ESTIMACIÓN BAYESIANA DEL MODELO DE COSTES HALF-NORMAL CON FRONTERA DETERMINISTA

Jesús Basulto Santos
Departamento de Economía Aplicada I
Universidad de Sevilla
e-mail: basulto@us.es

Fco. Javier Ortega Irizo
Departamento de Economía Aplicada I
Universidad de Sevilla
e-mail: fjortega@us.es

Resumen

El modelo Half-Normal (μ,σ) es un ejemplo de modelo no regular en el parámetro μ, que puede ser aplicado a muestras homogénea de empresas, con output iguales, que busquen minimizar sus costos, es decir a modelos homogéneos de costos con frontera determinista. La falta de regularidad del parámetro μ tiene dos consecuencias: (i) que su estimador máximo verosimilitud no se comporta, para grandes muestras como la teoría sostiene e (ii) la aplicación de la regla de Jeffreys, usada para calcular distribuciones no informativas, no puede ser utilizada en este caso.

En el presente trabajo aplicamos una regla generalizada de Jeffreys, que es también válida para el caso no regular, al modelo Half-Normal (μ,σ), que nos permite dar una solución al problema de estimar los parámetros (μ,σ) desde la aproximación Bayesiana. Ilustramos el trabajo con un primer ejemplo sobre el porcentaje de grasa corporal en una muestra de atletas y un segundo ejemplo sobre el mínimo costo por unidad de output en una muestra de empresa suministradoras de electricidad en Estados Unidos (Greene, 1990).

Palabras clave: Distribución a priori no informativa, Inferencia Bayesiana, Modelo Half-Normal, Modelo no regular, Modelo de costes con frontera determinista.

1. Introducción.

Si Z es una variable aleatoria normal tipificada, que lo indicamos por la expresión de que $Z \sim N(0,1)$, diremos que la variable aleatoria $X = |Z|$ sigue una distribución half-normal tipificada. La distribución half-normal tipificada es una distribución truncada de la variable Z a la que exigimos que sea no negativa.

Una generalización de la variable X es la variable aleatoria $Y = \mu + \sigma X$, que denominaremos variable aleatoria half-normal con parámetros μ y σ. La densidad de probabilidad de la variable aleatoria Y, es

$$f(y/\mu,\sigma) = \frac{2}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2} \right), \quad y \geq \mu, \quad (1)$$

donde $\sigma > 0$ y $-\infty < \mu < \infty$.

Vamos a representar a la distribución half-normal (1) por $Y \sim HN(\mu,\sigma)$. La variable aleatoria $Y^* = \mu - \sigma X$ sigue una distribución half-normal negativa y proviene de truncar una variable normal, $N(\mu,\sigma)$, con $Y^* \leq \mu$. Si consideramos la variable aleatoria $-Y^*$, entonces es fácil probar que se trata de una distribución half-normal (1) con parámetros $(-\mu,\sigma)$,es decir, $-Y^* \sim HN(-\mu,\sigma)$. En consecuencia, su análisis puede ser investigado a partir de la half-normal (1).

La distribución half-normal (1) es un modelo unimodal, siendo el valor de la moda igual a μ, con asimetría a la derecha. Una aplicación de gran interés de la distribución half-normal es a los problemas de fronteras deterministas o estocásticas en economía (Forsund et al, 1980). En el caso de suponer que la frontera es determinista, el problema consiste en minimizar, para una población de unidades de producción homogéneas, los costos para un output fijo. Las unidades de producción con el mismo output y que sus costos sean superiores a los de unidades de producción con igual output y el menor costo, se dice entonces que son unidades menos eficientes. Si ahora suponemos que los costos de las unidades de producción

parten de un mínimo, donde se concentran las unidades más eficientes, con una frecuencia máxima, y suponiendo que los costos van aumentando a medida que las unidades productivas son menos eficientes, y las frecuencias de las unidades productivas van disminuyendo hasta anularse, entonces podemos intentar modelar esta situación con un modelo half-normal. También, el modelo (1) con μ=0 ha sido usado para estimar los tamaños de poblaciones de animales en áreas cerradas por medio de diseños muestrales del tipo "Line Transect" en Rohana et al (1995).

El modelo half-normal (1) es un ejemplo de modelo no regular para el parámetro μ y regular para el parámetro σ (Ortega y Basulto, 2003). La existencia de no regularidad en el primer parámetro conduce a que el estimador máximo verosímil no se comporte, para grandes muestras, como la teoría indica (Rohatgi, 1976; página 384); también la aplicación de la regla de Jeffreys, para calcular las distribuciones no informativas de la inferencia Bayesiana, no puede ser utilizada, en general, en el caso de no regularidad. Ahora bien, en Ortega (2001) y en Ortega y Basulto (2003), se propone una generalización de la regla de Jeffreys, para el caso de un parámetro unidimensional no regular.

Así, en el presente trabajo, aplicaremos esta regla generalizada de Jeffreys para estimar los parámetros del modelo half-normal (μ,σ) por medio de la inferencia Bayesiana. Daremos intervalos probabilísticos a partir de las distribuciones a posteriores de cada uno de los parámetros, comprobando como estos intervalos tienen buenos comportamientos en repeticiones de muestras aleatorias, es decir serán intervalos de confianza, lo que será de gran utilidad, también, para aquellos investigadores que hacen uso de los métodos de inferencia clásicos.

A partir de aquí, en la sección 2, determinamos la distribución a priori conjunta para el modelo (1) por medio de la regla de Jeffreys generalizada. En la sección 3, calculamos las distribuciones a posteriores para cada uno de los parámetros. En la sección 4, obtenemos intervalos probabilísticos en cada una de las distribuciones marginales. En la sección 5 aportamos evidencia de que los intervalos probabilísticos se comportan como intervalos de confianza. El trabajo se ilustra, en la

sección 6, con una primera aplicación sobre el porcentaje de grasa corporal en una muestra de atletas y, una segunda aplicación, sobre la estimación del mínimo coste en empresas suministradores de electricidad en Estados Unidos. Por último, discutimos los resultados del trabajo en la sección 7. Todos los cálculos se han realizado con el programa Mathematical 4.

