Anuario de Psicología
1992, nº 55, 23-36

Aportaciones del análisis exploratorio de datos al estudio de la resistencia

Ramon Ferrer
Montserrat Freixa
Joan Guàrdia
Universidad de Barcelona
Eugene Horber
Universidad de Ginebra

El artículo presenta, a nivel introductorio, las técnicas agrupadas por Tukey y colaboradores bajo la denominación Análisis Exploratorio de Datos, renovadoras tanto en su concepción como procedimiento de la metodología de análisis de datos en nuestros días. Algunas de estas técnicas son consideradas ya como «clásicas» en recientes manuales de estadística. Sus características principales, ejemplificadas en diversos campos de aplicación a lo largo del texto, pueden resumirse en: a) potenciación de técnicas gráficas, b) maximización de la resistencia y robustez de los indicadores, c) minimización de los supuestos previos a la aplicación de técnicas de contraste y ajuste, d) facilitar el uso de transformaciones sobre los datos originales y e) atención especial a los residuales generados en el proceso de análisis, todo ello derivado de su focalización en índices descriptivos de posición, para simplificar el «trabajo de detective numérico» propugnado en sus trabajos.

Palabras clave: *Análisis Exploratorio de Datos (EDA), resistencia, robustez, promedio de cuartiles (Q), trimedia (TRIM), centrimedia (MID), mediana de las desviaciones absolutas (MAD), gráfico de tronco y hojas, gráfico de caja, ajuste de medianas, línea resistente.*

This paper introduces to «Exploratory Data Analysis» techniques, so-called by Tukey and cols. that renews today's data analysis methodology from both concept and proceedings. Some of them have already been considered «classical» on recent statistic handbooks. Their principal features, exemplified on different fields in the text, could be resumed in a) graphical techniques outstanding, b) index resistance and robustness maximization, c) previous assumptions to contrast and adjusting techniques minimization, d) rough data transformation strength and e) special focu-

Dirección de los autores: Ramon Ferrer, Montserrat Freixa y Joan Guàrdia. Departamento de Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Universidad de Barcelona. Adolf Florensa s/n. 08028 Barcelona. Eugene Horber. Département de Science Politique. Faculté des Sciences Economiques et Sociales. Université de Genève. 102, Boulevard Carl-Vogt. 1211 Genève 4, Suisse.

sing on residual outcoming from analysis process, all derived since its starting on statistical position descriptive index, to furnish the «numerical detective's job» proposed on their papers.

Key words: *Exploratory Data Analysis (EDA), Resistance, Robustness, Mid Spread Mean (Q), Trimmed Mean (TRI), Midmean (MID), Median of Absolute Deviations (MAD), Steam-and-Leaf Display, Box-Plot, Median Polish, Resistent Line.*

Tukey (1977) en su libro *Exploratory Data Analysis* (EDA), desarrolla una serie de nuevas técnicas gráficas y analíticas para conseguir *un conocimiento previo de los datos a analizar, siempre desde una perspectiva exploratoria*, y propugna *un cambio de actitud y de enfoque metodológico ante el análisis de datos*. Intenta descubrir en los datos patrones o modelos, incorporando nuevas técnicas gráficas y presenta estadísticos resistentes y robustos basados principalmente en el orden y centrados en la mediana.

De acuerdo con lo propuesto por Hoaglin, Mosteller y Tukey (1983) y por Velleman y Hoaglin (1981) se reconoce la existencia de cinco características principales del EDA:

1. Sus *representaciones gráficas* nos revelan, en una primera fase, el comportamiento de los datos y la estructura del conjunto.
2. Dedica mucha atención al análisis de *residuales*.
3. Utiliza la *transformación de los datos* para conseguir ajustar los valores originales a la escala que más simplifique y clarifique el análisis como, por ejemplo, mediante el uso de funciones matemáticas simples (raíz cuadrada, logaritmos, etc.).
4. Valora *la resistencia*, propiedad que presentan algunos estadísticos que les hace poco sensibles a la influencia de uno o varios valores marcadamente distantes de la mayoría de los datos de la distribución.
5. Busca estadísticos *robustos*, propiedad que presentan algunos estadísticos que les hace poco sensibles ante desviaciones de los supuestos básicos.

