las expresiones anteriores en (29) y se obtiene:

$$e^{gZ_p} = \frac{C^{\mu}(C^{\sigma Z_p} - 1)}{C^{\mu}(1 - C^{-\sigma Z_p})} = \frac{C^{\sigma Z_p} - 1}{C^{-\sigma Z_p}(C^{\sigma Z_p} - 1)} = C^{\sigma Z_p},$$

entonces al tomar logaritmo natural a ambos lados de esta expresión se obtiene que el valor del parámetro g es $\sigma \ln C$.

Como la distribución Log-normal está definida para $x > 0$ y dado que $x_p \cdot x_{1-p} = x_{0.5}^2$, en virtud del Teorema 3, se tiene que $h = 0$ y $B = Ag$.