2. Función de Verosimilitud, distribución a priori no informativa y distribución conjunta a posteriori.

Dada una muestra aleatoria **y** = $(y_1, y_2, ...y_n)$ de n observaciones que proviene del modelo $HN(\mu,\sigma)$, se sigue de (1) que la función de verosimilitud para los parámetros (μ,σ), es

$$L(\mu,\sigma / y) \propto \left(\frac{1}{\sigma}\right)^n \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right)\exp\left(-\frac{n(\mu-\bar{y})^2}{2\sigma^2}\right),\ \mu \le y_{(1)},\ \sigma > 0, \text{(1)}$$

donde $s^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2/(n-1)$, $\bar{y}$ es la media aritmética de los datos e $y_{(1)}$ el valor mínimo de los datos muestrales. Los valores de los parámetros que maximizan esta función de verosimilitud, denominados estimadores máximos verosímiles, son $\hat{\mu} = y_{(1)}$ y $\hat{\sigma} = \left(\sum_{i=1}^{n}\left(y_i - y_{(1)}\right)^2/n\right)^{\frac{1}{2}}$. El estudio del comportamiento de estos estimadores puede verse en Pewsey (2002, 2004). Un estadístico suficiente para el vector paramétrico (μ,σ) es $\left(\bar{y}, s, y_{(1)}\right)$**.** Con lo que los estimadores máximos verosímiles, $(\hat{\mu},\hat{\sigma})$ no son suficientes para el vector paramétrico (μ,σ)**.** La relación de los estimadores máximos verosímiles con el estadístico s^2, es $s^2 = \frac{n}{n-1}\left(\hat{\sigma}^2 - (\hat{\mu}-\bar{y})^2\right)$. Y si consideramos el estadístico auxiliar t, definido por,

$$t = \frac{\left(\bar{y} - y_{(1)}\right)\sqrt{n}}{s},$$

entonces el estadístico $(\bar{y}, s, t)$ es suficiente para (μ, σ). La relación del estadístico t con los estimadores s^2 y $\hat{\sigma}^2$ es $t^2 = \frac{n\hat{\sigma}^2}{s^2} - (n-1)$.

Para aplicar la regla de Jeffreys generalizada vamos a calcular las funciones de verosimilitudes condicionadas a cada parámetro, para así trabajar con parámetros unidimensionales.

Cuando suponemos que σ es conocido, entonces la función de verosimilitud para el parámetro μ, es

$$L(\mu/\sigma, y) \propto \exp\left(-\frac{n(\mu-\bar{y})^2}{2\sigma^2}\right),\ \mu \le y_{(1)}.$$

En este caso los rangos de la variable aleatoria Y están encajados unos dentro de otros, lo que facilita la aplicación de la regla generalizada de Jeffreys. En Ortega y Basulto (2003) probamos que en este caso la distribución no informativa para el parámetro μ se obtiene por la fórmula siguiente,

$$\pi(\mu/\sigma) \propto \left| E\left[\frac{\partial}{\partial\mu}\log L(\mu/\sigma, y)\right]\right|, \qquad (2)$$

donde denominando al logaritmo neperiano de la función de verosimilitud por la función $l(\mu/\sigma, y)$, obtenemos

$$l(\mu/\sigma, y) \propto -(\mu-\bar{y})^2,$$

y aplicando la fórmula (2), resulta la siguiente distribución no informativa,

$$\pi(\mu/\sigma) \propto 1.$$

Bajo el supuesto de considerar σ conocido, las distribución a posteriori de μ es,

$$\pi\left(\mu/\sigma,y\right)=\frac{\frac{1}{\sqrt{2\pi}}\frac{\sqrt{n}}{\sigma}\exp\left(-\frac{n\left(\mu-\bar{y}\right)^2}{2\sigma^2}\right)}{\left(1-\Phi\left(\frac{\left(\bar{y}-y_{(1)}\right)\sqrt{n}}{\sigma}\right)\right)},\ \mu\le y_{(1)},\ (3)$$

donde $\Phi(.)$ es la función de distribución de un modelo $N(0,1)$. Vemos que (3) es una distribución normal con parámetros $\bar{y}$ y $\sigma/\sqrt{n}$ que es truncada en el intervalo $\left(-\infty,y_{(1)}\right]$.

Cuando suponemos que μ es conocido, entonces la función de verosimilitud para el parámetro σ, es

$$L\left(\sigma/\mu,y\right)\propto\left(\frac{1}{\sigma}\right)^n\exp\left(-\sum_{i=1}^{n}\left(y_i-\mu\right)^2/\left(2\sigma^2\right)\right),\ \sigma>0.$$

Al ser ahora el parámetro σ regular, la regla introducida por Ortega y Basulto (2003) para generar no informativas se reduce en este caso a la regla de Jeffreys, cuya fórmula es

$$\pi\left(\sigma/\mu\right)\propto\left|E\left(\left[\frac{\partial}{\partial\sigma}\log L\left(\sigma/\mu,y\right)\right]^2\right)\right|,$$

que operando se obtiene como distribución no informativa,

$$\pi\left(\sigma/\mu\right)\propto\frac{1}{\sigma}$$

Bajo el supuesto de considerar μ conocido, las distribución a posteriori de σ es,

$$\pi\left(\sigma/\mu,y\right)=\frac{1}{\left(\Gamma\left(\frac{n}{2}\right)2^{\frac{n}{2}}\right)}\left(\frac{1}{\sigma}\right)^{n+1}\exp\left(-\sum_{i=1}^{n}\left(y_i-\mu\right)^2/\left(2\sigma^2\right)\right)\left[\sum_{i=1}^{n}\left(y_i-\mu\right)^2\right]^{\frac{n}{2}},\ (4)$$

es decir, se trata de una distribución Gamma Invertida con parámetros $n/2$ y $2/\sum_{i=1}^{n}(y_i-\mu)^2$.

Ahora la distribución conjunta de (μ,σ) a posteriori puede calcularse por la expresión:

$$\pi(\mu,\sigma/y)=\pi(\sigma/\mu,y)\pi(\mu/y),$$

donde $\pi(\sigma/\mu,y)$ está calculada en (4), mientras que la marginal $\pi(\mu/y)$ la obtenemos por medio de la siguiente expresión (Arnold et al, 1999):

$$\pi(\mu/y)=\frac{1}{\int_0^{\infty}\frac{\pi(\sigma/\mu,y)}{\pi(\mu/\sigma,y)}d\sigma}\propto\left[\sum_{i=1}^{n}(y_i-\mu)^2\right]^{-\frac{n}{2}}.$$

Como consecuencia de estos últimos resultados, obtenemos que la distribución conjunta a posteriori, es

$$\pi(\mu,\sigma/y)\propto\left(\frac{1}{\sigma}\right)^{n+1}\exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right)\exp\left(-\frac{n(\mu-\bar{y})^2}{2\sigma^2}\right),\ \mu\le y_{(1)},\sigma>0 \quad (5)$$

que al compararla con la función de verosimilitud (1), vemos que la distribución no informativa conjunta para los parámetros (μ,σ), es la función

$$\pi(\mu,\sigma)\propto\frac{1}{\sigma}.$$

La distribución conjunta a posteriori (5) es siempre es propia, ya que se trata de una distribución normal-gamma truncada a que $\mu\le y_{(1)}$.