En consecuencia, las técnicas EDA no sólo constituyen un complemento a las técnicas estadísticas clásicas sino 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 tanto las técnicas estadísticas exploratorias como las confirmatorias. Las técnicas exploratorias ayudan a comprobar las condiciones de aplicación de las pruebas de hipótesis, detectar errores o valores anómalos, establecer 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.

El presente artículo pretende ofrecer una perspectiva o presentación de las posibilidades de algunas de dichas técnicas, resaltando las propiedades vinculadas a su *resistencia*, sin que se pretenda efectuar una exhaustiva demostración de sus cualidades y/o limitaciones. Para ello pueden consultarse algunos de los manuales sobre el tema (Marsh, 1988; Freixa y cols., 1992).

Índices resistentes y análisis gráfico

El análisis exploratorio de datos aporta como novedad respecto a la estadística clásica que la descripción se efectúa en base a estadísticos resistentes.

Además de la mediana, considerada como índice de localización, hay otros estadísticos resistentes, basados asimismo en indicadores de posición, entre los que a continuación citamos:

PROMEDIO DE CUARTILES (Q)

Se calcula mediante:

$$Q = \frac{C_{25} + C_{75}}{2}$$

TRIMEDIA

calculada mediante:

$$TRI = \frac{Md + \bar{Q}}{2} = \frac{C_{25} + (2 \cdot Md) + C_{75}}{4}$$

CENTRIMEDIA O MEDIA INTERCUARTÍLICA

Se calcula promediando todos los valores *entre* el primer y tercer cuartiles.

$$MID = \frac{X_{iC25+1} + \ldots + X_{iC75-1}}{n_i}$$

Todas estas expresiones pueden considerarse casos particulares de medias recortadas. Entendiéndose como *media recortada* aquella media obtenida después de haber eliminado a ambos lados de la distribución una proporción α de valores, como por ejemplo el «5 % Trim» que nos ofrece el comando «Examine» del SPSS/PC+ (V.4.0). (Véase Tabla 2.)

MEDIANA DE LAS DESVIACIONES ABSOLUTAS (MAD)

Se calcula mediante:

$$MAD = Md \mid X_i - Md \mid$$

es decir, obteniendo la mediana (Md) de las diferencias, en valor absoluto, respecto de la mediana general de la serie. El proceso de cálculo de estos índices resistentes puede ejemplificarse mediante los datos que se exhiben en la siguiente tabla. Supongamos una situación simple en la que se ha registrado el tiempo de reacción de varios sujetos, sometidos a distintas intensidades de estimulación acústica:

TABLA 1. TIEMPOS DE REACCIÓN EN MSEG. ANTE DIVERSAS INTENSIDADES DE ESTIMULACIÓN ACÚSTICA

20 dB		30 dB	40 dB		65 dB	
161	172	158	146	164	151	160
165	180	164	155	166	152	168
168		170	160	174	153	
170		175	162	182	157	

El análisis descriptivo de los 24 sujetos que constituyen la muestra general ofrece los resultados que exponemos a continuación (Comando Examine del SPSS/PC+ V.4.0 y otros cálculos).

TABLA 2. RESULTADOS DE LOS ÍNDICES ESTADÍSTICOS RESISTENTES

Mean	163.8750	Std Err	1.8760	Min	146.0000	Skewness	.0935
Median	164.0000	Variance	84.4620	Max	182.0000	S E Skew	.4723
5 % Trim	163.8333	Std Dev	9.1903	Range	36.0000	Kurtosis	−.4057
				IQR	12.7500	S E Kurt	.9178
MAD	6.000	TRI	163.870	Q	163.7500	MID	163.830

Algunos de estos estadísticos tienen la ventaja de ser resistentes. Como ya se ha comentado, un estadístico es resistente si su resultado prácticamente no varía cuando reemplazamos una pequeña parte de los datos por otros diferentes. Los métodos resistentes dan gran importancia a la parte central de los datos y poca a los posibles valores alejados. Por ejemplo la media aritmética no es resistente, puesto que el cambio de un solo valor en la serie de datos hace variar el estadístico, en cambio la mediana es un valor resistente.

Para poder evaluar qué fracción de observaciones de la muestra puede cambiarse sin que el estadístico varíe, el EDA propone el punto de colapso o de ruptura. El punto de colapso (*breakdown point*) limita el número de valores que pueden ser alterados sin que el estadístico varíe.

