ESTADISTICA ESPAÑOLA
Vol. 33, Núm. 127, 1991, págs. 291 a 303

# Un método gráfico de análisis de la hipótesis de normalidad

SANTIAGO VELILLA CERDAN
Departamento de Economía
Universidad Carlos III de Madrid

## RESUMEN

En este trabajo se propone un método gráfico de análisis de la hipótesis de normalidad cuando los datos proceden de una muestra aleatoria. Las propiedades del método se ilustran con tres ejemplos de datos reales.

*Palabras clave:* Función cuantílica; método kernel de estimación de densidades; transformación Box-Cox.

*Clasificación AMS:* 62F05; 62G30.

## 1. INTRODUCCION

Sea X una variable aleatoria con función de distribución desconocida F. Sea $\{h(.,\lambda)\}$ una familia de transformaciones indexada por el parámetro $\lambda \in \Lambda$, donde $\Lambda$ es un subconjunto no vacío de $\mathbb{R}^m$. Un método para modelar F es suponer que, para algún $\lambda \in \Lambda$ desconocido, $h(X,\lambda) \sim N(\mu,\sigma^2)$. Cuando $m=1$, una familia usual de transformaciones es la familia de Box-Cox (1964):

$$X^{(\lambda)} = \begin{cases} \dfrac{X^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log X, & \lambda = 0. \end{cases} \qquad (1.1)$$

En (1.1), se supone que $X > 0$. De lo contrario, X se reemplaza por $X+c$, donde $c$ es tal que $X+c > 0$. El modelo es, entonces,

$$X^{(\lambda)} \sim N(\mu,\sigma^2). \qquad (1.2)$$

Dada una muestra aleatoria $X_1, \ldots, X_n$ de F, el modelo (1.2) puede usarse para construir un procedimiento gráfico de análisis de la hipótesis de normalidad. La sección 2 resume los aspectos principales de la teoría de los tests de normalidad. La sección 3 motiva y presenta las características y propiedades del nuevo método gráfico. La sección 4 ilustra las aplicaciones con ejemplos basados en datos reales. La sección 5 contiene los comentarios finales.

## 2. TESTS DE NORMALIDAD

En esta sección se comentan, brevemente, las propiedades y características de algunos tests de normalidad.

### 2.1 Test de Shapiro y Wilk

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una función de distribución F. Sean $X=(X_{(1)}, \ldots, X_{(n)})$, el vector de estadísticos de orden de la muestra, $m \in \mathbb{R}^n$ el vector de medias y $V_0$ la matriz $n \times n$ de covarianzas de los estadísticos de orden de una muestra de tamaño $n$ de una N(0,1). Un gráfico de probabilidad normal es un gráfico $(X_{(j)}, m_j)$. La linealidad del gráfico sugiere la normalidad de la muestra. La extracción de conclusiones de un gráfico conlleva, de forma inevitable, un cierto gráfico de subjetividad. Es natural, por tanto, proponer medidas numéricas que cuantifiquen el

comportamiento del gráfico. Se obtiene un test de normalidad considerando el cuadrado del coeficiente de correlación de los pares $(X_{(j)}, m_j)$,

$$W' = \frac{(X'm)^2}{(m'm)\ \Sigma(X_j-\bar{X})^2} \quad . \tag{2.1}$$

En (2.1), el vector de medias $m$ se reemplaza por una aproximación adecuada. El estadístico W' es, a su vez, una aproximación del estadístico W de Shapiro y Wilk (1965),

$$W = \frac{(X'V_0^{-1}m)^2 \,/\, (m'V_0^{-2}m)}{\sum_{j=1}^{n} (X_j-\bar{X})^2} \quad . \tag{2.2}$$

La distribución nula bajo normalidad de W y W' es intratable y el cálculo de los puntos críticos se hace por simulación (Shapiro y Wilk (1965) y Shapiro y Francia (1972)). Leslie, Stephens y Fotopoulos (1986) han obtenido la distribución asintótica de W. De Wet y Venter (1972) estudian también el comportamiento asintótico de estadísticos de contraste de normalidad de estructura similar a (2.1) y (2.2).

## 2.2 Test basado en la transformación de Box-Cox.

Sea $L(\mu,\lambda,\sigma)$ el logaritmo de la verosimilitud de una muestra obtenida de acuerdo con el modelo (1.2). Se define $L_{max}(\lambda)=\underset{\mu,\sigma}{máx} L(\mu,\lambda,\sigma)$. Sea $\hat{\lambda}$ el estimador de máxima verosimilitud del parámetro $\lambda$. Bajo el modelo (1.2), el conjunto de los valores de $\lambda$ tales que

$$2\left\{ L_{max}(\hat{\lambda}) - L_{max}(\lambda) \right\} \leq \chi^2_{1,\alpha} \tag{2.3}$$

es un intervalo asintótico para el parámetro $\lambda$ de coeficiente aproximado $1-\alpha$. La hipótesis de normalidad se rechaza a un nivel aproximadamente igual a $\alpha$ cuando $\lambda=1$ no pertenece a (2.3). Ver Krishnaiah (1980, p. 291) y Peña (1987, p. 264).