3. Distribuciones marginales a posteriores de los parámetros (μ,σ).

La distribución a posteriori marginal del parámetro μ es la siguiente:

$$\pi(\mu/y) \propto \left[\sum_{i=1}^{n}(y_i-\mu)^2\right]^{-\frac{n}{2}} \propto \left[1+\frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right]^{-\frac{n}{2}},\ \mu \le y_{(1)},\ (6)$$

donde la moda, $y_{(1)}$,es precisamente el estimador máximo verosímil. Esta distribución marginal de μ sigue una distribución truncada de un modelo T de Student con esperanza igual a $\bar{y}$, precisión $h = n/s^2$ y grados de libertad (n-1).

Si ahora hacemos la transformación $v = (\bar{y}-\mu)\sqrt{n}/s$, obtenemos,

$$\pi(v/y) \propto \left[1+\frac{v^2}{(n-1)}\right]^{-\frac{n}{2}} \propto \frac{g_{T,n-1}(v)}{1-G_{T,n-1}(t)},\ v>t,\ (7)$$

donde $t = (\bar{y}-y_{(1)})\sqrt{n}/s$ es el estadístico auxiliar introducido en la sección anterior, que además depende de la variable aleatoria Z □ $N(0,1)$ y $g_{T,n-1}(v)$ es la función de densidad de una T de Student con esperanza cero, precisión unidad y con n-1 grados de libertad. La distribución a posteriori (7) es un modelo truncado de una variable aleatoria T de Student con esperanza cero, precisión unidad y con n-1 grados de libertad, con T > t.

La longitud del intervalo $\left[y^*, y_{(1)}\right]$, que contiene a μ con probabilidad $1-\alpha$, es de la forma:

$$\square = y_{(1)} - y^* = (\bar{y} - \frac{t\,s}{\sqrt{n}}) - (\bar{y} - \frac{t^* s}{\sqrt{n}}) = \frac{s}{\sqrt{n}}(t^* - t),$$

donde el intervalo $\left[t^*, t\right]$ contiene a la variable aleatoria v (7) con probabilidad $1-\alpha$, y t^* es función del estadístico auxiliar t. El valor esperado de la longitud del intervalo $\left[y^*, y_{(1)}\right]$, condicionado al estadístico auxiliar t, es de la forma:

$$E[\square / t] = \sigma E\left[\frac{s}{\sigma\sqrt{n}} / t\right]\left(t^* - t\right).$$

Vemos que la longitud esperada del intervalo $\left[y^*, y_{(1)}\right]$, condicionado al estadístico t, depende de $\left(t^* - t\right)$, que disminuye cuando t aumenta, y de $E\left[\frac{s}{\sigma\sqrt{n}} / t\right]$, que también disminuye cuando t aumenta, para cualquier valor de σ. En resumen, el valor esperado de la longitud esperada del intervalo $\left[y^*, y_{(1)}\right]$, condicionado al estadístico t, disminuye cuando el estadístico auxiliar aumenta.

Para un tamaño de n =50, y con 5000 repeticiones, recogemos en el siguiente gráfico la relación entre $E\left[\frac{s}{\sigma\sqrt{n}} / t\right]$ y el estadístico auxiliar t,

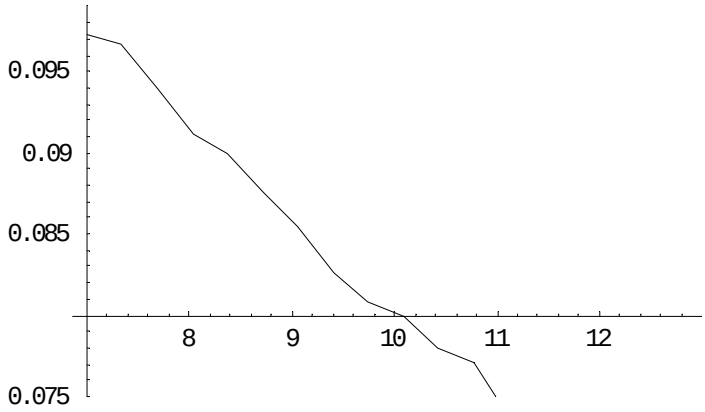

Gráfico 1: Relación entre $E\left[\left(s/(\sigma\sqrt{n})\right)\middle| t\right]$ y el estadístico auxiliar t

donde hemos estimado las esperanzas matemáticas por simulación para intervalos de anchura 0.05 de los valores del estadístico auxiliar t, entre los valores de 7 a 11. Vemos que el gráfico confirma la disminución de la esperanza $E\left[\frac{s}{\sigma\sqrt{n}} / t\right]$ cuando aumenta el estadístico auxiliar t.

La distribución a posteriori marginal del parámetro σ es la siguiente:

$$\pi(\sigma/y) \propto \left(\frac{1}{\sigma}\right)^n \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right)\Phi\left(-\frac{(\bar{y}-y_{(1)})\sqrt{n}}{\sigma}\right),\ \sigma>0\,,\ (8)$$

donde $\Phi(.)$ es la función de distribución de un modelo $N(0,1)$. Haciendo la transformación $w=\frac{(n-1)s^2}{\sigma^2}$, la distribución (8) se transforma en una Ji-cuadrado afectada (Rao,1994) por la siguiente función peso:

$$\Phi\left(-\frac{t\sqrt{w}}{\sqrt{n-1}}\right),$$

es decir, la distribución de w es:

$$\pi(w/y) \propto w^{\frac{n-1}{2}-1}\exp\left(-\frac{w}{2}\right)\Phi\left(-\frac{t\sqrt{w}}{\sqrt{n-1}}\right),\ \text{w>0.}\ (9)$$

4. Intervalos Probabilísticos para los parámetros μ y σ.

Un intervalo $\left[y^*, y_{(1)}\right]$ que contiene al parámetro μ con probabilidad $1-\alpha$ debe verificar la propiedad siguiente:

$$\Pr\left[y^* < \mu < y_{(1)}/y\right] = \int_{y^*}^{y_{(1)}} \pi_\mu(u/y)du = 1-\alpha\,,$$

donde $\pi_\mu(u/y)$ es valor de la distribución a posteriori marginal (6) para $\mu = u$.

