Revista Investigación Educativa - N.º 21 - 1993 (P. 125-136)

# ESTUDIO DE LAS CALIFICACIONES ESCOLARES MEDIANTE ANÁLISIS EXPLORATORIO DE DATOS

*por*
*Lluis Salafranca Cosialls, Montserrat Freixa Blanxart*
*y Fco. Javier Ormazábal Unzué*

**RESUMEN**

Tukey 1977 en su libro «Exploratory Data Analysis» expone un conjunto de nuevas técnicas estadísticas resistentes y robustas que intentan descubrir patrones o modelos, anomalías o errores en los datos, para ello se valen de importantes innovaciones principalmente gráficas. Estas técnicas no sólo constituyen un complemento a las clásicas sino también una valiosa alternativa.

En este artículo, se evidencian las ventajas de aplicar algunas de estas técnicas a las calificaciones escolares.

**ABSTRACT**

Tukey in his book «Exploratory Data Analysis» exposed in 1977 an ensemble of new statistical techniques, both robust and resistent, intending to find data models and outlines, errors and anomalies. With this purpose new devises are introduced most of they graphical. These are not only complementary to the classical ones but also they are valuable alternatives.

In this paper are shown the promising possibilities of the use of such technical procedures in school's problems.

## INTRODUCCIÓN

Tukey en su libro «Exploratory Data Analysis» (1977) E.D.A., desarrolla una serie de nuevas técnicas gráficas y analíticas para conseguir un conocimiento previo de los datos a analizar, siempre desde una perspectiva exploratoria.

El análisis exploratorio de datos propugna un cambio de actitud y de enfoque metodológico ante el análisis de datos.

Postula que es necesario *explorar* detenidamente los datos antes de empezar cualquier análisis.

La Estadística Descriptiva clásica se ocupa en recoger, ordenar y representar los datos, normalmente en forma de tablas y agrupando los datos en intervalos para representarlos gráficamente. Calcula principalmente estadísticos basados en la distancia, con datos medidos en escala de intervalo y toma como índice de referencia la media.

El E.D.A. tiene los mismos objetivos, pero pretende ademas *detectar anomalías o errores* en las distribuciones univariantes de los datos de forma que éstos no incidan o invaliden posteriores análisis. También intenta *descubrir en los datos patrones o modelos*. Para ello incorpora nuevas técnicas gráficas y busca estadísticos resistentes y robustos basados principalmente en el orden y tomando como referencia la mediana.

Su sencillez y rapidez de cálculo la hacen sumamente útil en Ciencias Sociales, Humanas y de la Salud para explorar distribuciones univariantes, así como estructuras de relación entre variables.

El E.D.A. tiene cinco características principales:

1) *Sus representaciones gráficas* nos revelan visualmente el comportamiento de los datos y la estructura de conjunto.
2) Se sirve de *la transformación de los datos*, que consiste en encontrar la escala que más simplifique y clarifique el análisis, como por ejemplo con el uso de funciones matemáticas simples como raíz cuadrada, logaritmos, etc.
3) Valora *la resistencia,* propiedad que presentan ciertos estadísticos de ser poco sensibles a la influencia de unos valores muy distantes de la mayoría de los de la distribución.
4) Busca estadísticos *robustos*, propiedad que presentan algunos estadísticos que les hace poco sensibles a desviaciones de los supuestos básicos.
5) Pone mucha atención en el análisis de *residuales,* es decir en las diferencias que hay entre los datos reales y el resultado de un ajuste exploratorio a un modelo previamente determinado o subyacente.

Cuando buscamos relaciones entre las variables, el E.D.A. es especialmente adecuado en las disciplinas anteriormente mencionadas donde los modelos sustantivos son complejos y las variables han sido medidas en todo tipo de escalas, nominal, ordinal, de intervalo y de razón y los datos están sujetos a gran variabilidad.