## 2.3 Tests de bondad de ajuste y tests basados en medidas descriptivas

Los tests de bondad de ajuste proporcionan, al particularizarse al caso normal, una familia natural de tests de normalidad. Tal es el caso, por ejemplo, del test $\chi^2$ de Pearson o del test de Kolmogorov-Smirnov.

Los coeficientes muestrales de apuntamiento y simetría proporcionan también un test de normalidad muy conocido.

Los tests de normalidad más relacionados con el método que se presenta en este trabajo son los descritos en 2.1 y 2.2. Para una visión más detallada del problema de contraste de normalidad, ver Krishnaiah (1980, cap. 9) y Wetherill (1986, cap. 8).

## 3. CONSTRUCCION DEL GRAFICO

### 3.1 Resultados previos

Dada una variable aleatoria con función de distribución F, la función cuantílica de X se define por

$$Q(u) = inf \{ x: F(x) \geq u \} \qquad 0 < u < 1. \tag{3.1}$$

Se supone que el lector está familiarizado con las propiedades más importantes de la función Q. (ver Parzen (1979, secs. 2, 3 y 4) y Serfling (1980, sec. 1.1.4.) ). Sea $X_1, ..., X_n$ una muestra aleatoria de F. La función cuantílica muestral se obtiene sustituyendo, en (3.1), F por $F_n$, la función de distribución empírica:

$$Q_n(u) = inf \{ x: F_n(x) \geq u \} \qquad 0 < u < 1. \tag{3.2}$$

Si $X_{(1)} < X_{(2)} < ... < X_{(n)}$ son los estadísticos de orden de la muestra, se tiene $Q_n(u)=X_{(j)}$, $(j-1)/n < u \leq j/n$ $(1 \leq j \leq n)$. En lo que sigue, se supone que F es absolutamente continua con densidad *f*.

Es conocido que las transformaciones en la familia de Box-Cox son crecientes y continuas para cada $\lambda$ fijo. Por consiguiente, bajo el modelo (1.2), se tiene, si $X > 0$, $\mu^{\lambda}\sigma^{-1}(u)=[Q(u)]^{(\lambda)}$, donde $\phi^{-1}(u)$ es la inversa de la función de distribución de N(0,1). Se puede probar que, para todo $\lambda$, $([Q(u)]^{(\lambda)})' = q(u)\ [Q(u)]^{\lambda-1}$, donde $q(u)=Q'(u)$. Entonces, $f[Q(u)] = (1/\sigma)\,[Q(u)]^{\lambda-1}\,\vartheta\,[\phi^{-1}(u)]$, donde $\vartheta$ es la densidad de la distribución normal unitaria. Al tomar logaritmos en esta última relación, se obtiene:

$$\log \frac{f[Q(u)]}{\vartheta[\phi^{-1}(u)]} = -\log\sigma + (\lambda-1)\log Q(u)\,. \tag{3.3}$$

La expresión (3.3) es una ligera modificación de un resultado de Parzen (1979, sec. 12).

### 3.2 Motivación del gráfico

Si en (3.3), se sustituye Q(.) por $Q_n(.)$ y, posteriormente, se pone $u=j/n$, $j=1,2, ..., n$, la relación entre las cantidades

$$U_j = \log\{ f(X_{(j)}/\vartheta[\phi^{-1}(j/n+1)]\} \quad \text{y} \quad V_j = \log X_j, \tag{3.4}$$

debe ser, aproximadamente, lineal, es decir,

$$U_j \tilde{-} -\log\sigma+(\lambda-1)V_j, \tag{3.5}$$

para $j=1, ..., n$.

De (3.5) se deduce el gráfico de los pares $(V_j,U_j)$ puede usarse como herramienta exploratoria para analizar la validez del modelo (1.2). En particular, el caso $\lambda=1$ corresponde a la situación en la que los datos son normales.

### 3.3 Propiedades y consideraciones prácticas

(1) Los posibles patrones de un gráfico $(V_j,U_j)$ construido bajo el modelo (1.2) aparecen recogidos en la figura 1. La figura 1. **a)** corresponde al caso $\lambda<1$ y la figura 1. **b)** al caso $\lambda>1$. Una nube de puntos sin tendencia, como en la figura 1. **c)**, que oscila en torno al punto $-\log\sigma$ corresponde al caso de normalidad ($\lambda=1$).

Fig. 1. a)

Fig. 1. b)

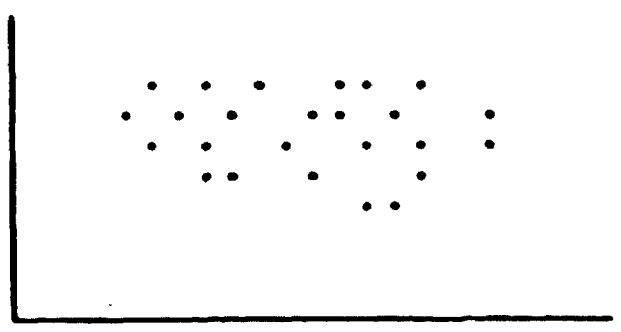


Fig. 1. c)

(2) De acuerdo con (3.4), la definición de los $U_j$ depende de la densidad $f$ desconocida. Por tanto, la construcción efectiva del gráfico requiere la sustitución de $f$ por un estimador $\hat{f}_n(.)$ calculado con la muestra $X_1, ..., X_n$. En notación obvia, $U_j$ se reemplaza, entonces, por $\hat{U}_j$.