Si ahora consideramos la variable aleatoria siguiente,

$$F_\mu(\mu^*/y) = \int_{-\infty}^{\mu^*} \pi_\mu(u/y)du\,,\ \ (10)$$

donde μ^* es el valor verdadero del parámetro μ, y suponemos que la variable aleatoria (10) sigue un modelo uniforme (0,1), vamos a probar que el intervalo

bayesiano $\left[y^*, y_{(1)} \right]$ es un intervalo de confianza con coeficiente de confianza igual a $1-\alpha$. Veamos el resultado:

$$\Pr\left[y^* < \mu^* < y_{(1)} / \mu^*, \sigma \right] = \Pr\left[F_\mu\left(y^* / y\right) < F_\mu\left(\mu^* / y\right) < F_\mu\left(y_{(1)} / y\right) / \mu^*, \sigma \right] =$$
$$= \Pr\left[\alpha < F_\mu\left(\mu^* / y\right) < 1 / \mu^*, \sigma \right] = 1-\alpha \qquad .(11)$$

Si hacemos la transformación $v = \dfrac{(\bar{y}-\mu)\sqrt{n}}{s}$, la variable aleatoria (10) se puede calcular por la expresión siguiente:

$$F_\mu\left(\mu^* / y\right) = 1 - \int_t^{v^*} \pi_v\left(u / y\right) du ,$$

donde $v^* = \dfrac{\left(\bar{y}-\mu^*\right)\sqrt{n}}{s}$ y $\pi_v\left(u / y\right)$ es el valor de la función (7) para v = u. Así, la función de distribución,

$$F_v\left(v^* / y\right) = \int_t^{v^*} \pi_v\left(u / y\right) du , \quad (12)$$

es uniforme (0,1) como consecuencia de suponer que (10) sea uniforme (0,1).

En la sección siguiente vamos a discutir el supuesto de que la variable aleatoria (12) sea uniforme (0,1). Ahora bien, (12) es una variable que es idéntica para todos los valores de los parámetros μ y σ, con lo que basta que calculemos sus valores a partir de muestras aleatorias generadas del modelo (1) con μ=0 y σ=1.

Un resultado más interesantes es suponer que la variable aleatoria (10), cuando se condiciona al estadístico auxiliar “t”, sigue un modelo uniforme (0,1), para todo valor del estadístico auxiliar t. De nuevo, este supuesto conduce a probar que los intervalos de confianza condicionados al estadístico auxiliar “t” contienen al parámetro μ con una probabilidad de $1-\alpha$. Este supuesto de uniformidad condicionada al estadístico “t” es equivalente a suponer que el estadístico (12) sea,

condicionado al estadístico auxiliar “t”, uniforme (0,1), En la sección siguiente volveremos sobre estos supuestos de uniformidad.

Un intervalo central que contenga al parámetro σ con probabilidad $1-\alpha$ es de la forma $[\sigma_1, \sigma_2]$ y verifica las siguientes condiciones:

$$\Pr[\sigma < \sigma_1 / y] = \int_0^{\sigma_1} \pi_\sigma (s/y) ds = \int_{w_1}^{\infty} \pi_w (s/y) ds = \frac{\alpha}{2}$$

y

$$\Pr[\sigma > \sigma_2 / y] = \int_{\sigma_2}^{\infty} \pi_\sigma (s/y) ds = \int_0^{w_2} \pi_w (s/y) ds = \frac{\alpha}{2},$$

donde $w_i = (n-1)s^2 / \sigma_i^2$, para i = 1,2; $\pi_\sigma (./y)$ es la función de densidad a posteriori (8) de σ, mientras que $\pi_w (./y)$ es la de la variable transformada w (9).

Si para el valor $\sigma = \sigma^*$, suponemos que la variable aleatoria

$$\int_0^{\sigma^*} \pi_\sigma (s/y) ds, \quad (13)$$

es uniforme (0,1), entonces el intervalo $[\sigma_1, \sigma_2]$ es un intervalo de confianza que contiene al verdadero valor σ^* de σ con un coeficiente de confianza de $1-\alpha$. La demostración es equivalente a la que hicimos en el caso del parámetro μ.

También, suponer que la variable aleatoria $\int_0^{\sigma^*} \pi_\sigma (s/y) ds$ es uniforme (0,1) es equivalente a que la variable aleatoria $\int_0^{w^*} \pi_w (s/y) ds$ sea uniforme (0,1), donde $w^* = (n-1)s^2 / (\sigma^*)^2$.

Un resultado más interesantes es suponer que la variable aleatoria (13), cuando se condiciona al estadístico auxiliar “t”, sigue un modelo uniforme (0,1), para

todo valor del estadístico auxiliar t. De nuevo, este supuesto conduce a probar que los intervalos centrados de confianza condicionados al estadístico auxiliar "t" contienen al parámetro σ con una probabilidad de $1-\alpha$.

5. Los supuestos de Uniformidad sobre las funciones de distribución a posteriores para cada uno de los parámetros μ y σ.

Para analizar el comportamiento del intervalo bayesiano del parámetro μ, vamos a seleccionar muestras aleatorias, de tamaño n, del modelo (1) con μ=0 y σ=1, calculando para cada muestra el valor de la variable aleatoria (12), siendo dicho valor igual a,

$$F_v\left(v^* / y\right) = \frac{G_{T,n-1}\left(\frac{\overline{y}\sqrt{n}}{s}\right) - G_{T,n-1}\left(t\right)}{\left(1-G_{T,n-1}\left(t\right)\right)} \quad , (14)$$

donde $t = \left(\overline{y} - y_{(1)}\right)\sqrt{n}\Big/ s$ es un estadístico auxiliar.

Con los valores calculados para cada una de las m muestras de tamaño n del modelo (1) con μ=0 y σ=1, haremos un gráfico recogiendo en el eje OX los valores de $F_v\left(v^* / y\right)$, ordenados de menor a mayor, y en el eje OY las correspondientes frecuencias relativas acumuladas. Ahora, la hipótesis de que la variable aleatoria (14) sigue un modelo uniforme (0,1), será sustentada a medida que los puntos calculados se aproximen a la diagonal principal del primer cuadrante.