Así, dado el desconocimiento de los verdaderos modelos y teorías que generalmente subyacen en estos campos, los análisis mediante E.D.A. ayudan a descubrir tendencias, patrones de conducta, conductas diferenciales, formación de actitudes y evaluación del cambio.

Cabe destacar así mismo que las técnicas E.D.A. no sólo constituyen un complemento a las técnicas estadísticas clásicas si no también una valiosa alternativa en caso de incumplimiento de alguna condición de aplicación, puesto que no son tan restrictivas en sus supuestos.

En realidad el investigador necesita usar las técnicas estadísticas exploratorias y confirmatorias. Las técnicas exploratorias ayudan a combrobar las condiciones de aplicación de las pruebas de hipótesis, a detectar errores o valores anómalos, a buscar la mejor transformación cuando es necesaria, etc. En general dan una visión distinta, previa, pero complementaria a la confirmatoria. Todo ello repercute en una mejor calidad del análisis de datos globalmente entendido.

## DIAGRAMA DE TALLO Y HOJAS

*Tukey (1977)* idea una representación gráfica, llamada diagrama de «tallo y hojas» (Steam and Leaf), para variables cuantitativas.

El diagrama de tallo y hojas es un procedimiento semi-gráfico de presentar la información de variables cuantitativas. Es en realidad, una representación visual de la distribución de una variable, que intenta respetar su información cuantitativa al máximo.

Para construir el diagrama de tallo y hojas se hace una tabla con dos columnas separados por una línea y cada dato se desglosa en varias cifras.

Para datos con dos dígitos, como por ejemplo 29, se escribe a la izquierda de la línea los dígitos de las decenas —que forman el tallo— y a la derecha las unidades serán las hojas.

Ejemplo: 29 se representa

| | |
|---|---|
| 2 | 9 |
| tallo | hoja |

Para datos con tres dígitos el tallo puede estar formado por los dígitos de las centenas y decenas que se escribirán a la izquierda, separadas de las unidades.

Ejemplo: 567 se representa

| | |
|---|---|
| 56 | 7 |
| tallo | hoja |

Lo primero que hay que hacer es buscar la unidad que es la hoja. Cada tallo define una clase y se escribe una sola vez. El número de hojas representa la frecuencia de dicha clase.

Ejemplo 54, 57, 69, 67, 43, 59, 62, 36, 74, 67, 46, 75,

| | | | |
|---|---|---|---|
| 4 | 36 | 4 3 | representa 43 |
| 5 | 479 | | |
| 6 | 2.779 | n=12 | |
| 7 | 45 | | |
| tallo | hojas | | |

El número de hojas es igual al número de casos.

En Estadística Descriptiva clásica el histograma es una representación gráfica muy frecuente en la que toda la información se concentra en unos intervalos.

El diagrama de tallo y hojas se parece a un histograma que retiene y ordena todos los datos, aunque estén repetidos, no perdiendo ninguno de ellos y nos da una buena representación de la forma de la distribución.

Las ventajas del Diagrama de tallo y hojas sobre el histograma se pueden concretar en tres niveles.

1) Es más fácil de preparar manualmente.
2) Puede examinarse más detalladamente que el histograma por que las barras de un histograma pueden ocultar puntos dentro de ellas pero esto no ocurre con el diagrama de tallo y hojas ya que este retiene los valores numéricos de los datos.
3) El diagrama de tallo y hojas es un gráfico muy flexible, que permite varios tipos, mientras que el histograma es un gráfico más rígido.

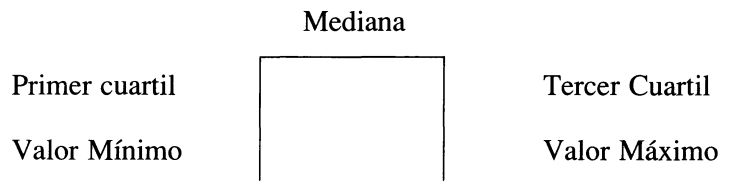
Existen diferentes representaciones del diagrama de «tallo y hojas», a base de diferentes subdivisiones o cambios de unidad. Véase B. Erickson y T. Nosanchuk. (1983).