(3) El estimador $\hat{f}_n(.)$ puede escogerse de acuerdo con los siguientes criterios:

*(i)* Si los datos que se consideran son razonablemente simétricos en torno a un punto y no contienen valores aberrantes ("outliers") el estimador kernel

$$\hat{f}_n(x) = (nh_n)^{-1} \sum_{i=1}^{n} K[h_n^{-1}(x-X_i] \tag{3.6}$$

es una elección adecuada. En (3.6), K es el núcleo gaussiano 40/2) y $h_n$ es la amplitud de banda. Siguiendo las indicaciones de Silverman (1986, cap. 3) ésta última puede tomarse siguiendo un criterio automático del tipo

$$h_n=0.9\, n^{-1/5}\min(R/1.34,s), \tag{3.7}$$

donde R y $s$ son, respectivamente, el rango intercuantílico y la desviación típica de la muestra, o elegirse por el método de validación cruzada mínimo-cuadrática. Ver Silverman (1986, cap. 3) y Cuevas (1989).

*(ii)* Cuando los datos que se consideran son datos de cola larga es aconsejable tomar el estimador kernel adaptado *(adaptive)*

$$\hat{f}_n(x)=n^{-1} \sum_{i=1}^{n} (\lambda_i h_n)^{-1} K\{(\lambda_i h_n)^{-1}(x-X_i)\}. \tag{3.8}$$

En (3.8), $h_n$ es como en (3.7), K como en (3.6) y las constantes $\lambda_i$ son los factores de amplitud de banda locales definidos por $\lambda_i=\{\tilde{f}_n(X_i)/g\}^{-1/2}$, donde $g$ es la media geométrica de los $\tilde{f}_n(X_i)$ y $\tilde{f}_n(.)$ es un estimador piloto de la densidad que puede tomarse como en (3.6) con $h_n$ dada por (3.7). (Silverman (1986, cap. 5) ).

## 4. EJEMPLOS

En esta sección se ilustran las aplicaciones del gráfico descrito en la sección 3. En los ejemplos 4.1 y 4.2, el cálculo de la amplitud de banda por el método de validación cruzada mínimo-cuadrática se efectúa mediante un programa de ordenador que implementa el algoritmo descrito en Silverman

(1986, cap. 3, sec. 3.5). El algoritmo utiliza la transformada de Fourier rápida de una familia de constantes $\{\xi_k\}$ que depende de cierto proceso de discretización de los datos.

## 4.1 Precipitación de nieve en Buffalo

La tabla 1 recoge los datos de precipitación anual de nieve en Buffalo durante los 63 inviernos de 1910 a 1972. Este conjunto de datos ha sido analizado, con otros propósitos, en Parzen (1979) y Silverman (1986). Los datos han sido proporcionados al autor por Carmichael en comunicación personal. Este último también analiza este conjunto de datos en su tesis doctoral de 1976 que aparece referenciada en Parzen (1979).

| Año | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1910 | 126.4 | 82.4 | 78.1 | 51.1 | 90.9 | 76.2 | 104.5 | 87.4 | 110.5 | 25.0 |
| 1920 | 69.3 | 53.5 | 39.8 | 63.6 | 46.7 | 72.9 | 79.6 | 83.6 | 60.3 | |
| 1930 | 79.0 | 74.4 | 49.6 | 54.7 | 71.8 | 49.1 | 103.9 | 51.6 | 82.4 | 83.6 |
| 1940 | 77.8 | 79.3 | 89.6 | 85.5 | 58.0 | 120.7 | 110.5 | 65.4 | 39.9 | 40.1 |
| 1950 | 88.7 | 71.4 | 83.0 | 55.9 | 89.9 | 84.8 | 105.2 | 113.7 | 124.7 | 114.5 |
| 1960 | 115.6 | 102.4 | 101.4 | 89.8 | 71.5 | 70.9 | 98.3 | 55.5 | 66.1 | 78.4 |
| 1970 | 120.5 | 97.0 | 110.0 | | | | | | | |

Tabla 1. Precipitación anual de nieve en Buffalo 1910-1972

La aplicación del método kernel (3.6) para la construcción de los $\hat{U}_j$ conduce a las figuras 2. a) y 2. b). En 2. a) la amplitud de banda se toma de forma automática, $h_n$=9.322. En 2. b) la amplitud de banda se obtiene por validación cruzada mínimo-cuadrática y es $h_n$=6.889.