Un ejemplo es el siguiente gráfico,

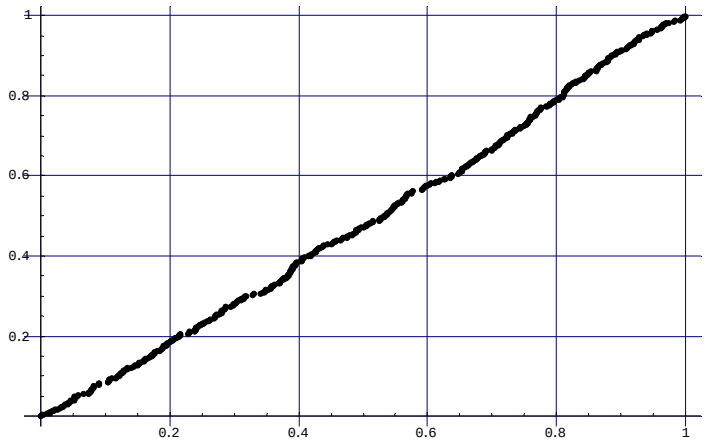

Gráfico 2. Ajuste a un modelo uniforme. n = 2, m = 500.

donde el tamaño n = 2 y el número de muestras replicadas es m = 500. Se observa que los puntos se aproximan a la diagonal principal, lo que sustenta la hipótesis de uniformidad (0,1). El estadístico de la prueba de Kolmogorov-Smirno calculado vale 0,0416, que es menor que el valor teórico 0,061 para un nivel de significación del 5%, lo que sustenta la hipótesis de uniformidad (0,1).

Recogemos en la Tabla 1 los valores de la prueba de Kolmogorov-Smirnov para distintos m y n.

Tabla 1. Valores del estadístico de Kolmogorov-Smirnov (KS)

m\ n	2	4	8	15	20	50	5%
300	0,0416	0,039	0,028	0,0369	0,030	0,0679	0,0785
500	0,0349	0,0325	0,0239	0,0413	0,0515	0,0341	0,061
1000	0,0326	0,0244	0,0185	0,0301	0,0329	0,023	0,043

En la última columna de la Tabla 1 recogemos los valores teóricos del la prueba para un nivel de significación del 5%. Vemos que en todos los casos aceptamos la hipótesis de uniformidad (0,1).

A continuación, recogemos en la Tabla 2 los valores de la prueba Kolmogorov-Smirnov sobre la hipótesis de que la variable aleatoria (14) condicionada al estadístico $t = \left(\overline{y} - y_{(1)}\right)\sqrt{n}\,/\,s$, sigue un modelo uniforme (0,1). La Tabla 2 recoge para una muestra de intervalos de valores del estadístico t, el valor de la prueba de Kolmogorov-Smirnov (KS) para n = 4 y m = 500. En la Tabla 2 recogemos las marcas de clase de los intervalos del estadístico t cuyas amplitudes se han tomado igual a 0.02.

Tabla 2. Valores del estadístico de Kolmogorov-Smirnov (KS)

Intervalos de t	KS	Intervalos de t	KS
1.02	0.026	2,00	0.038
1.06	0.041	2.02	0.025
1,10	0.045	2.06	0.031
1.16	0.024	2.16	0.027
1.26	0.032	2.26	0.028
1.36	0.044	2.36	0.020
1.40	0.035	2.46	0.031
1.46	0.033	2.50	0.028
1.50	0.038	2.56	0.049
1.56	0.047	2.62	0.026
1.62	0.053	2.66	0.032
1.66	0.036	2.76	0.035
1.76	0.022	2.86	0.025
1.82	0.029	2.92	0.033
1,86	0.030	2.96	0.024
1.96	0.042	3,00	0.032

El valor teórico del estadístico de la prueba de Kolmogorov-Smirnov (KS), para el nivel de significación del 5%, es 0.061. Vemos que en todos los casos aceptamos la hipótesis de uniformidad (0,1) de la variable (14) condicionadas al estadístico t.

Para analizar el comportamiento del intervalo bayesiano del parámetro σ, vamos a seleccionar muestras aleatorias, de tamaño n, del modelo (1) con μ=0 y σ=1, calculando para cada muestra el valor de la variable aleatoria (13), que es igual a la integral siguiente,

$$F_w\left(w^* / y\right)=\frac{\int_0^{w^*} w^{\frac{n-1}{2}-1}\exp\left(-\frac{w}{2}\right)\Phi\left(-\frac{t\sqrt{w}}{\sqrt{n-1}}\right)dw}{\int_0^{\infty} w^{\frac{n-1}{2}-1}\exp\left(-\frac{w}{2}\right)\Phi\left(-\frac{t\sqrt{w}}{\sqrt{n-1}}\right)dw} \quad , (15)$$

donde $w^*=(n-1)s^2$, ya que las muestras provienen del modelo (1) con parámetro μ=0 y σ=1. Para n = 4 y m = 300, presentamos el gráfico siguiente cuya construcción es la misma que la del gráfico 2,

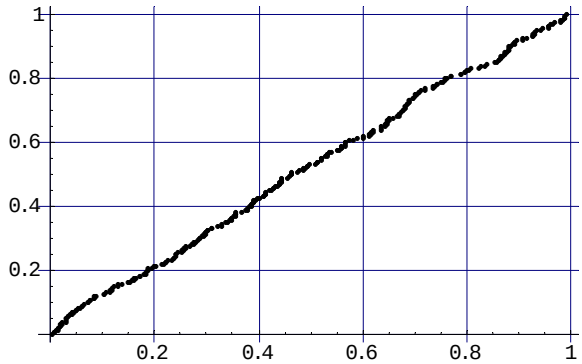

Gráfico 3. Ajuste a un modelo uniforme. n = 4, m = 300.

el valor de la prueba de Kolmogorov-Smirnov es, KS = 0.0548, que sustenta la hipótesis de uniformidad (0,1) al estar por debajo del valor de 0.061 (5%).

Recogemos en la Tabla 3 los valores de la prueba de Kolmogorov-Smirnov para distintos m y n.

