ANÁLISIS EXPLORATORIO DE DATOS: SUS POSIBILIDADES EN LA ENSEÑANZA SECUNDARIA

C. Batanero, A. Estepa y J. D. Godino

Suma, nº 9, 1991: 25-31

En la actualidad la enseñanza de la Estadística se realiza de forma gradual desde séptimo de E.G.B. hasta los primeros cursos universitarios. en donde existen asignaturas de Estadística Aplicada en distintas licenciaturas: Medicina, Biología, Farmacia, Economía, Psicología, Ciencias de la Educación, ingeniería, etc.

Debido al espectacular desarrollo de la informática y a la disponibilidad de paquetes de cálculo, fácilmente manejables y accesibles, asistimos en nuestros días a una demanda cada vez mayor de formación estadística, lo que sin duda ha contribuido a que en los diseños curriculares propuestos para la Reforma de las Enseñanzas no Universitarias estos contenidos reciban mayor peso. Baste citar que de los cinco bloques de contenido que se proponen para la Enseñanza Secundaria Obligatoria, dos están relacionados con la Estadística: el *4, "Interpretación. representación y tratamiento de la información"* el 5, "*Tratamiento del azar"*. (Diseño Curricular Base). Este fenómeno no es exclusivo de nuestro país: la enseñanza de los contenidos referidos a estadística y probabilidad se acentúa en los nuevos planes de estudio de diferentes países. Como ejemplos podemos citar el Curriculum Nacional de Inglaterra y Gales (Department of Education and Science and the Welsh Office, 1989) puesto en funcionamiento en 1988 y las recomendaciones sobre el tema incluidas en los Curriculum and Evaluation Standards del N.C.T.M (1989) en Estados Unidos.

Pero si queremos aprovechar las posibilidades de renovación cumcular que se manifiesta en la intención de la reforma, desde las primeras tomas de contacto con el tema, debiera tratarse el análisis de datos desde un punto de vista exploratorio, como se está iniciando en otros países, máxime teniendo a nuestra disposición herramientas tan poderosas como los microordenadores.

Enfoques confirmatorio y exploratorio en el análisis de datos

Las capacidades de cálculo y representación gráfica de los ordenadores actuales, permiten de una forma sencilla, la obtención de una amplia variedad de gráficos y estadísticos diferentes y han hecho posible la aparición de una nueva filosofía en los estudios estadísticos: el análisis exploratorio de datos, introducido por Tukey (1977).

Anteriormente a este enfoque, el análisis de datos se basaba fundamentalmente en el cálculo de estadísticos, conduciendo a dos consecuencias: En primer lugar se disminuía la importancia visual de la representación de los datos, dándosela exclusivamente a los cálculos y en segundo se equiparaba el análisis con el modelo confirmatorio.

En este tipo de análisis el conjunto de valores de las variables observadas se supone que se ajusta a un modelo preestablecido, calculando los estadísticos para aceptar o no una hipótesis, que es previa a la toma de las observaciones, las cuales han sido recogidas con el único propósito de poner tal hipótesis a prueba. Al contemplar solamente dos alternativas, confirmación o no de la hipótesis, los datos no se suelen explorar para extraer cualquier otra información que pueda deducirse de los mismos.

Para entender los principios por los que se guía el análisis exploratorio, se ha de tener en cuenta que los datos están constituidos por dos partes:
la "regularidad" y las "desviaciones". La regularidad indica la estructura simplificada de un conjunto de observaciones (en una nube de puntos, por ejemplo, es la recta a la cual se ajusta). Las diferencias de los datos con respecto a esta estructura (diferencia en nuestro caso respecto a la recta), representan las desviaciones o residuos de los datos, que usualmente no tienen por qué presentar una estructura determinada.