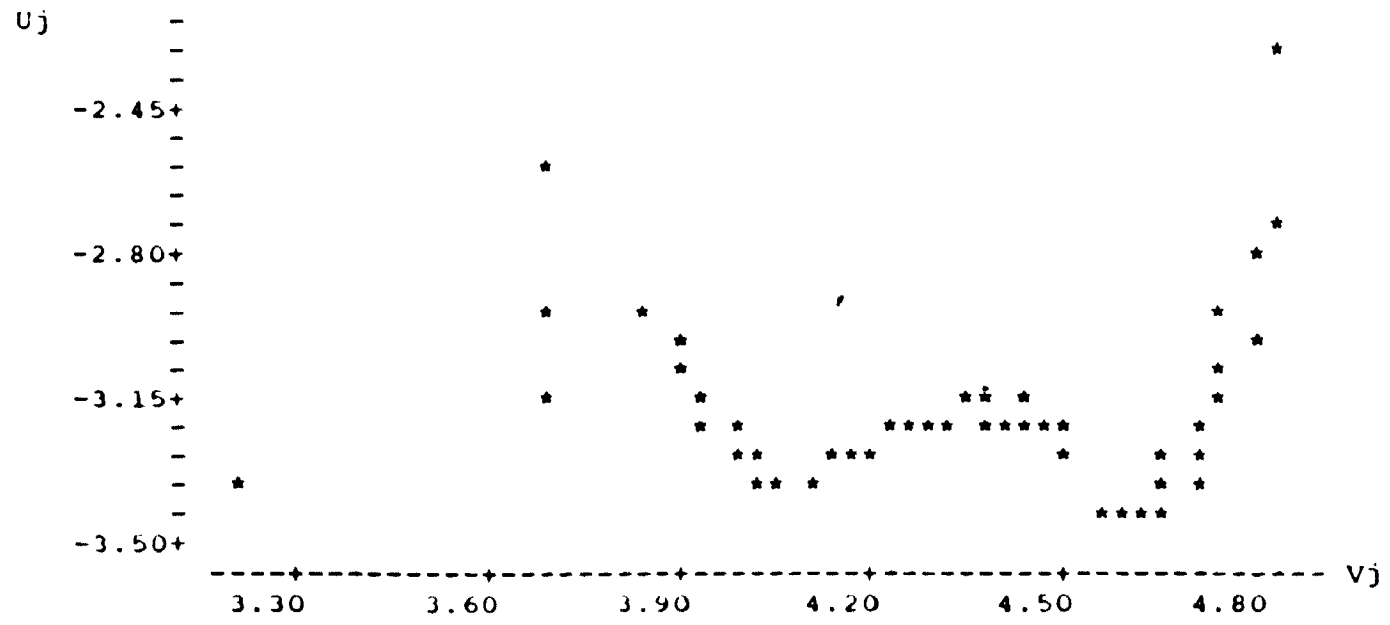


Fig. 2 a)