Tabla 3. Valores del estadístico de Kolmogorov-Smirnov (KS)

m\ n	3	4	8	15	20	50	5%
300	0.0542	0.0548	0.0385	0.0379	0.0550	0.0420	0,0785
500	0.0438	0.0221	0.0236	0.0397	0.0418	0.0200	0,061
1000	0.0286	0.0200	0.0263	0.0255	0.0289	0.0362	0,043

En la última columna de la tabla 3 recogemos los valores teóricos del la prueba para un nivel de significación del 5%. Vemos que en todos los casos aceptamos la hipótesis de uniformidad (0,1).

A continuación, recogemos en la Tabla 4 los valores de la prueba Kolmogorov-Smirnov sobre la hipótesis de que la variable aleatoria (15) condicionada al estadístico $t = \left(\overline{y} - y_{(1)}\right)\sqrt{n}\Big/ s$, sigue un modelo uniforme (0,1). La Tabla 4 recoge para una muestra de intervalos de valores del estadístico t, el valor de la prueba de Kolmogorov-Smirnov (KS) para n = 6 y m = 300. En la Tabla 4 recogemos las marcas de clase de los intervalos del estadístico t cuyas amplitudes se han tomado igual a 0.02.

Tabla 4. Valores del estadístico de Kolmogorov-Smirnov (KS)

Intervalos de t	KS	Intervalos de t	KS
1.16	0.0702	2.16	0.0353
1.26	0.0629	2.26	0.0461
1.36	0.0614	2.36	0.0584
1.40	0.0400	2.46	0.0305
1.46	0.0329	2.50	0.0441
1.50	0.0568	2.56	0.0617

1.56	0.0440	2.62	0.0381
1.62	0.0540	2.66	0.0557
1,86	0.0586	3,10	0,0481
1.96	0.0473	3,20	0,0747
2,00	0.0597	3,30	0,0496
2.02	0.0332	3,40	0,0457
2.06	0.0770	3,50	0,0406
2.76	0.0386	3,60	0,0626
2.86	0.0378	3,70	0,0357
2.92	0.0421	3,80	0,0443
2.96	0.0536	4,00	0,0468
3,00	0.0502	4,5	0,0413

6. Un primer ejemplo ilustrativo.

Los datos que vamos a analizar corresponden a n = 102 atletas, que han sido entrenados por el Instituto Australiano de Deportes, y la variable de interés mide el porcentaje de grasa corporal de cada uno de los atletas. Lo datos han sido tomados del libro de Cook y Weisberg (1994).

En la Tabla 5 hemos recogido los estadísticos que serán necesarios en nuestro análisis.

Tabla 5

Estadísticos	Valores
$\overline{y}$	9,25088 %
$y_{(1)}$	5,63 %
s	3,1846 %
$\hat{\sigma}$	4,8118 %
t	11,4833

A partir de simular 3000 muestras del modelo (1) con $\mu = 0$ y $\sigma = 1$ estimamos el recorrido del estadístico auxiliar t como (10,49; 16,94), observando que la probabilidad de que el estadístico auxiliar t sea menor que 11,4833 es 0,025, un valor pequeño que nos hace dudar sobre la hipótesis de que los datos siguen el modelo (1).

La Distribución a posteriori (6) para μ, es

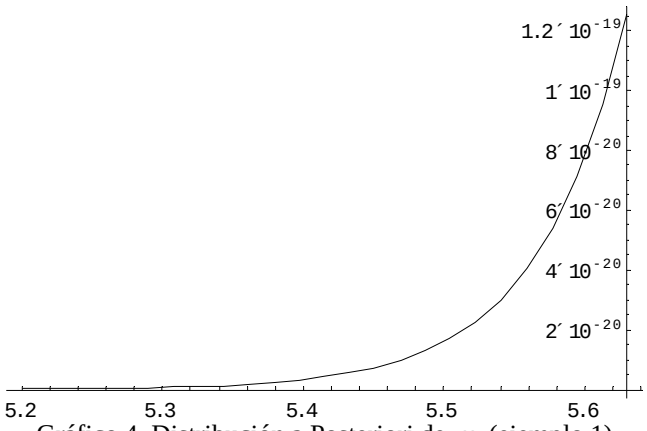

Gráfico 4. Distribución a Posteriori de μ (ejemplo 1)

que se trata de una T de Student, con esperanza 9,25088%, precisión igual a 10,0575 y grados de libertad 101, truncada en el intervalo $(-\infty, 5{,}63]$. Un intervalo probabilístico, con probabilidad igual a 0.95, es [5,441, 5.63]. Es decir, el porcentaje de grasa corporal mínimo teórico, para estos atletas, está entre 5,441% y 5,63%, con una probabilidad del 95%. Este intervalo también se puede considerar como un intervalo de confianza con un coeficiente de confianza del 95%. El intervalo de confianza, [5,41 ; 5,63], calculado por Pewsey (2002, 2004), es prácticamente igual al calculado por nosotros, aunque un poco más amplio.

Para n = 102, y con una simulación de 576 muestras, hemos estimado la esperanza $E\left(\frac{s}{\sigma\sqrt{102}} / t = 11,4833\right) = 0,0649628$, para valores del estadístico auxiliar t en el intervalo [11,4333; 11,5333]. Los valores observados del estadístico $\frac{s}{\sigma\sqrt{102}}$ han oscilado entre 0,050634 a 0,0810704.

El valor esperado de la longitud del intervalo probabilístico $\left[y^*, y_{(1)}\right]$, condicionado a que t = 11,4833 (que pertenezca al intervalo [11,4333; 11,5333]), es en el presente ejemplo,

$$E[\square / t] = \sigma\ E\left[\frac{s}{\sigma\sqrt{n}} / t\right]\left(t^* - t\right) = \sigma\, 0,0649628.0,59938 = \sigma\ 0,03893,$$

donde vemos que es una función del parámetro σ que estimaremos a partir de la estimación de σ.