Hoaglin, Mosteller y Tukey (1983) y Velleman y Hoaglin (1981) coinciden en que el diagrama de tallo y hojas presenta las siguientes características que lo convierten en un gráfico muy útil para «ver» en los datos las siguientes características:

a) La dispersión de los valores.
b) En dónde están concentrados los valores.
c) La simetría de la serie.
d) Si se presentan agujeros (vacíos en donde no se encuentran valores).

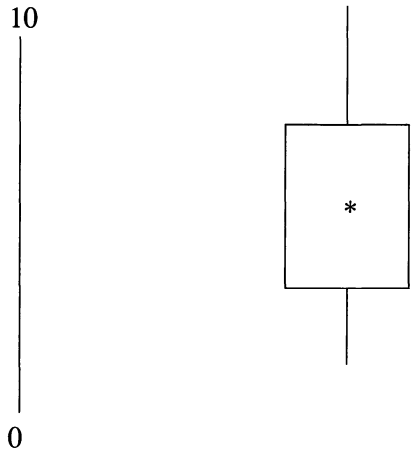
## Diagrama de Caja

Tukey (1977) presenta una nueva representación gráfica que denomina «diagrama de caja» (Box Plot). El esqueleto del más simple de los «diagramas de caja» se construye a partir de la Mediana, los cuartiles y los valores máximo y mínimo.




Dentro de la «caja» se encuentra el 50% central de valores de la distribución.

Se suele disponer de la siguiente manera:




Valor máximo = 10

Tercer Cuartil = 8,45

Mediana = 6

Primer Cuartil = 5

Valor Mínimo = 2,1

Sin embargo normalmente el diagrama de caja se hace más completo ya que detecta los valores alejados (outliers) y estudia hasta que punto se alejan de la normalidad.

El diagrama de caja nos muestra la estructura de la serie de datos. A partir de esta representación gráfica se puede ver claramente, la dispersión, la simetría, el aspecto y el alcance de las colas y los valores alejados de una distribución, así como la localización de un valor determinado.

El diagrama de caja resulta especialmente útil para comparar varias distribuciones a la vez. Dibujando en paralelo el diagrama de caja de cada distribución, podemos disponer de una impresión visual rápida de las similitudes y diferencias entre las distribuciones, así como de las características de las mismas. El diagrama de caja sugiere con frecuencia la mejor transformación de los datos cuando ésta sea necesaria.

## Ejemplo aplicado a las calificaciones escolares

Es ya práctica habitual en muchos colegios que en el boletín de calificaciones después de cada evaluación, además de las notas y actitud en cada asignatura, alguna información estadística, dirigida a los padres, que pretende comparar la calificación del alumno con la media del grupo.

Normalmente la información es como sigue

Estudio compartivo

| | media | def. | insuf. | suf. | bueno | not. | sob. |
|---|---|---|---|---|---|---|---|
| cal. alumno | 3,5 | ******** | | | | | |
| med. grupo | 5,46 | ********** | | | | | |

En esta información se observa la diferencia entre la nota media del grupo y la calificación del alumno.

No obstante dado que no se acostumbra a dar ninguna medida de dispersión es imposible ver la posición del alumno dentro del grupo.

Además en este estudio comparativo se utiliza normalmente la media como índice descriptivo. La media es un estadístico no resistente ya que puede estar afectada por las notas altas o bajas de dos o tres sujetos no representativos del grupo.