Tradicionalmente el estudio se ha concentrado en la búsqueda de un modelo que exprese la regularidad de las observaciones. Por el contrario. el análisis exploratorio de datos es básicamente el

desglose de los mismos en las dos partes que hemos citado. En lugar de imponer, en hipótesis, un modelo a las observaciones, se genera dicho modelo desde las mismas. Por ejemplo, cuando se estudian las relaciones entre dos variables, el investigador no solamente necesita ajustar los puntos a una línea recta, sino que estudia los estadísticos, compara la línea con los residuos, estudia la significación estadística del coeficiente de correlación u otros parámetros para descubrir si la relación entre las variables se debe o no al azar. Aunque los estadísticos calculados presenten un valor estadísticamente significativo (en el ejemplo, el coeficiente de correlación sea significativamente distinto de cero), la relación entre las variables puede no ajustarse bien a una línea recta. En este caso al investigador le faltaría descubrir algo importante: el modelo latente no es el esperado.

Características del análisis exploratorio de datos

Como hemos indicado, nos encontrarnos ante una nueva filosofía en la aplicación de los métodos de análisis de datos, aunque unida a ella se han desarrollado también algunas técnicas concreta para su aplicación. Esta filosofía consiste en el estudio de los datos desde todas las perspectiva, y con todas las herramientas posibles, incluso las ya existentes. El propósito es extraer cuanta información sea posible, generar hipótesis nuevas, en el sentido de conjeturar sobre las observaciones de las que disponemos.

Como contrapartida, tales "hipótesis" no quedan contrastadas en el sentido estadístico del término al finalizar el análisis, por lo que seria preciso la toma de nuevos datos (una replicación) sobre el fenómeno y efectuar sobre ellos un análisis estadístico tradicional con el fin de contrastarlas. Por ello, el análisis exploratorio se utiliza especialmente en las fases iniciales del estudio experimental en las diversas ciencias –Biología, Ciencias Humanas, Economía, en las que se dispone de poca información sobre los objetos bajo estudio, siendo especialmente útil en el denominado paradigma cualitativo de investigación.

Al considerar la conveniencia o no de incluir un tema como objeto de enseñanza hemos de tener en cuenta su utilidad y que este tema se halle al alcance de los alumnos. Además de la utilidad, ya razonada, el análisis exploratorio de datos tiene las siguientes características que lo hacen un tema apropiado de estudio en la enseñanza secundaria:

- ***Posibilidad de generar situaciones de aprendizaje referidas a temas de interés para el alumno***. Lo usual es trabajar sobre un fichero de datos que han sido codificados previamente e introducidos en el ordenador, ya que se pretende estudiarlos mediante cuantas perspectivas y técnicas tengamos a nuestro alcance. Estos conjuntos de datos pueden ser obtenidos por los mismos estudiantes, mediante la realización de una encuesta a sus compañeros sobre temas diversos, como características físicas, aficiones, empleo del tiempo libre, etc., o incluyendo valores de variables relacionadas con otras áreas curriculares obtenidos en anuarios o publicaciones estadísticas.

- **Fuerte apoyo en representaciones gráficas**: *"Una idea fundamental del análisis exploratorio de datos es que al usar representaciones múltiples de los datos se convierte en un medio de desarrollar nuevos conocimientos **y** perspectivas. Esto puede ejemplificarse al pasar de tablas a gráficos, de lista de números a representaciones como la del "tronco", reduciendo los números a una variedad discreta en un mapa estadístico para facilitar la exploración de la estructura total, construyendo gráficos, como el de la "caja" que hace posible la comparación de varias muestras"*. (Biehler, 1998a,pg.2).

- **Empleo preferente de los estadísticos de orden**, porque son sensibles a la mayor parte de los datos y con ellos se disminuye el efecto producido por los valores atípicos, escasos y muy alejados de la norma.

- **No necesita una teoría matemática compleja,** *"Como el análisis de datos no supone que estos se distribuyen según una ley de probabilidad clásica (frecuentemente la normal, no utiliza sino nociones matemáticas muy elementales y procedimientos gráficos fáciles de realizar.*

- *Hasta aquí es, pues,. bastante parecida a la estadística descriptiva tradicional, pero se aleja de ella por su intención. Pues, al contrario que en ella, la representación o el cálculo no son en el análisis exploratorio de datos un fin, sino un medio de descubrir la información oculta en los mismos".* Jullien y Nin 1989. págs. 30-3 1.)