La Distribución a posteriori (8) para σ, es

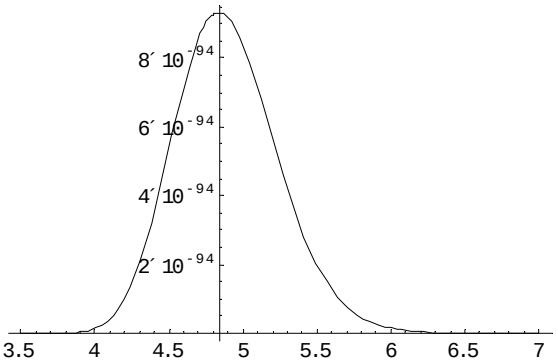

Gráfico 5. Distribución a Posteriori de σ (ejemplo 1)

donde un intervalo probabilístico centrado, con probabilidad igual a 0.95, es [4,269 ; 5.64]. Es decir, el parámetro σ está entre 4,269% y 5,64%, con una probabilidad del 95%. Este intervalo también se puede considerar como un intervalo de confianza con un coeficiente de confianza del 95%. El intervalo de confianza, [4,25 ; 5,61], calculado por Pewsey (2002,2004), es prácticamente igual al calculado por nosotros. La moda de esta distribución a posteriori es 4,84 %, que es prácticamente igual al estimador máximo verosímil 4,8118 %.

Ahora, un intervalo probabilístico para $E[\square / t = 11,4833]$, es [0,1662; 0,2196], que contiene la amplitud observada 0,180 del intervalo bayesiano [5,441, 5.63] de μ. Vemos que además de que el valor del estadístico auxiliar t =11,4833 es pequeño, cuando se compara con su máximo observado de 16,94, también, para este valor de t, el valor observado de la amplitud del intervalo [5,441, 5.63] podría haber sido más pequeño, ya que podríamos haber observado un valor tan pequeño como 0,146 (este valor ha sido estimado por el producto 4,8118 *0,050634*0,59938, estimando σ por 4,8118 y tomando el valor más pequeño observado de $\frac{s}{\sigma\sqrt{102}}$ en las muestras generadas con t =11,4833 o con mayor precisión, con t perteneciente al intervalo [11,4333; 11,5333]).

Un análisis, que nos informa sobre la validez de la hipótesis de que los datos siguen el modelo (1), es el proporcionado por el siguiente gráfico Q-Q:

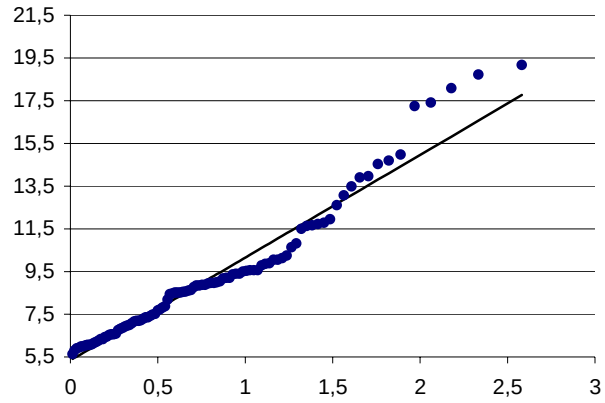

Gráfico 6. Ajuste Q-Q de los datos al modelo Half-Normal (ejemplo 1)

donde observamos que el ajuste se deteriora para valores grandes de la variable, con lo que el modelo (1) no ajusta todo el recorrido de la variable considerada. El gráfico Q-Q nos informe de que los datos tienen la cola de la derecha del histograma empírico por debajo de la cola teórica del modelo (1).

7. Un segundo ejemplo ilustrativo.

Los datos que vamos a analizar corresponden a una muestra de 119 empresas suministradoras de electricidad en Estados Unidos. Lo datos han sido tomados del artículo de Greene, W. H. (1990), donde de sus 123 empresas hemos eliminado 4 empresas. La variable Y bajo estudio es ln[C/pf]-ln[Q], donde C son los costes (10^6 $), pf es el precio del fuel y Q es la cantidad de ouput (10^6 Kwh). Suponemos que los costes por unidad de output, la variable Y, siguen un modelo $HN(\mu,\sigma)$.

En la Tabla 6 hemos recogido los estadísticos que serán necesarios en nuestro análisis.

Tabla 6

Estadísticos	Valores
$\overline{y}$	-8,6145
$y_{(1)}$	-9,0769
s	0,23663
$\hat{\sigma}$	0,501902
t	21,3185

El valor tan elevado del estadístico t nos lleva a sospechar de que el modelo (1) no es muy adecuado en este ejemplo.

La Distribución a posteriori para μ, es

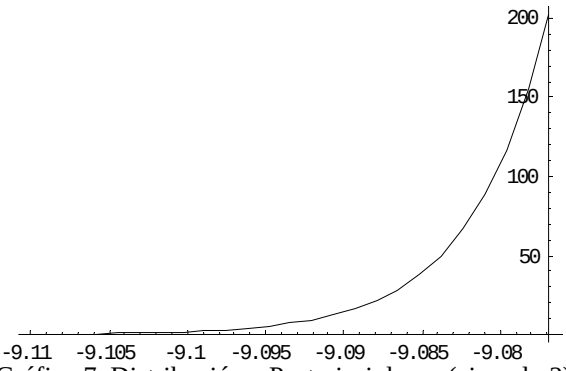

Gráfico 7. Distribución a Posteriori de μ (ejemplo 2)

que como sabemos es una t de Student truncada, con esperanza -8,6145 , precisión 2121,24 y grados de libertad 118. Un intervalo Bayesiano con probabilidad del 95% de contener μ es [-9,091;-9,076].

La Distribución a posteriori para σ, es

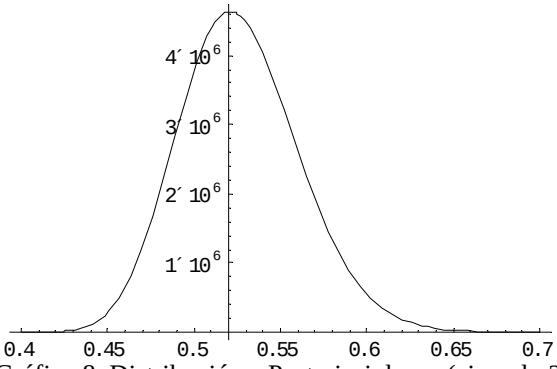

Gráfico 8. Distribución a Posteriori de σ (ejemplo 2)

donde la moda es 0,521, que está próxima al valor del estimador máximo verosímil $\hat{\sigma}$=0,501902. Un intervalo simétrico con probabilidad 0,95 de contener σ es [0,464; 0,6001].