Estudiemos dos clases con las puntuaciones y los estadísticos que se detallan en la tabla 1. Las dos clases tienen prácticamente la misma media (media clase 1 = 5,47 media clase 2 = 5,46). Entonces dos alumnos, uno de la clase 1 y otro de la clase 2 con una puntuación de 3,5, por ejemplo, tendrían una gráfica de estudio comparativo igual en las dos clases. No obstante si analizamos más detenidamente las distribuciones veremos que tienen dispersión distinta y que una misma nota sitúa a los alumnos en muy distinta posición dentro del grupo.

Clase A

| | | | | |
|---|---|---|---|---|
| 4,1 | 3,5 | 4,2 | 5,5 | 5,5 |
| 5,0 | 6,0 | 7,0 | 6,0 | 6,0 |
| 5,5 | 6,2 | 4,7 | 4,8 | 5,8 |
| 5,4 | 6,1 | 6,2 | 5,5 | 4,8 |
| 5,0 | 5,0 | 3,9 | 5,6 | 4,9 |
| 4,5 | 4,5 | 5,0 | 6,7 | 6,0 |
| 5,9 | 7,0 | 5,8 | 6,7 | 6,0 |
| 7,1 | 5,0 | 4,0 | 6,7 | 5,5 |
| 4,2 | 4,8 | 4,5 | 6,9 | 6,8 |

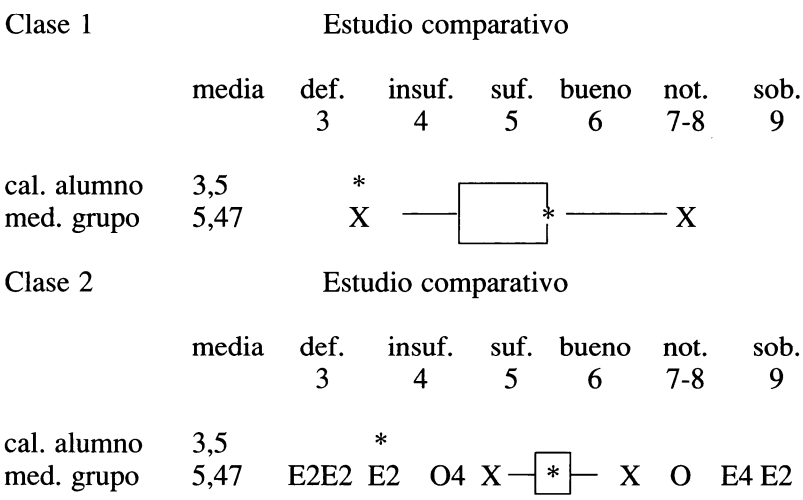
Clase B

| | | | | |
|---|---|---|---|---|
| 2,0 | 3,4 | 2,7 | 1,8 | 2,7 |
| 3,2 | 1,5 | 2,4 | 5,0 | 3,6 |
| 3,2 | 3,7 | 5,5 | 5,0 | 5,5 |
| 5,0 | 6,0 | 5,9 | 5,8 | 5,8 |
| 5,9 | 6,0 | 6,8 | 5,0 | 5,0 |
| 5,0 | 5,0 | 5,0 | 6,3 | 5,5 |
| 5,0 | 6,2 | 5,8 | 6,8 | 5,9 |
| 6,8 | 5,9 | 5,9 | 8,9 | 7,9 |
| 9,9 | 8,7 | 9,0 | 8,5 | 9,9 |

En efecto, el alumno que tiene puntuación 3,5 puede ser el último de la clase 1 o bien ocupar un lugar próximo al primer cuartil en la clase 2, lo que se evidencia en la gráfica del estudio comparativo. Lo mismo pasaría si comparamos dos notas altas en las dos clases. Un alumno con nota 7,1 es el mejor del grupo en la clase 1 o puede estar detrás de un número relativamente grande de alumnos mejores que él en la clase 2.

Todo ello evidencia que el estudio estadístico que se viene realizando de las calificaciones escolares, que pretende ser comparativo, no sólo es insuficiente sino que también puede ser engañoso en algunos casos.