- **Uso de diferentes escalas o reexpresión**: La escala en la que una de las variables es observada y registrada no es única. A veces, transformando los valores originales de la variable a una nueva escala se puede lograr que dichos valores sean más manejables. De este modo se incluye también el empleo de otros contenidos matemáticos, especialmente los referidos al concepto de función **y** el estudio de las propiedades de las funciones elementales.

En resumen, como indica Biehler 1998b, pg**.** 5: *'El currículum tradicional de Estadística Descriptiva debiera transformarse en dirección al análisis exploratorio de datos. Seria esencial, sin embargo, dar apoyo sustancial a la actitud investigadora, contra la tendencia de la mayor parte de las transposiciones didácticas de reducir el conocimiento a la técnica.*

Técnicas elementales de análisis exploratorio de datos

Aunque, como hemos indicado, lo usual en este enfoque seria trabajar con un ordenador, muchas de estas técnicas son sencillas de realizar incluso a mano. En lo que sigue, y para ejemplificar algunas de ellas, trabajaremos con el conjunto d datos que se muestra en la Figura 1, y que supondremos ha sido recogido en clase por los propios alumnos.

Peso en Kg.

Varones	Hembras
55 64 70 74 75 70	60 45 46 50 47 55
64 93 60 62 70 80	49 52 50 46 50 52
61 60 62 68 65 65	52 48 52 63 53 54
66 68 70 72 72 71	54 54 53 55 57 44
	56 56 56 53 60 65
	67 61 68 55 64 60

Figura 1

Gráfico del tronco

El gráfico del tronco (en inglés -stem and leaf-) fue descrito por Tukey y es utilizado para la representación de distribuciones de variables cuantitativas, consiguiendo con él, además de una gráfica de la distribución, la visualización de los valores de los datos que estamos estudiando. Para realizar este gráfico procederemos de la siguiente forma:

- Se redondean los datos a dos o tres cifras, expresando los valores con números enteros. En nuestro ejemplo, puesto que los datos disponibles constan sólo de dos cifras, este paso no es necesario.

- Se ordenan de menor a mayor, como se muestra en la Figura 2.

44 45 46 46 47 48 49 50 50 50 52 52 52 52 53
53 53 54 54 54 55 55 55 55 56 56 56 57 60 60
60 60 60 61 61 62 62 63 64 64 64 65 65 65 66
67 68 68 68 70 70 70 70 71 72 72 74 75 80 93

Figura 2

- Se separan por la izquierda uno o más dígitos de cada dato, según el número de filas que se quiera obtener, en general no más de 12 ó 15. Cada uno de estos valores se escriben uno debajo del otro, trazando una línea a la derecha de los números escritos. Estas cifras constituyen el “tronco”. En nuestro caso tomaremos la primera cifra para formar con ella el “tronco”.

- Para cada dato original se buscan los dígitos escritos en el tronco y a la derecha de los mismos se escriben las cifras que nos habían quedado. Estas cifras forman las “hojas”.

De este modo obtenemos el gráfico del tronco para nuestros datos (Figura 3).

Gráfico del tronco del peso de los alumnos

4	4 5 6 6 7 8 9
5	0 0 0 2 2 2 2 3 3 3 4 4 4 5 5 5 5 6 6 6 7
6	0 0 0 0 0 1 1 2 2 3 4 4 4 5 5 5 6 7 8 8 8
7	0 0 0 0 12 2 4 5
8	0
9	3

Figura 3

Como se observa, el resultado es, en la práctica, un histograma de amplitud de intervalo 10, que además de mostrarnos la forma de la distribución, presenta todos los datos ordenados. Esta representación puede ser ampliada o condensada para aumentar o disminuir el número de filas, subdividiendo o fundiendo dos o más filas adyacentes. Por ejemplo, para extender el gráfico de la Figura 3, podemos subdividir en dos cada fila de la siguiente forma: marcamos con un asterisco las filas cuyos dígitos de la derecha varian de 0 a 4 y con un punto las filas cuyos dígitos de la derecha varían de 5 a 9. Este nuevo diagrama, que podemos observar en la Figura 4, recibe el nombre de gráfico del tronco extendido.