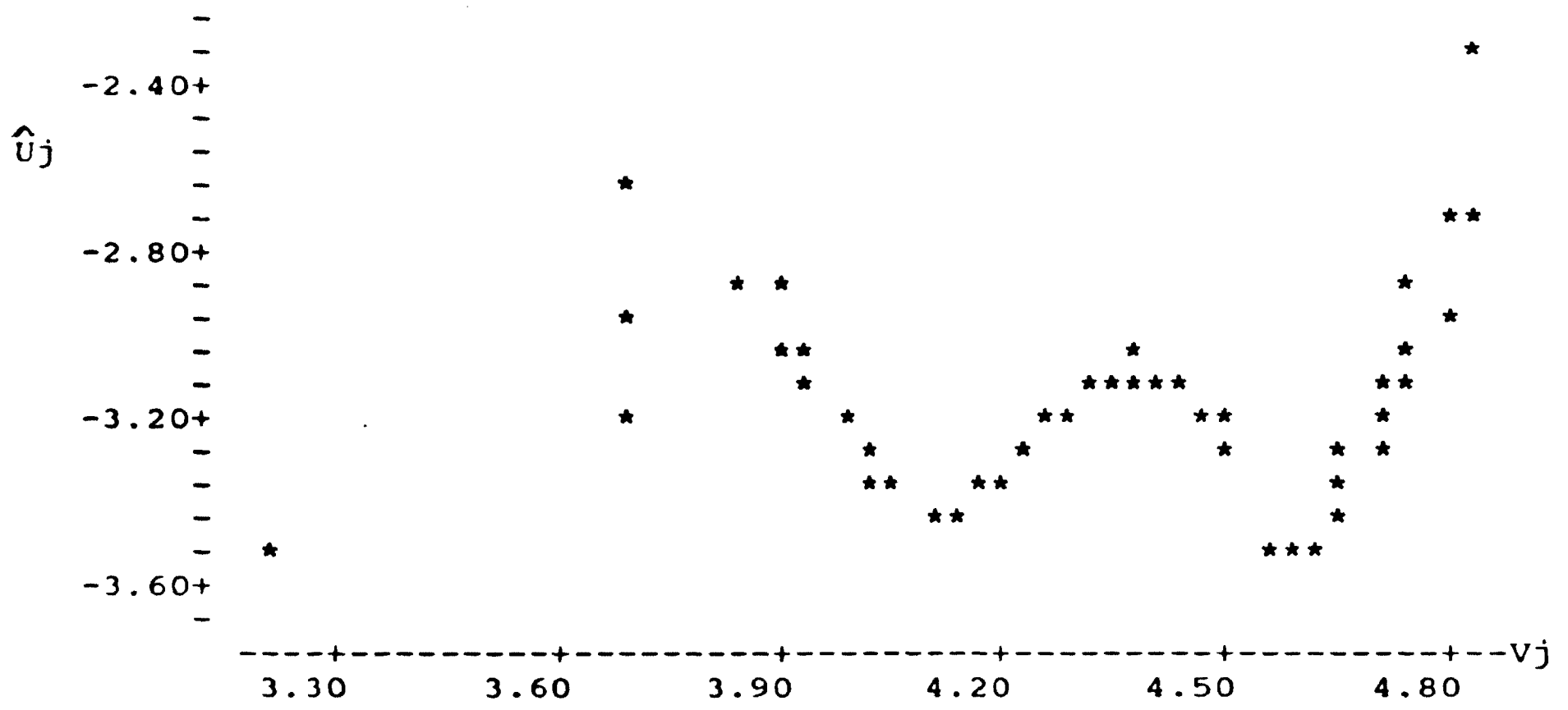


Fig. 2 b)

El núcleo central en las figuras 2. a) y 2. b) es una nube de puntos sin tendencia que oscila en torno al valor -logs=-3.166, donde $s$=23.72 es la desviación típica de los datos de la tabla 1. La conclusión que puede extraerse es que los datos de precipitación de nieve en Buffalo son, aproximadamente, normales. Silverman (1986, págs. 44 y 45) obtiene la misma conclusión a partir de un gráfico de un estimador kernel del tipo (3.6) asociado a una amplitud de banda $h_n$=12. No obstante, cuando la amplitud de banda es $h_n$=6, la estructura del gráfico de $\hat{f}_n(.)$ sugiere la posibilidad de modelar la distribución de los datos de la tabla 1 con una mixtura de tres densidades normales.

Obsérvese que la elección automática y la elección por validación cruzada mínimo-cuadrática de $h_n$ conducen al mismo efecto visual en el gráfico. De acuerdo con (3.4), la expresión de $\hat{U}_j$ es

$$\hat{U}_j = \log \hat{f}_n[X_{(j)}] + (1/2)\log 2\pi + (1/2)\,[\phi^{-1}(j/n+1)].$$

En los puntos de baja densidad, el primer sumando toma valores negativos de alto valor absoluto. Por otra parte, $|\phi^{-1}(t)|^2 \longrightarrow \infty$ cuando $t \longrightarrow 0$ ó $t \longrightarrow 1$. Esto explica el ruido que aparece en las colas de las figuras 2. a) y 2. b).

## 4.2 Concentración de ozono en el centro de Los Angeles

Los datos de este ejemplo son 78 mediciones de la concentración de ozono en la atmósfera del centro de la ciudad de Los Angeles realizadas durante los veranos de 1966 y 1967. Los datos aparecen recogidos en Bhattacharyya y Johnson (1977, pág. 15). La desviación típica es $s$=2.003. La figura 3. a) es el gráfico $(V_j, \hat{U}_j)$, construido con $\hat{f}_n(.)$ como en (3.6) y $h_n$=0.754 elegida de modo automático, y la figura 3. b) el mismo gráfico con $h_n$=0.489 tomada por validación cruzada. En ambos casos, el resultado es una nube de puntos cuya parte central oscila, alrededor de -logs=-0.695. Se sigue la normalidad aproximada de los datos de concentración de ozono.