Un análisis, que nos informa sobre la validez del modelo (1) , es el proporcionado por el siguiente gráfico Q-Q:

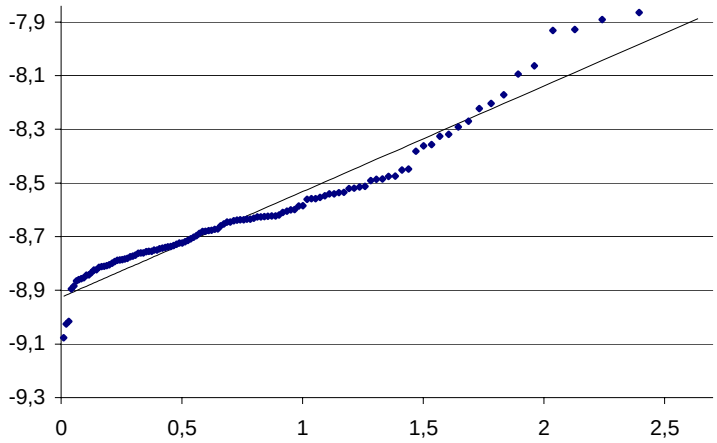

Gráfico 9. Ajuste Q-Q de los datos al modelo Half-Normal (ejemplo 2)

donde observamos que el ajuste se deteriora para valores grandes y pequeños de la variable, con lo que el modelo (1) no ajusta todo el recorrido de la variable considerada. El gráfico Q-Q nos informa de que los datos tienen la cola de la derecha del histograma empírico por debajo de la cola teórica del modelo (1), en cambio la cola de la izquierda está por encima del modelo teórico.

8. Discusión.

En el presente trabajo hemos estimado los parámetros del modelo Half-Normal (1) por medio de la inferencia Bayesiana. Al ser el parámetro μ no regular, hemos aplicado la regla de Jeffreys generalizada (Ortega y Basulto, 2003) para obtener una distribución no informativa para el vector de parámetros (μ,σ). A partir de calcular las distribuciones a posteriores de cada uno de los parámetros, hemos obtenido intervalos de probabilidad para cada parámetro, aportando evidencia sobre el comportamiento de estos intervalos como intervalos de confianza en muestras repetidas. En el proceso de nuestro análisis hemos encontrado un estadístico auxiliar, que junto con los estimadores máximos verosímiles, definen un estadístico suficiente para el vector de parámetros (μ,σ). Este estadístico auxiliar lo hemos usado para criticar la validez del modelo (1), así como para evaluar la importancia de cada muestra sobre la precisión con que hemos estimado cada uno de los parámetros. También, este estadístico auxiliar ha sido usado para estudiar el comportamiento de

los intervalos probabilísticos cuando se condicionado a un valor observado de dicho estadístico.

El modelo Half-Normal (1) admite generalizarlo a un modelo no homogéneo al introducir variables explicativas relacionadas con la esperanza matemática de la variable Y, por ejemplo, $\mu_t = \alpha + \beta x_t$ define la parte determinista de un modelo de regresión con perturbaciones del tipo $HN(0,\sigma)$. Otra generalización del modelo (1) es el modelo truncado de la forma $Y \square N(\mu,\sigma)$ con $Y \geq \lambda$, y cuando el parámetro $\lambda = 0$ entonces se reduce al modelo (1). Cuando $\lambda \leq \mu$, entonces el modelo truncado tiene una moda igual a λ, y así el comportamiento de la función de densidad arranca de un valor, a la izquierda de λ, hasta alcanzar un máximo en el valor de λ para comenzar, a continuación, a disminuir. El modelo truncado ha sido propuesto por Stevenson (1980) para modelar la frontera determinista o estocástica en problemas de estimación de la eficiencia económica en muestras de empresas.

Por último, el modelo Half-Normal (1) es usado en problemas de frontera estocástica (Forsund et al, 1980), que para nuestro caso homogéneo (1) (un modelo de frontera estocástica) sería de la forma $Y = \mu + \varepsilon + \sigma|Z|$, donde $Z \square N(0,1)$. Una hipótesis es suponer que las variables aleatorias ε y $|Z|$ son independientes, donde la variable aleatoria ε es de la forma $\varepsilon \square N(0,\omega)$, siendo ω su varianza. Este último modelo permite que todos sus parámetros sean regulares y, así, se evita el problema del estimador máximo verosímil ante la no regularidad. Ahora bien, este nuevo modelo presenta el problema de cómo separar los efectos de las variables aleatorias ε y $|Z|$ (Forsund et al, 1980) sobre la variable dependiente.

Bibliografía.

1. Arnold, B.R.; Castillo, E. and Sarabia, J.M. (1999), Conditional Specification of Statistical Models. Springer Series in Statistics, Springer.Verlag, New York.

2. Cook, R.D. and Weisberg, S. (1994), An Introduction to Regression Graphics. Wiley: New York.

3. Greene, W. H. (1990), A Gamma-Distributed Stochastic Frontier Model. Journal of Econometric. 46, 141-163.

4. Forsund, F.R.; Nnox Lovell, C.A. and Schmidt, P. (1980), *A Survey of Frontier Production Function and of Their Relationship to Efficiency Measurement.* Journal of Econometrics, 13, 5-25.

5. Ortega, F.J. (2001), Obtención de Distribuciones a Priori Infomativas Usando Medidas de Información. Aplicación a la Evaluación de la Revistas Científicas. Tesis Doctoral Universidad de Sevilla. Spain.

6. Ortega, F.J. y Basulto, J. (2003), *Distribuciones a priori unidimensionales en Modelos No Regulares*. Estadística Española. Vol. 45, Núm. 154, pp. 363-383.

7. Pewsey, A. (2002), *Large-Sample Inference for the Geneal Half-Normal Distribution*. Communication in Statistics. Theory and Methods. Vol. 33, No. 2, pp. 197-204.

8. Pewsey, A. (2004), *Improved Likelihood Based Inference for the Geneal Half-Normal Distribution*. Communication in Statistics. Theory and Methods. Vol. 33, No. 2, pp. 197-204.

9. Pitman, E.J. (1979), Some Basic Theory for Statistical Inference, Chapman and Hall, London.

10. Rao, C.R. (1994), Estadística y Verdad. Aprovechando el Azar. PPU. Barcelona.

11. Rohana. J. Karunamuni and Terrance J. Quinn II (1995), *Bayesian Estimation of Animal Abundance for line Transect Sampling.* Biometrics, 51, pp. 1325-1337.

12. Rohatgi, V.K. (1976), An Introduction to Probability Theory and Mathematical Statistics. Wiley.

13. Stevenson, R.E. (1980), *Likelihood Functions for Generalized Stochastic Frontier Estimation.* Journal of Econometrics, 13, 57-66.