Gráfico del tronco extendido del peso de los alumnos

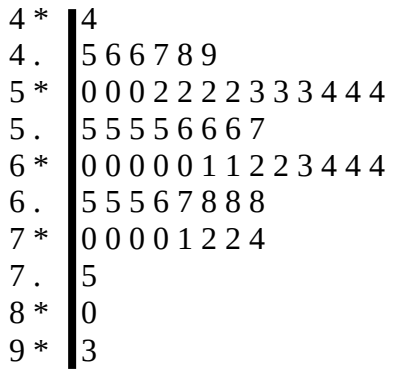

Figura 4

Al comparar el gráfico del tronco con un histograma de frecuencias observamos las siguientes ventajas:

- Su fácil construcción, especialmente con papel cuadriculado.

- Se pueden observar los datos con más precisión que en el histograma, pues los rectángu1os pueden ocultar diferencias importantes entre los valores, mientras que en el gráfico del tronco estas lagunas pueden ser fácilmente detectadas y observadas.

- Pueden obtenerse a partir de él rápidamente los estadísticos de orden, como los valores máximo y mínimo, la mediana, cuartiles, percentiles y sus rangos, así como la moda.

Como contrapartida observamos que no podemos elegir la amplitud del intervalo. como en el caso del histograma, sino que viene impuesta por el sistema de numeración. Tampoco podemos escoger la escala de la representación gráfica, que viene impuesta por el espaciado del papel empleado.

Además de emplear este gráfico para el estudio de una variable aislada, también puede servir para establecer comparaciones entre dos distribuciones, observando claramente las diferencias y semejanzas en las mismas. Este hecho es importante a nuestro juicio, pues permite desde los comienzos de los estudios de la estadística una iniciación intuitiva para el estudio posterior del contraste de hipótesis, como ilustraremos con el estudio separado de los pesos para chicos y chicas en nuestro ejemplo.

Construyendo un gráfico del tronco separado para cada uno de estos conjuntos de datos y uniéndolos en la forma expuesta en la Figura 5 podemos hacer ver a los alumnos qué quiere decir que las chicas pesan menos que los chicos: no se trata de todos los casos, pues hay varones y hembras de igual peso. Tampoco se trata de casos aislados: al comparar las frecuencias relativas de pesos en los mismos intervalos observamos que esta frecuencia es distinta en los dos sexos, al menos para algunos intervalos. Estas diferencias de frecuencias, también se corresponden con otras en los distintos estadísticos: máximo, mínimo, moda, mediana, media.

Estas diferencias son en este ejemplo muy acentuadas y pueden servir también, como una primera introducción intuitiva al concepto de asociación y dependencia estadística entre variables. Cuando en lugar de tomar en clase observaciones de dos variables distintas (peso y sexo en nuestro caso) hemos recogido varias sobre los mismos individuos (sexo, peso, altura, longitud de brazos, fumar o no, practicar o no deporte, pulsaciones por minuto en reposo, etc.) se puede hacer ver los distintos grados de dependencia de una de las variables con las

restantes, especialmente cuando combinamos el uso de estos métodos con las técnicas más clásicas: histogramas, cálculo de estadísticos, tablas de contingencia, representación de nubes de puntos y cálculo de la recta de regresión.

Gráfico del tronco extendido del peso para varones y hembras

CHICOS		CHICAS
	4*	4
	4.	566789
	5*	000222233344
5	5.	5556667
4422100	6*	000134
88655	6.	578
42210000	7*	
5	7.	
0	8*	
3	9	

Figura 5

Gráfico de la «Caja»

El gráfico de la caja fue descrito por Tukey [denominándolo *"box and whiskers"*. Para su construcción se utilizan 5 estadísticos de la distribución de frecuencias: el mínimo, el primer cuartil Q_1, la mediana, el tercer cuartil Q_3, y el máximo, procediéndose de la forma siguiente:

- Se traza una línea vertical u horizontal de longitud proporcional al recorrido de la variable, que llamaremos eje (Véase la Figura 6). Los extremos del eje serán el mínimo y el máximo de la distribución, que en nuestro caso son 44 y 93 kilos. En el interior del eje se señalarán las subdivisiones que creamos necesarias, para formar una escala.