A título ilustrativo, el punto de colapso de la mediana se puede reflejar en las siguientes expresiones:

$(1/2) - (1/n)$ si n es par
$(1/2) - (1/2n)$ si n es impar

Veamos la utilización de este concepto estadístico con la siguiente serie simulada de datos:

3 5 6 9 10 10 11 11 13 16

El punto de colapso correspondiente a la mediana de esta serie de $n=10$ se puede obtener del siguiente modo:

$$(1/2) - (1/n) = (1/2) \cdot (1/10) = 0.5 - 0.1 = 0.4$$

lo cual conlleva que puede alterarse el 40 % de valores en cada una de las colas de la distribución, sin que la mediana varíe.

Por lo que se refiere a las novedades gráficas destacan principalmente las aportaciones del diagrama de tronco y hojas y del diagrama de caja que, de hecho, son ya conocidas en el ámbito estadístico e incorporadas en algunos programas informáticos de tratamiento de datos.

El diagrama de tronco y hojas es un procedimiento semi-gráfico de representar la distribución de una variable cuantitativa. Para elaborar el diagrama de tronco y hojas se construye una tabla con dos columnas separadas por una línea y cada dato se desglosa en sus unidades.

De este modo el primer paso necesario es establecer la unidad que es el tronco. Cada tronco define una clase y se escribe una sola vez. El número de hojas representa la frecuencia de dicha clase. El número de hojas es igual al número de casos. El diagrama de tronco y hojas de nuestros datos acerca de los 24 sujetos adoptaría la forma siguiente:

Frequency	*Stem*	&	*Leaf*
1.00	14	|	6
6.00	15	|	123578
10.00	16	|	0012445688
5.00	17	|	00245
2.00	18	|	02
Stem width:	10		
Each leaf:	1 case(s)		

Gráfica 1. Diagrama de Tronco y Hojas.

El diagrama de tronco y hojas es un gráfico parecido al histograma pero que muestra los valores numéricos de los datos y permite ver rápidamente la dispersión de la distribución, la simetría de la serie, dónde están los valores concentrados y si se presentan vacíos en la distribución. Es por otro lado mucho más flexible que el histograma y muy útil para comparar distribuciones.

Otro gráfico importante es el denominado diagrama de caja que nos muestra la estructura de la serie de datos en la cual se puede evaluar la dispersión, simetría, y el aspecto y alcance de las colas y los valores alejados, así como la localización de un valor determinado.

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 (Freixa y cols., 1992; Tufte, 1987, 1991; Buja y Tukey, 1991).

Normalmente el diagrama de caja se hace más completo ya que estudia especialmente los valores alejados del 50 % central de la distribución, y hasta

qué punto se alejan de éste. Una especial utilización del diagrama de caja es la de efectuar comparaciones entre distintas distribuciones. Si sometemos los datos anteriormente citados a un análisis de varianza (ANOVA) clásico obtendremos evidencia empírica tendente a la hipótesis nula ($F_{(3,21)}=2.0877$; $p=0.1324$). Sin embargo, si representamos gráficamente, y de forma paralela, los diagramas de caja de los cuatro grupos, obtendremos una figura como la que sigue:

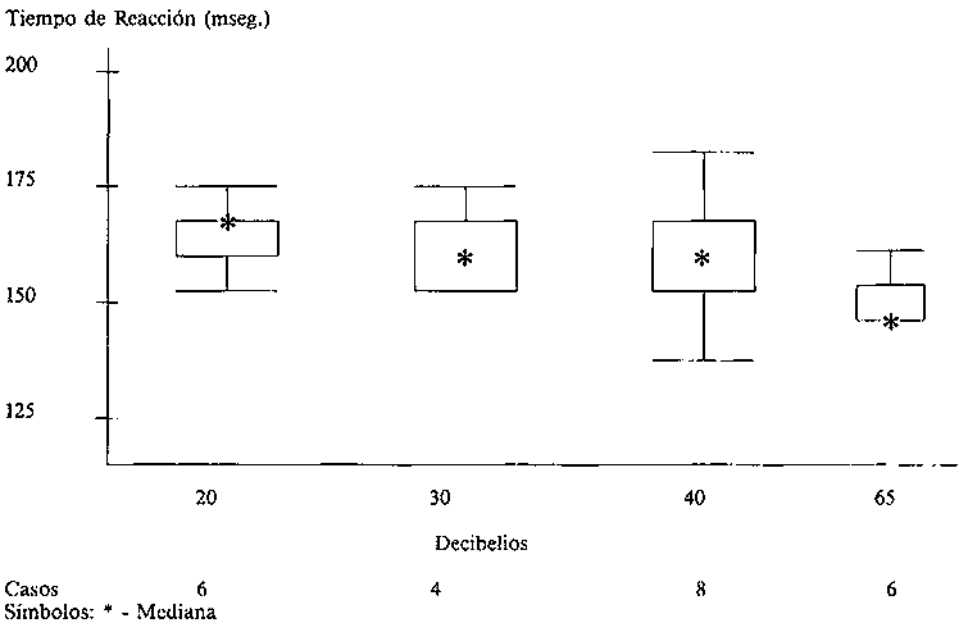

Figura 1. Diagramas de caja paralelos de los cuatro grupos de la Tabla 1.

No es difícil comprobar que la no significación estadística del análisis de la varianza, no se opone, sino que se complementa con el diagnóstico del comportamiento de los cuatro grupos a través de los diagramas de caja. Parece lógico pensar en una cierta tendencia a la disminución del tiempo de reacción al aumentar la intensidad del estímulo acústico; lo cual, por otra parte, es conocido dentro del campo de la reacciometría.

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 res-

puesta (dependiente, desde una perspectiva experimental) cuantitativa. Esta estructura de datos conocida como tablas de dos factores o diseño factorial se utiliza frecuentemente para estudiar cómo cada uno de los factores varía regular y separadamente 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 contrastadas tradicionalmente en estadística clásica mediante el análisis de la varianza de dos factores. El ajuste simple de medianas (*Median Polish*) descompone los efectos de la variable dependiente del siguiente modo:

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

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

$$Y = X + \alpha + \beta + \varepsilon$$

Aunque en principio el ajuste de medianas usa un modelo aditivo similar al del análisis de la varianza, ajustando éste a partir de las medianas a través de un proceso iterativo, pone el énfasis en el análisis de los residuales.

El ajuste de medianas requiere, a menudo, varias iteraciones hasta alcanzar una mediana de los residuales cercana a cero. Para ello se sigue el procedimiento de cálculo que a continuación se esquematiza:

1. Se halla la mediana de cada fila.
2. Se halla la mediana de las medianas de las filas.
3. A cada valor se le resta la mediana correspondiente a su fila.
4. Se hallan las medianas de los residuos de cada columna.
5. A cada residuo se le resta la mediana de los residuos correspondientes a su columna. Así se obtiene una segunda tabla de residuos después de haberles restado el efecto de la fila y de la columna.
6. A cada mediana de cada fila se le resta la mediana común.

Con ello ha terminado la primera iteración. Se pueden hacer más iteraciones por el mismo procedimiento, repitiendo los pasos 3 y 4 hasta que el residuo sea despreciable o lo más cercano a cero posible.

Debe advertirse, sin embargo, que el análisis de medianas puede dar resultados un poco diferentes si el proceso de análisis empieza por filas o por columnas y que también se ve afectado por el número de iteraciones que se hagan, sin alterar las conclusiones globales que permite extraer.

Resumiendo, podemos decir que la técnica que introducimos ofrece (Freixa y cols., 1992) las siguientes ventajas para la exploración de tablas de dos factores:

a) No es preciso asumir los rígidos supuestos de un modelo lineal.

b) Puede analizarse con todo tipo de datos (puntuaciones directas, porcentajes, proporciones, medias, medianas, etc.).

c) Puede efectuarse el análisis con datos incompletos (casillas vacías).

d) Es resistente.

e) Explora la estructura aditiva entre las variables y, mediante otras técnicas EDA, establece la transformación más adecuada para conseguirla.

f) 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.

g) Es, en general, más flexible y por tanto tiene gran diversidad y riqueza de análisis y aplicaciones.

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.