Si representamos las calificaciones con diagramas de caja se observa enseguida si el alumno está dentro del 50% central de sujetos o es un valor alejado en alguno de los extremos. Es decir, además de la información de la nota del alumno y la mediana del grupo, se da la posición del alumno dentro del grupo.



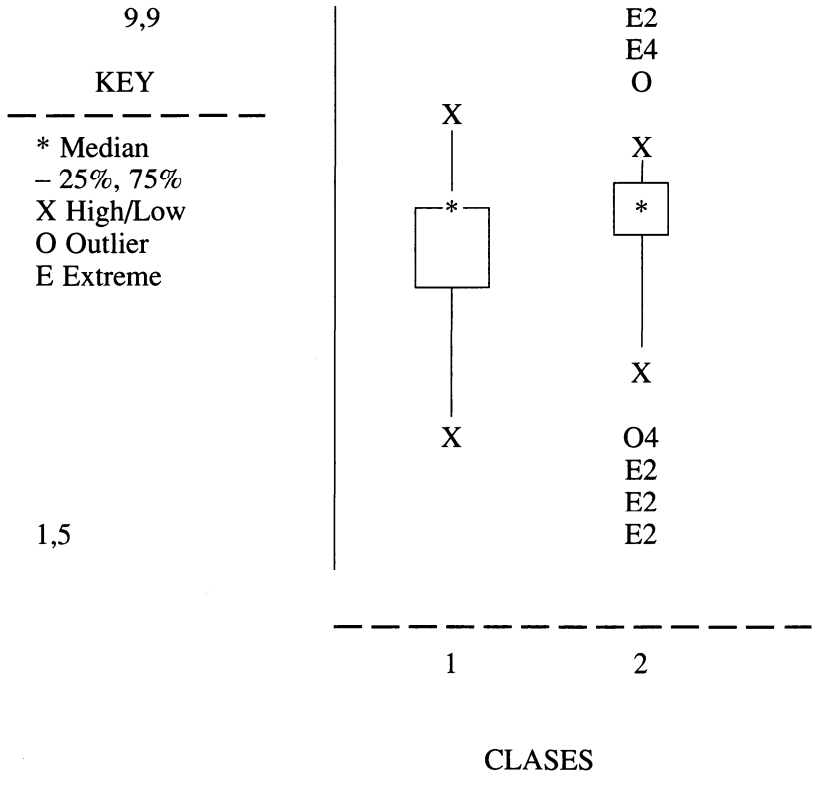

## Comparación de distribuciones

Las representaciones gráficas del análisis exploratorio de datos pueden utilizarse para comparar distribuciones, en este caso las de las dos clases.

Diagrama de tallo y hojas

| | Clase 1 | | Clase 2 |
|---|---|---|---|
| 3 | 59 | 1 | 58 |
| 4 | 0122 | 2 | 0477 |
| 4 | 55578889 | 3 | 22467 |
| 5 | 000004 | 4 | |
| 5 | 555556889 | 5 | 00000000055588899999 |
| 6 | 00000122 | 6 | 0023888 |
| 6 | 77789 | 7 | 9 |
| 7 | 001 | 8 | 579 |
| | | 9 | 099 |
| | N = 45 | | N = 45 |

El diagrama de tallo y hojas no esconde información. Se observa, por ejemplo, que no hay ningún cuatro en la clase 2, etc.



## Ajuste de Medianas

El Análisis Exploratorio de Datos ofrece una serie de técnicas resistentes y robustas para examinar relaciones entre dos o más variables (independientes desde una perspectiva experimental) cualitativas (no necesariamente) y una variable respuesta (dependiente desde una perspectiva experimental) cuantitativa. Esta estructura de datos conocida como tablas de dos factores o diseño factorial se presenta mucho para estudiar como cada uno de los factores varía regularmente y separademente del otro y para observar los valores que va tomando la variable respuesta según las diferentes combinaciones de los niveles y de los factores.