Paralelamente al eje se construye una caja rectangular con altura arbitraria y cuya base abarca desde el primer cuartil al tercero. Como vemos esta "caja" indica gráficamente el intervalo de variación del cincuenta por ciento de valores centrales en una distribución que, para el peso de los estudiantes, abarca desde 53 a 66.5.

- La caja se divide en dos partes, trazando una línea a la altura de la mediana (60 kg. en nuestro raso). Cada una de estas partes indica pues el intervalo de variabilidad de una cuarta parte de los datos. De este modo, en el ejemplo dado, una cuarta parte de los alumnos tiene un peso comprendido entre 44 y 53, estando incluidas las otras cuartas partes en los siguientes intervalos de peso: 53 a 60. 60 a 66.5 y 66.5 a 93.

- A la caja así dibujada se añaden dos guías paralelas al eje, una a cada lado, de la forma siguiente: el primero de estos segmentos se prolonga desde el primer cuartil hasta el valor máximo entre el mínimo de la distribución y la diferencia entre el primer cuartil y una vez y media el recorrido intercuartilico. Como en nuestro caso el peso mínimo es 44 kilos, y el recorrido intercuartilico es 66.5 - 53 = 13.5, al restar al primer cuartil, Q_1= 53 una vez y media el recorrido intercuartílico obtenemos:

$$Q_1 - 1.5\ RI = 53 - 20.25 = 32.75$$

El máximo entre 44 y 32.75 es 44, por lo que el segmento inferior que debe dibujarse en el gráfico de la caja debe llegar hasta 44, como se muestra en la Figura 6.

El segmento dibujado al otro lado de la caja abarca desde el tercer cuartil hasta el mínimo entre el mayor de los datos y la suma del tercer cuartil con una vez y media el recorrido intercuartílico. En el peso de los alumnos el máximo es 93 kilos y, al sumar una vez y media el recorrido intercuartilico al cuartil superior 66.5. obtenemos:

$$Q_3 + 1.5\ RI = 66.5 + 20.25 = 86.75$$

De este modo, el extremo superior del segmento debe prolongarse ahora sólo hasta 86.75

Si alguno de los datos queda fuera del intervalo cubierto por la caja y estos segmentos, como ocurre en el ejemplo con el alumno que pesa 93 kg, se señala en el gráfico mediante un asterisco o cualquier otro símbolo, como puede verse en la Figura 6.

Estos datos son los llamados valores atípicos (“outliers” en la terminología anglosajona), que son valores muy alejados de los valores centrales de la distribución. En la distribución normal, fuera del intervalo que resulta de extender los cuartiles en una vez y media el recorrido intercuartilico, sólo
aparecen un uno por ciento de los casos, por lo que estos valores, si no son debidos a errores, suelen ser casos excepcionales.

Gráfico de la caja para el peso de los alumnos

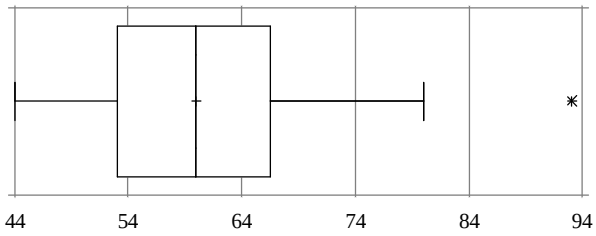

Figura 6

Como vemos en el ejemplo dado, este gráfico nos proporciona, en primer lugar, la posición relativa de la mediana, cuartiles y extremos de la distribución. En segundo lugar, nos proporciona información sobre los valores atípicos, sugiriendo la necesidad o no de utilizar estadísticos robustos. En tercer lugar, nos informa de la simetría o asimetría de la distribución, y posible normalidad o no de la misma.

El gráfico de la caja también se puede utilizar para comparar la misma variable en dos muestras distintas, como se muestra en la Figura 7 al comparar los pesos de chicos y chicas.