Para poner de manifiesto la utilidad de esta técnica planteamos la siguiente situación: se realizó un estudio longitudinal con 1.500 niños nacidos en el año 1958 en la provincia de Barcelona, sobre el condicionamiento social de las aptitudes intelectuales y su influencia sobre el rendimiento escolar (Freixa, 1983). Uno de los objetivos de dicha investigación era averiguar cuáles eran las variables que mejor predicen el logro o éxito en cuanto al nivel de estudios alcanzado. Para ello, entre otros, se siguió el rendimiento académico de los sujetos durante trece años, midiéndose además variables tales como el tipo de colegio, nivel sociocultural de los padres y nivel de aspiración de los sujetos. Cuando los sujetos cumplieron 25 años, 259 de la muestra inicial finalizaron licenciaturas o diplomaturas universitarias, obteniéndose la siguiente tabla de frecuencias:

TABLA 3. NÚMERO DE SUJETOS CON LICENCIATURA O DIPLOMATURA UNIVERSITARIA SEGÚN LA CLASE SOCIOCULTURAL DE LOS PADRES Y SEGÚN TIPO DE COLEGIO (Tomado de Freixa, 1983)

Nivel sociocultural padres		Tipo de colegio							
		Públicos				Privados			
		Com. 1	Cint. 2	Urb. 3	Bcn. 4	Com. 5	Cint. 6	Urb. 7	Bcn. 8
Muy baja	1	2	2	3	3	3	4	4	6
Baja	2	3	3	3	4	4	5	6	8
Media	3	4	4	5	5	6	7	9	10
Alta	4	5	5	6	7	7	8	10	12
Muy alta	5	6	6	7	7	9	12	14	15

Com. = Comarcas Barcelona.
Cint. = Cinturón Barcelona.
Urb. = Ciudades provincia Barcelona.
Bcn. = Barcelona ciudad.

Después de dos iteraciones de ajuste de medianas se obtiene la siguiente tabla (Programa *Statgraphics*):

TABLA 4. RESIDUALES DESPUÉS DE 2 ITERACIONES A PARTIR DE LA TABLA 3

	1	2	3	4	5	6	7	8	EF.
1	0.75	0.75	1.00	0.00	0.00	0.00	−2.00	−1.25	−2.50
2	0.75	0.75	0.00	0.00	0.00	0.00	−1.00	−0.25	−1.50
3	0.00	0.00	0.25	−0.75	0.25	0.25	0.25	0.00	0.25
4	−0.25	−0.25	0.00	0.00	0.00	0.00	0.00	0.75	1.50
5	−0.75	−0.75	−0.50	−1.50	0.50	2.50	2.50	2.25	3.00
EC.	−1.75	−1.75	−1.00	0.00	0.00	1.00	3.00	4.25	5.50

ECOM

EF. = Efectos Fila; EC. = Efectos Columna; ECOM. = Efecto Común.

de tal manera que el logro (puntuación de la fila 5 y columna 8) se puede descomponer del siguiente modo:

$$15 = 5.5 + 3.00 + 4.25 + 2.25$$

retomando las categorías definidas en la tabla de datos inicial:

$$\text{logro} = \text{efecto común} + \text{efecto nivel padres} + \text{efecto colegio} + \text{residual}$$

Del análisis de los residuales de la Tabla 4 se desprende, obviamente, que en aquellas casillas en las que el residual es 0 o cercano a ese valor; la variable dependiente (en este caso el logro) quedaría explicada por la actuación de las variables independientes. Análogamente, aquellas casillas con residual alto (véase, por ejemplo, las últimas columnas de la octava fila) ponen de manifiesto la existencia de algún efecto interactivo, o la existencia de otra variable no considerada en el modelo ajustado, y que es relevante para la descomposición exhaustiva de la variable dependiente.

Este tipo de estrategia se ve completada con la aplicación de los mismos presupuestos en el caso de las medidas repetidas, que complica un tanto el desarrollo. Siguiendo la filosofía general de las técnicas EDA, el estudio de la tabla de residuales obtenida ofrece información muy valiosa al investigador. Por ejemplo, la distribución interna de los signos de esos residuales informa de la posible estructura interna de los datos originales. Por otro lado, esos mismos residuales son el punto de partida de ulteriores análisis, como el gráfico de diagnóstico, que permiten la evaluación de la aditividad del modelo propuesto. La vinculación de esta técnica con el concepto de ajuste se ve fomentada con la utilización de la línea resistente (tema que abordaremos en este trabajo a continuación) como elemento tendente a un mejor ajuste. Un análisis más exhaustivo de esta temática se ofrece en Freixa y cols. (1992).