Estas tablas son analizadas tradicionalmente en Estadística clásica con el análisis de la varianza de dos factores.

El ajuste simple de medianas (Median Polish) descompone los efectos de la variable dependiente:

$$Y = \text{efecto común} + \text{efecto fila} + \text{efecto columna} + \text{residual}$$

Un ajuste Y para tablas de dos factores describe los datos a través de la ecuación

$$Y = X \beta + E$$

Aunque en principio el ajuste de medianas usa un modelo aditivo similar al del Análisis de la Varianza (ANOVA), ajustando éste a partir de las medianas, a través de un proceso iterativo, pone mucho énfasis en mirar y analizar los residuos.

Resumiendo, podemos decir que la técnica que introducimos ofrece, (Freixa, Salafranca, Guardia, Ferrer, Turbany, 1992) para explorar tablas de dos factores las siguientes ventajas:

— No es preciso asumir los rígidos supuestos de un modelo lineal.
— Puede analizarse con todo tipo de datos (puntuaciones directas, porcentajes, proporciones, medias, medianas, etc...).
— Puede realizarse el análisis con datos incompletos (casillas vacías).
— Es resistente.
— Explora la estructura aditiva entre las variables y mediante otras técnicas Eda se busca la transformación más adecuada para conseguirla.
— Detecta patrones de comportamiento de los datos analizando los residuales. Mediante la descomposición de los datos intenta detectar sus patrones de comportamiento, complementando la búsqueda de estos patrones con el estudio de residuales.
— Es en general más flexible y por tanto tiene gran diversidad y riqueza de análisis y aplicaciones.

Debe advertirse, sin embargo, que el análisis de Medianas puede dar resultados un poco diferentes si el proceso de análisis se empieza por filas o por columnas.

También puede dar resultados un poco distintos según el número de iteraciones que se hagan,

pero todo ello no efectaría a las conclusiones globales extraídas.

Aunque el análisis de medianas puede usarse como técnica alternativa al ANOVA, puede plantearse como estrategia exploratoria, aportando una visión distinta y previa al análisis confirmatorio.

Se podría decir que las informaciones provinientes del análisis exploratorio resistente ayudan a señalar indicios de estructura aditiva o de aproximación a patrones o modelos, siendo ello especialmente interesante en investigaciones en Ciencias Humanas, Sociales y de la Salud.

Ejemplo:

| | Clase 1 | Clase 2 | Clase 3 | Clase 4 |
|---|---|---|---|---|
| Asignatura A | 5,8 | 7,2 | 7,3 | 9,1 |
| Asignatura B | 5,4 | 7 | 7,8 | 8,8 |
| Asignatura C | 6,2 | 8,8 | 7,4 | 8,1 |
| Asignatura D | 6,7 | 9,2 | 9,3 | 7,2 |

El ajuste de medianas requiere a menudo varias iteraciones. En la siguiente tabla se presentan los residuos después de dos iteraciones realizadas mediante el paquete estadístico Statgrafics.

| | Clase 1 | Clase 2 | Clase 3 | Clase 4 | Efecto Asignatura |
|---|---|---|---|---|---|
| Asignatura A | 0,1875 | 0,3625 | –0,2250 | 0,8125 | –0,275 |
| Asignatura B | –0,2262 | –0,6125 | 0,2250 | 0,4625 | –0,225 |
| Asignatura C | 0,3125 | 0,9662 | –0,4000 | –0,4625 | –0,000 |
| Asignatura D | –0,1875 | 0,3625 | 0,5000 | –2,3625 | 1 |
| Efecto Clase | –1,9370 | 0,0125 | –0,025 | 0,7375 | 7,82 |

Reproducimos el valor original

$$5{,}8 = 7{,}82 + (-0{,}275) + (-1{,}937) + 0{,}1875$$

En la clase 4 y en la asignatura D hay un residual alto lo cual evidencia signos de interacción. La nota promedio de la Asignatura D en la clase 4 presenta un valor menor que la mediana común, cuando lo esperado por efectos principales es que

ésta se incrementara por el hecho de ser la asignatura D y por el hecho de ser la clase 4.