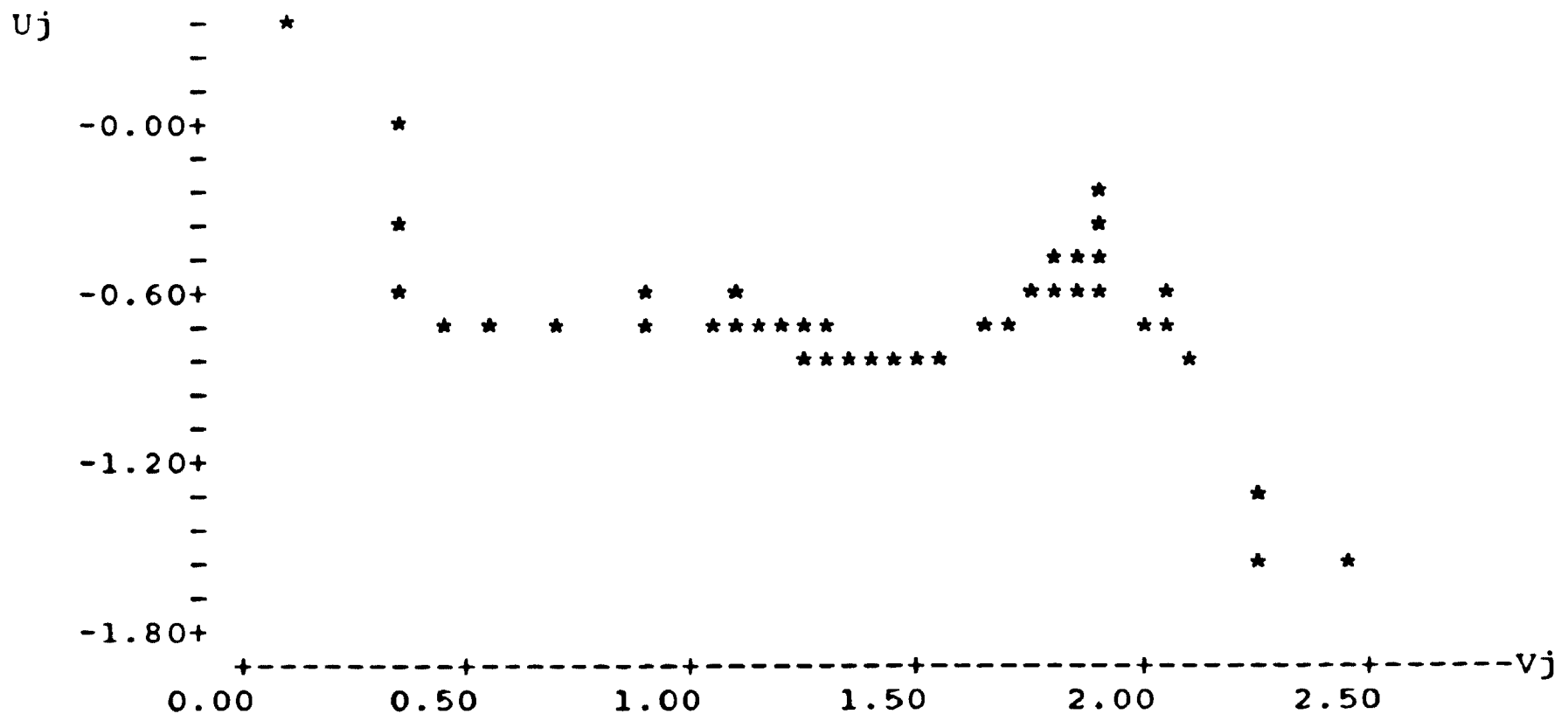


Fig. 3. a)

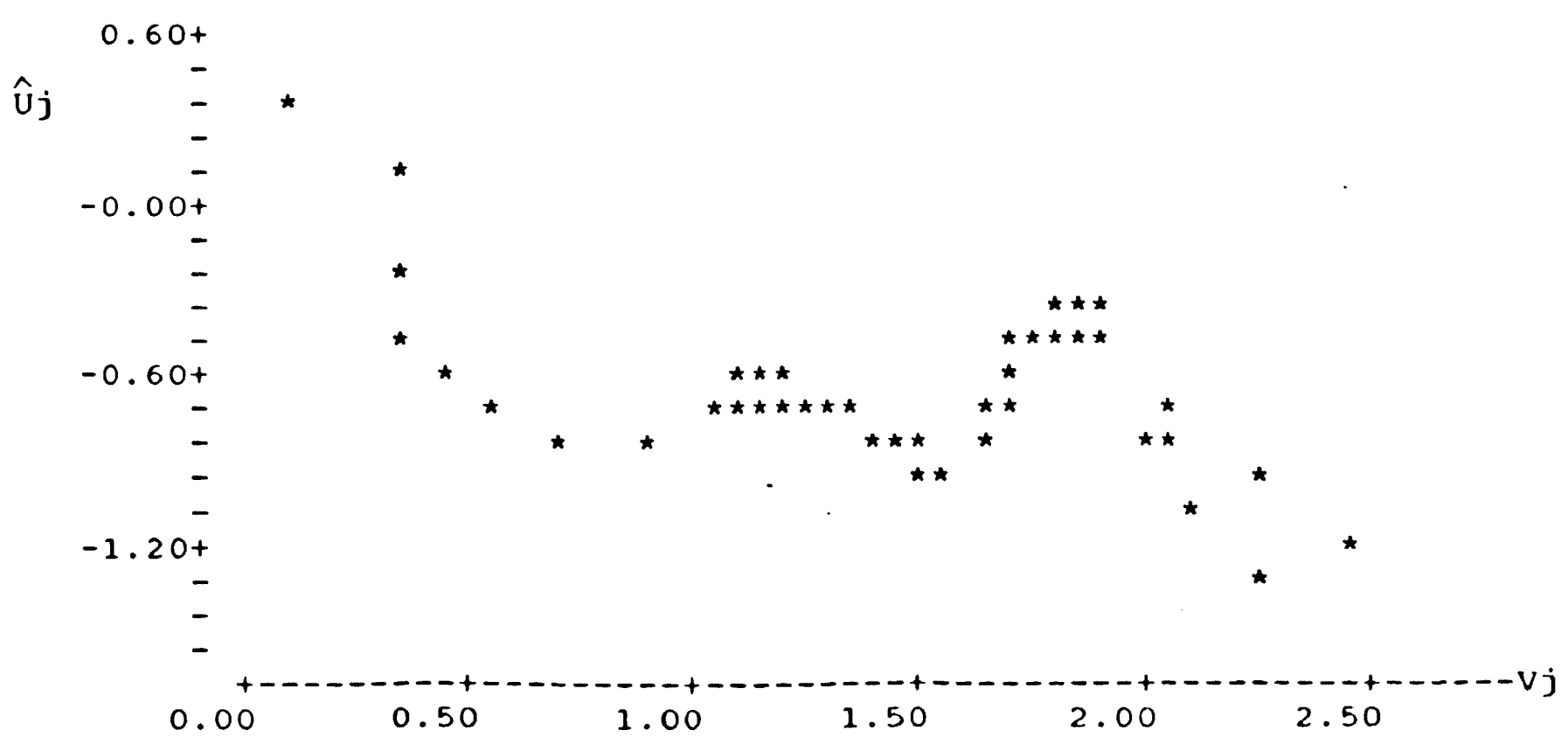


Fig. 3. b)