Por otro lado, una de las ventajas del ajuste de medianas es, como ya se ha comentado, la resistencia ante valores alejados. Pero, ¿cuántos valores pueden ser alterados sin que se modifique la estructura o forma de los residuales

hallados? De este modo, el concepto anteriormente planteado de *punto de colapso* adquiere su aplicación en esta técnica, a pesar de ser algo más compleja su definición e interpretación estadística en este ámbito.

Para aclarar este aspecto deberemos desglosar, de acuerdo con distintos autores (Hoaglin, Mosteller & Tukey, 1982), dos conceptos vinculados al punto de colapso general. Por una parte, el denominado *worst-case breakdown bound* (que podríamos traducir por límite de colapso en el caso más desfavorable —WCBB—), definido por la siguiente expresión:

$$\text{WCBB} = \frac{1}{2 \text{ máx } (I,J)} - \frac{2 - d(\min (I,J)}{2(IJ)}$$

siendo I el número de filas y J el número de columnas de la tabla de datos y adoptando *d* los siguientes valores:

$$d(n) = \begin{cases} 0 \text{ si } n \text{ es par} \\ 1 \text{ si } n \text{ es impar} \end{cases}$$

Los datos de la tabla anterior (Tabla 5 filas (I) × 8 columnas (J)), presentan el siguiente valor de límite de colapso para los casos desfavorables:

$$\text{WCBB} = \frac{1}{2 \text{ máx } (5,8)} - \frac{2 - d\ [\text{mín } (5,8)]}{2\ (40)} = \frac{1}{16} - \frac{1}{80} = \frac{2}{40}$$

Es decir, si más de 2 de los 40 valores son anómalos o varían (siempre en el mismo sentido) y están todos situados en una misma columna *j*, la estimación del efecto de esa columna estará especialmente afectada. De ello se desprende que en esta situación es especialmente importante el análisis de la ubicación de esos valores anómalos. Debe advertirse, pues, que en el caso de tablas con casillas vacías (*missings*) el proceso iterativo es más largo y el punto de colapso se altera.

También, de forma análoga a lo visto hasta este momento, se puede plantear el límite de colapso de la configuración más favorable (WPBB) (*well-placed*) que, en general, es de más difícil interpretación.

Si llamamos B al número de observaciones desfavorables que el ajuste de medianas puede tolerar, con la condición de que esas observaciones estén situadas en la configuración más favorable posible entonces, en general, la razón (B/IJ) se constituye como el punto de colapso de la configuración más favorable que antes hemos mencionado.

De acuerdo con Hoaglin, Mosteller y Tukey (1982), pueden plantearse diversas situaciones para el cálculo de este punto de colapso particular, dependiendo, en todos los casos, de la configuración I × J de la tabla inicial.

En nuestro caso, utilizando como antes la configuración de la tabla inicial, el WPBB adopta el siguiente valor (teniendo en cuenta que I es impar y J es par y que $J<2I$):

$$WPBB = B = ((1/2)IJ) - I = (1/2)(5)(8) - 5) = 15$$

Como se ha dicho, la interpretación de los límites de colapso es compleja pero, siguiendo a sus autores y siendo consecuentes con la existencia de limitaciones en cuanto al número de valores anómalos que la/s mediana/s puede/n tolerar, hemos de considerar, en primer lugar, el límite de valores anómalos (WCBB) y mitigar, en cierta manera, este límite, considerando el número máximo de valores bien situados (WPBB). Es decir, podemos definir un intervalo que incluya la fracción de valores anómalos que el ajuste de medianas puede tolerar.

Con ello se plantea de nuevo la necesidad de estudiar con exhaustividad la estructura de la tabla inicial de los datos, pero ello excede a las pretensiones de este trabajo ilustrativo. No obstante, debemos resaltar el uso del WCBB, puesto que pone de manifiesto que, incluso en técnicas resistentes como la que nos ocupa, los valores anómalos pueden alterar el resultado final de su aplicación.