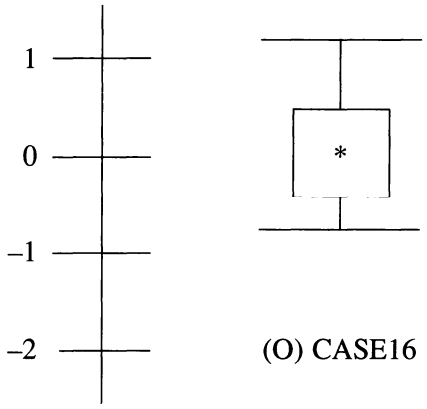



Diagrama de Caja de los residuales

## CONCLUSIÓN

Con las técnicas y gráficas E.D.A. obtenemos más información de una manera rápida y sencilla. En este sencillo ejemplo los padres y tutores podrán ver la situación relativa del alumno dentro del grupo de una forma fiable. Esta información tiene gran importancia no sólo desde el punto de vista academico sino también personal.

Al Director del Centro y a los tutores estos fáciles cálculos le servirán para comparar los grupos de cada curso, que en principio se suponen homogéneos y sin grandes diferencias. Informaciones tales como que hay un tutor que no pone cuatros o notas altas, o que los alumnos menos capacitados o los mejores están en un mismo grupo, etc. se desprenden con sólo mirar los gráficos E.D.A.

Finalmente y desde un punto de vista más crítico y amplio, este ejemplo nos lleva a reflexionar sobre la utilidad y representatividad de los estadísticos clásicos que en algunos casos no son los más adecuados. Sirva el estudio, para que el usuario estadístico tome conciencia de que hay otras muchas pruebas estadísticas adecuadas a las características de cada estudio, complementarias o alternativas a las clásicas.

Tukey (1977) afirma en su libro que el E.D.A. es un *trabajo de detective númerico* para evitar confundir, mentir o cometer errores al utilizar la Estadística.

Las técnicas presentadas en este artículo son algunas de las que el E.D.A. ofrece siendo todas ellas muy útiles para analizar datos en varios contextos.

## BIBLIOGRAFÍA

BATISTA, J. M. & VALLS, M. (1985a): Nuevas técnicas de análisis estadístico de datos: Tabulación y síntesis numérica Análisis Exploratorio de Datos). *Qüestió, 9,* 2, 105-119.

BATISTA, J. M. & VALLS, M. (1985b): Técnicas gráficas en Análisis Exploratorio de Datos. *Qüestió, 9, 3,* 163-176.

ERICKSON, B. & NOSANCHUK, T. (1979): *Understanding Data.* Open Univ. Press: Milton Keynes.

FREIXA, M.; SALAFRANCA, Ll.; GUARDIA, J.; FERRER, R. & TURBANY, J. (1992): *Análisis Exploratorio de Datos: Nuevas Técnicas Estadísticas.* Barcelona: PPU.

HOAGLIN, D.; MOSTELLER, F. y TUKEY, J. W. (1983) (Eds.):. *Understanding Robust and Exploratory Data Analysis.* New York: John Wiley & Sons.

HORBER, E. (1991): *Manual del paquete estadístico EDA.* Faculté des Sciences Politiques. Ginebra.

HARTWIG, F. y DEARING, B. R. (1979): *Exploratory Data Analysis.* London: Sage.

TUKEY, J. W. (1977): *Exploratory Data Analysis.* Reading, Massachussets: Addison-Wesley.

VELLEMAN, P. F. y HOAGLIN, D. C. (1981): *Applications, Basics and Computing of Exploratory Data Analysis.* Boston: Duxbury.