### 4.3 Volumen de madera

En Bhattacharyya y Johnson (1977, pág. 225), se presenta un conjunto de datos obtenidos al medir el volumen de madera en 49 partes distintas de un bosque seleccionadas al azar. Los datos son de cola larga por lo que se elige el estimador adaptado de (3.8), con $h_n$=3.897, para la construcción de los $\hat{U}_j$. La tendencia lineal del gráfico $(V_j,\hat{U}_j)$, en la figura 4, indica que los datos considerados no son normales. Al considerar la metodología Box-Cox, el estimador de máxima verosimilitud bajo el modelo (1.2) es $\hat{\lambda}_M$=0.25.

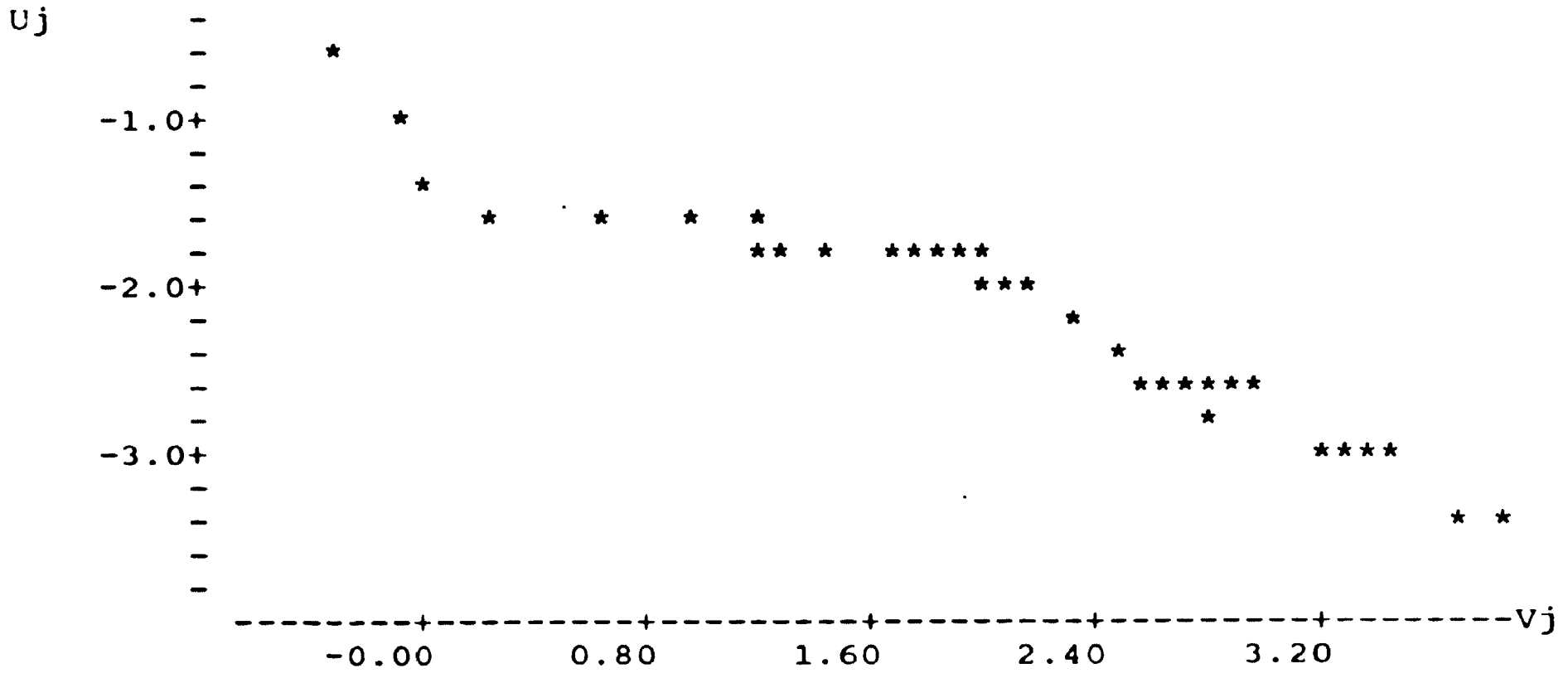


Fig. 4

## 5. DISCUSION Y SUMARIO

En este artículo se presenta un método de análisis gráfico de la hipótesis de normalidad. El método es de carácter semiparamétrico porque combina un fundamento paramétrico basado en la transformación de Box-Cox y utiliza técnicas no paramétricas de estimación de densidades.