Línea resistente

Presentamos a continuación una técnica dedicada al estudio de relaciones bivariables entre variables con escalas ordinales, intervalo o de razón. Es decir, si el ajuste de medianas incorpora variables categorizadas, esta propuesta contempla el uso de variables como mínimo de escala ordinal. Se trata, en términos generales, de una estrategia para ajustar una recta a una nube de puntos bivariables, usando para ello las ya conocidas características de ausencia de supuestos y de resistencia. Lógicamente, las similitudes entre el modelo lineal de la regresión y la línea resistente son más que aparentes y, en consecuencia, será preciso no olvidar la primera en la exposición de la segunda (Freixa y cols., 1992). Por lo dicho hasta este momento es evidente que la expresión general de la línea resistente se ajustará a la siguiente fórmula:

$$Y = b_0 + b_1X$$

La existencia de la expresión general $Y=f(x)$ nos acerca a los presupuestos clásicos del Modelo Lineal General, pero sin los supuestos propios del mismo y con la relajación de no ser precisa la existencia de escalas de intervalo o de razón en las variables implicadas. Veamos, esquemáticamente, cómo se procede para la obtención de los dos coeficientes resistentes (b_0 y b_1):

a) Se ordenan los pares de valores de menor a mayor según el dominio de la variable X.

b) Se establecen tres grupos (inferior, medio y superior), cada uno de 1/3 de la muestra total aproximadamente (Velleman y Hoaglin, 1981).

c) Se calculan las medianas en X e Y para cada tercio, obteniéndose así seis puntos resumen:

Tercio Inferior	X_i	Y_i
Tercio Medio	X_m	Y_m
Tercio Superior	X_s	Y_s

Con estos valores, se plantea la forma más simple de cálculo de los coeficientes:

$$b_1 = (Y_s - Y_i) / (X_s - X_i)$$
$$b_0 = 1/3\ (b_{0i} + b_{0m} + b_{0s})$$

siendo

$$b_{0i} = Y_i - b_1 X_i$$
$$b_{0m} = Y_m - b_1 X_m$$
$$b_{0s} = Y_s - b_1 X_s$$

Existen formas más complejas de cálculo de los coeficientes (Johnstone y Velleman, 1982; Velleman y Hoaglin, 1981: Emerson y Hoaglin, 1985, entre otras).

Se puede plantear el uso de la línea resistente como forma exploratoria previa y complementaria del modelo lineal de la regresión y ello ofrece, sin duda, posibilidades muy sugerentes para un análisis estadístico original. Sin embargo, aquí nos ocuparemos en presentar un índice surgido de la propia línea resistente y que se dedica a la evaluación de la posible linealidad de la nube de puntos inicial, aspecto éste crucial en el análisis confirmatorio. Este indicador se denomina semipendiente y queda definido por la expresión:

$$\text{Semipendiente } ½(b) = b(\text{inf})/b(\text{sup})$$

donde

$$b(\text{inf}) = (Y_m - Y_i) / (X_m - X_i)$$
$$b(\text{sup}) = (Y_s - Y_m) / (X_s - X_m)$$

Es relativamente fácil unir la interpretación de la semipendiente y la actuación estadística con las variables originales para transformarlas y conseguir la linealidad necesaria en el análisis confirmatorio. El siguiente esquema resume estos aspectos (tomado de Freixa y cols., 1992):

CUADRO 1. CRITERIOS DE INTERPRETACIÓN DE LA SEMIPENDIENTE

Valores de semipendiente	Tratamiento de la nube original
$0.9 \leq \frac{1}{2}(b) \leq 1$	La relación es lineal.
$0.5 \leq \frac{1}{2}(b) < 0.9$	Una transformación adecuada permitirá la linealidad de la nube inicial.
$0 \leq \frac{1}{2}(b) < 0.5$	Se puede plantear una transformación en X o Y. Si el valor de $\frac{1}{2}(b)$ es muy cercano a 0, es factible que ni con transformaciones se consiga la linealidad.
$\frac{1}{2}(b) < 0$	No es factible ninguna transformación. Existe un cambio de dirección en la función teórica de la nube.