Gráfico de la caja para los pesos de chicos y chicas

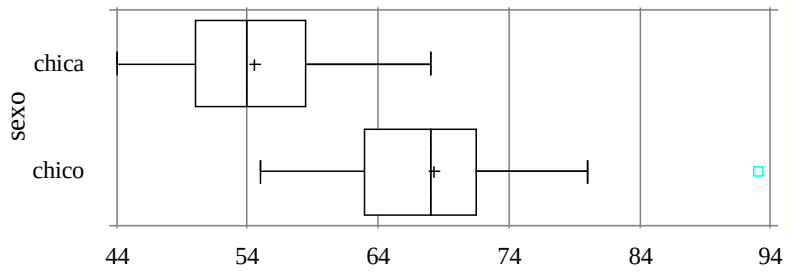

Figura 7

Además de estas representaciones gráficas existen otras específicas, tanto para datos univariantes, como para representar conjuntamente un grupo de variables. Una buena exposición de estos gráficos y su utilización se puede estudiar en Chambers y cols. (1983). Asimismo, para el estudio conjunto de dos variables se emplean, además de la regresión lineal ordinaria y diversos tipos de regresión no lineal, la recta de regresión respecto a la mediana, es decir, cuando la suma de cuadrados de residuos que se trata de minimizar se refiere a la mediana (Hartwing y Dearing 1979). En el análisis multivariante, que de momento creemos queda fuera de este nivel de enseñanza, la mayor parte de las técnicas pueden ser empleadas tanto para realizar un análisis clásico como con un enfoque exploratorio.

Programas disponibles para el análisis exploratorio de datos

Como hemos indicado, lo habitual en este enfoque es trabajar con un ordenador, para aligerar el trabajo de cálculo y representación gráfica, ya que el interés se centra en la comparación de todas las representaciones posibles sobre unos mismos datos. En la actualidad, muchos paquetes estadísticos profesionales, como pueden ser el BMDP o el S.P.S.S. están incorporando representaciones como el gráfico de la caja o del tronco en algunos de sus subprogramas. Con fines de enseñanza, existen paquetes concebidos para este fin exclusivo, como el distribuido por el N.C.T.M. [8], que viene acompañado con un libro de actividades, o los que se describen en Biehler (1998b). Por nuestra parte, hemos incorporado la posibilidad de realizar el gráfico del tronco y el de la caja en el paquete PRODEST, que se describe en Batanero,Godino y Estepa (1987).

REFERENCIAS

Batanero, C., Estepa, A. y Godino, J. D. (1991). Estrategias y argumentos en el estudio descriptivo de la asociación usando microordenadores. *Enseñanza de las Ciencias,* 9(2), 145-150.

Batanero, C., Godino, J. D. Y Estepa, A. (1987). Un paquete didáctico de programas para el laboratorio de estadística. Actas del 1 Simposio Internacional de Educación e Informática. I.C.E. de la Universidad Autónoma de Madrid, pp. 380-386.

Biehler, R. (1988.b). A selected bibllography on teaching Exploratory data analysis at school level. Sixth lnternational Congress on Mathematical Education.

Biehler. R. (l988.a). Educational perspectives on exploratory data analysis. Sixth International Congress on Matheinatical Education.

Chambers, J. M., Cleveland, W. S., Kleiner, B y Tukey, P. A. (1983). Graphical methods for data analysis. Duxbury Press.

Department of Education and Science and the Welsh Office (1989). Mathematics for ages 5 to 16. Proposals of the Secretary of State for Education and Science and the Secretary of State for Wales.

Hartwing, F. y Dearing, B. F. (1979). Exploratory data analysis. Sage University Press.

National Council of Teachers of Mathematics (1988). Exploring data.National Council of Teachers of Mathematics (1989). Curriculum and Evaluation Standars for School Mathematics.

Jullien, M. y Nin. G. (1989). L' E.D.A. au secours de I'OG.D. ou quelques remarques concernant l'enseignement de la Statistique dans les colleges. Petit X, 19: 29-41.

Tukey, J. W. (1977). Exploratory Data Analysis. Addisson Wesley.