La construcción de una medida que resuma el comportamiento de los gráficos introducidos en la sección 3 puede hacerse siguiendo los resultados obtenidos en Velilla (1991). Sea $\hat{\lambda}_n$ el estimador de mínimos cuadrados de $\lambda$ en el modelo de regresión lineal simple (3.5). Bajo el marco probabilístico del modelo (1.2), se puede probar que $\hat{\lambda}_n=\hat{A}_n+B_n$ donde $B_n \longrightarrow c \in \mathbb{R}$, c.s., $c\tilde{}0$, y $\hat{A}_n-1$ es el estimador de la pendiente en la regresión de los

puntos $\log\hat{f}_n(X_j)$ sobre los puntos $\log X_j$. Puede razonarse que la distribución asintótica de

$$n^{1/2}(\hat{A}_n-\lambda), \tag{5.1}$$

es aproximadamente normal de media 0 y varianza $(5/2)/\sigma^2$. Se prueba también que si $D_n=\sum_{j=1}^{n}(V_j-\bar{V})^2$, entonces, $D_n/n \longrightarrow d$, c.s., donde $d\tilde{-}\sigma^2$. La estudentización de (5.1) con el estimador de escala $(5/2)/(D_n/n)$ conduce a $(2D_n/5)^{1/2}(\hat{A}_n-\lambda) \tilde{\sim} N(0,1)$. Por tanto, un test asintótico para el contraste de la hipótesis $H_0:\lambda=1$ (i.e. los datos son normales) es rechazar cuando

$$|(2D_n/5)^{1/2}(\hat{A}_n-1)| > z_{\alpha/2}. \tag{5.2}$$

En (5.2), $z_{\alpha/2}$ es tal que $\Pr[N(0,1)>z_{\alpha/2}]=\alpha/2$. Los $p$-valores asociados al test (5.2) son $p$-valor$=2\Pr[N(0,1)>\Delta_n]$ donde $\Delta_n$ es el izquierdo de (5.2). En el caso de los ejemplos considerados en la sección 3 los $p$-valores respectivos son 0.549, 0.764 y 0.013. En los ejemplos 1 y 2 se ha considerado el caso de $\hat{A}_n$ construido con el estimador kernel automático. El caso del estimador construido por el método de validación cruzada conduce a conclusiones similares.

Las aproximaciones que conducen a (5.2) son correctas, en principio, para valores pequeños de $\lambda$ y, por consiguiente, el $p$-valor obtenido debe interpretarse con precaución. No obstante, a la vista de los ejemplos considerados, el procedimiento (5.2) es de utilidad en la práctica.

Una última observación es que el método descrito es aplicable únicamente cuando los datos son positivos. Ante la presencia de datos negativos, un procedimiento inmediato es tomar como nuevo conjunto de datos $Y_i=X_i+c_0$, donde $c_0$, donde $c_0$ es una constante conocida y tal que los $Y_i$ son todos positivos. Evidentemente, los $(Y_i)$ son normales si, y sólo si, los $(X_i)$ lo son.

## REFERENCIAS


BHATTACHARYYA, G. K. y JOHNSON, R. A. (1977). *Statistical Concepts and Methods*. Nueva York: J. Wiley.

BOX, G. E. P. y COX, D. R. (1964). An analysis of transformations. *J. Roy. Stat. Soc.*, Serie B, **26**, 211-252.

CUEVAS, A. (1989). Una revisión de resultados recientes en estimación de densidades, *Estadística Española*, **120**, 7-62.

DE WET, T. y VENTER, J. H. (1972). Asymptotic distributions of certain test criteria of normality, *South African Statist. J.*, **6**, 135-149.

KRISHNAIAH, P. R. (1980). *Handbook of Statistics, vol.* I. Amsterdam: North Holland.

LESLIE, J. R., STEPHENS, M. A. y FOTOPOULOS, S. (1986). Asymptotic distribution of the Shapiro-Wilk W for testing for normality, *The Annals of Statistics*, **14**, 1497-1506.

PEÑA, D. (1987). *Estadística: Modelos y Métodos, vol.* I. Madrid: Alianza Editorial.

PARZEN, E. (1979). Nonparametric statistical data modeling, *J. A. S. A.*, **74**, 105-131.

SERFLING, R. (1980). *Approximation Theorems for Mathematical Statistics*. Nueva York: J. Wiley.

SHAPIRO, S. S. y FRANCIA, R. S. (1972). An approximate analysis of variance test for normality, J. A. S. A., **67**, 215-216.

SHAPIRO, S. S. y WILK, M. B. (1965). An analysis of variance test for normality (complete samples), *Biometrika*, **52**, 591-611.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Londres: Chapman and Hall.

VELILLA, S. (1991). A quantile approach to the Box-Cox transformation in random samples, Working Paper 91-12, U. Carlos III de Madrid.

WETHERILL, G. B. (1986). *Regression Analysis with Applications*. Londres: Chapman and Hall.

# A GRAPHICAL METHOD FOR THE ANALYSIS OF THE NORMALITY HYPOTHESIS

## SUMMARY

In this paper a graphical methods for analysing the normality hipothesis when the data arise from a random sample is proposed. The properties of the method are illustrated with three real data sets.

*Key Words:* Box-Cox transformation; kernel density estimation; quantile function.

*AMS Classification: 62F05; 62G30.*