Una forma muy sencilla de ejemplificar el uso de la semipendiente se puede efectuar con los datos siguientes, provenientes de una situación ampliamente conocida. Supongamos que en una muestra de 10 sujetos fóbicos, evaluamos su intensidad mediante dos instrumentos clásicos: un cuestionario conductual estructurado en base a 100 situaciones ordenadas de menor a mayor intensidad y un autoinforme de ansiedad con un rango de 0 a 10. Sería factible obtener una nube de puntos con los siguientes valores

TABLA 5. VALORES DE DOS REGISTROS EN SUJETOS FÓBICOS

Sujeto	Cuestionario	Autoinforme
01	52	5
02	54	5
03	58	6
04	62	6
05	69	6
06	72	7
07	80	8
08	84	8
09	86	9
10	88	9

Con estos datos, y considerando como variable dependiente a los valores del autoinforme, se obtienen los siguientes resultados:

$b(inf) = 0{,}0909$
$b(sup) = 0.1613$
$\frac{1}{2}(b) = 0.5635$

Con este dato, la linealidad de la nube está comprometida, pero sugiere que una ligera transformación hará que la nube adquiera la condición necesaria para su análisis confirmatorio. Por ejemplo, si sometemos a la variable X (valores del cuestionario conductual) a una transformación logarítmica (base 10), el

nuevo valor de la semipendiente se sitúa en 0.8901. Ello nos permite decidir que esa ligera transformación nos llevaría a una linealidad mucho menos comprometida que con los datos iniciales.

Es importante señalar que ésta es sólo una posible utilización de la línea resistente y, por supuesto, es preciso un análisis más exhaustivo de su uso que el que aquí hemos presentado.

Comentario final

A través de las anteriores líneas creemos que hemos puesto de manifiesto con suficiente claridad que con las técnicas y gráficos EDA obtenemos información estadística de una manera rápida y sencilla. Estos ejemplos, desde un punto de vista más crítico y amplio, nos llevan a reflexionar sobre la utilidad y representatividad de los estadísticos clásicos que en algunos casos no son los más adecuados, puesto que no gozan de la propiedad de la resistencia en este trabajo planteada. Sirva el estudio, pues, para que el usuario estadístico tome conciencia de que hay otras muchas técnicas empíricas que pueden ser adecuadas a las características de cada estudio, complementarias o alternativas a las clásicas.

Tukey (1977) afirma, en una frase convertida ya en una declaración de intenciones, que las técnicas EDA se centran en un *trabajo de detective numérico* para evitar confundir, mentir o cometer errores al utilizar la estadística.

Las técnicas presentadas en este artículo son algunas de las que el análisis exploratorio de datos ofrece, siendo todas ellas muy útiles para analizar datos en varios contextos, y de forma particular, en la investigación psicológica.

Referencias

Buja, A. & Tukey, P.A. (1991). *Computing and Graphics in Statistics*. New York: Springer-Verlag.

Emerson, J.D. & Hoaglin, D.C. (1985). Resistant Multiple Regression, one variable at a time. In D.C. Hoaglin; F. Mosteller & J.W. Tukey (Eds.), *Exploring Data Tables, Trends and Shapes*, pp. 241-280. New York: John Wiley & Sons.

Freixa, M. (1983). *El condicionamiento social de las aptitudes intelectuales y su influencia sobre el rendimiento escolar*. Tesis doctoral no publicada. Universidad de Barcelona.

Freixa, M., Salafranca, Ll., Guàrdia, J., Ferrer R. y Turbany, J. (1992). *Análisis Exploratorio de Datos: Nuevas Técnicas Estadísticas*. Barcelona: PPU.

Hoaglin, D., Mosteller, F. & 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. Genève.

Johnstone, I. & Velleman, P.F. (1982). Tukey's resistant line and related methods: asymptotics and algorithm. 1981 *Procedings of the Statistical Computing Section*. Washington D.C.: American Statistical Association, pp. 218-223.

Marsh, C. (1988). *Exploring Data. An Introduction to Data Analysis for Socials Scientist*. Polity Press: Cambridge.

Tufte, E.R. (1987). *Envisioning information*. Chesire: Graphics Press.

Tufte, E.R. (1991). *The Visual Display of Quantitative Information*. Chesire: Graphics Press.

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

Velleman, P.F. & Hoaglin, D.C. (1981). *Applications, Basics and Computing of Exploratory Data Analysis*. Boston: Duxbury.