ANÁLISIS DE DATOS A TRAVÉS DE MÉTODOS GRÁFICOS

Mª Aurelia Noda Herrera
Mª Candelaria Espinel Febles

E.U. de Formación del Profesorado
Universidad de La Laguna

I. INTRODUCCIÓN

La exploración de datos es el primer paso para encontrar relaciones y sugerir hipótesis. De la misma forma que se adquieren habilidades de conteo y cálculo, los alumnos también deberían aprender a encontrar la información contenida en un conjunto de datos. Una forma de aproximarse a esos datos es mediante gráficos.

Las técnicas más frecuentes de análisis de datos se han desarrollado al amparo del ordenador. La mayoría de tales técnicas no se pueden enseñar a cualquier nivel; piénsese, en una matriz de correlación, ANOVA, etc. Sin embargo, existen algunas técnicas elementales que, combinadas con métodos gráficos, pueden ser utilizadas como una forma de iniciar al alumno en las modernas técnicas de análisis de datos mediante ordenador.

Los gráficos que mostramos facilitan el rápido reconocimiento de características de los datos y permiten interpretaciones e inferencias para situaciones más complejas. Los modelos aquí propuestos son: diagrama de *tallo y hoja*, diagrama de *caja* y *dendrograma*.

II. DIAGRAMA DE TALLO Y HOJA

El diagrama de *tallo y hoja* (Stem-and-leaf plot) fue introducido por John W Tukey en 1977. Se trata de una mezcla ingeniosa de tabla y gráfico.

Para construir el diagrama se divide primero cada dato en dos partes; una corresponderá al tallo y la otra a la hoja. Se colocan en columna los distintos valores que toma el tallo, y a la derecha, al lado de la respectiva columna, los que toman las hojas.

En función de los valores de la variable que se analice, se pueden construir diferentes versiones de este tipo de diagrama.

Como gráfico, se asemeja al histograma. Las diferencias significativas respecto de él son las siguientes: los valores numéricos de los datos están todos presentes; muestra de una forma más "visual" el recorrido del conjunto de datos, los valores extremos y los anómalos (outliers), si los hay; permite identificar huecos en los datos, así como reflejar el perfil de la distribución.

EJEMPLO 1. En la tabla 1 se muestran las edades de los padres de 16 alumnos de 3° de B.U.P. de un Centro de Santa Cruz de Tenerife.

ALUMNO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
PADRE	55	48	47	42	58	53	47	41	35	42	39	44	78	42	41	50
MADRE	50	40	47	50	56	47	46	40	39	36	35	37	--	41	39	50

Tabla 1

La fig. 1 recoge el diagrama de tallo y hoja para la edad de los padres y para la edad de las madres, respectivamente, con los datos de la tabla 1.

Fig. 1

PADRE							MADRE				
						3					
				9	5	3	5	6	7	9	9
4	2	2	2	1	1	4	0	0	1		
			8	7	7	4	6	7	7		
				3	0	5	0	0	0		
				8	5	5	6				
						6					
						6					
						7					
					8	7					

Fig. 2

La fig. 2 muestra el diagrama de *tallo y hojas* para los mismos datos. Este permite una mejor comparación entre las edades de los padres y las madres. Además, el tallo se ha separado en más intervalos y los datos de las hojas se presentan ordenados.

III. DIAGRAMA DE CAJA

Otro de los diagramas propuestos por el mismo autor es el de *cajas*. Esencialmente, un diagrama de caja muestra el valor central y los valores extremos de un conjunto de datos, combinando la parte gráfica con la información numérica.

La representación gráfica incluye algunos términos como: mediana, borde superior e inferior, recorrido, frontera superior e inferior y valores anómalos.

El diagrama se puede construir en horizontal o en vertical. Para un diagrama horizontal, primero se marca una escala que cubra el recorrido del conjunto de datos. Debajo de esta escala se traza un rectángulo o caja cuyo largo lo determina el recorrido intercuartílico. Los bordes superior e inferior de la caja toman valores similares a los cuartiles superior e inferior, que corresponden a la mediana de la mitad superior e inferior de los datos, respectivamente.

En la caja se marca la mediana mediante una línea vertical.

Si la mediana está exactamente en el centro de la caja, significa que los datos están distribuidos de forma simétrica. La asimetría de la caja informa sobre el sesgo de los datos.

A partir del borde superior, hasta el dato que tenga el máximo valor, están la cuarta parte de los datos.

La otra cuarta parte está entre el dato de valor mínimo y el borde inferior de la caja.

La falta de simetría en estas dos últimas regiones puede indicar que existen datos anómalos. Dichos datos pueden corresponder a medidas erróneas o a valores posibles pero poco probables.

Para identificar los valores anómalos: se fijan dos valores llamados frontera superior e inferior.

Frontera inferior = borde inferior - amplitud de la caja

Frontera superior = borde superior + amplitud de la caja

Los datos que caen fuera de estos dos valores frontera se consideran anómalos y en el diagrama se marcan mediante asteriscos.

El diagrama se completa dibujando una línea horizontal entre los bordes de la caja y las fronteras. A este último diagrama, en inglés, se le llama box-and-whisker plots y se puede traducir por *"caja y bigote"*.

En ocasiones, dependiendo de los valores de la variable que se analice, es conveniente multiplicar la amplitud de la caja por 1'5 o por 2, a la hora de construir las fronteras, con la finalidad de que queden fuera de los *"bigotes"* sólo los valores que se consideren anómalos. Además, si el valor de la frontera superior es mayor que el valor máximo de la variable motivo de estudio, o el valor de la frontera inferior es menor que el valor mínimo de la variable, el *bigote* se dibuja sólo hasta el valor máximo o mínimo, respectivamente.

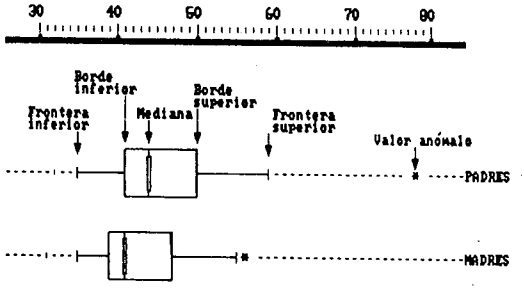

Fig. 3

La fig. 3 muestra dos diagramas de caja obtenidos con los datos de la tabla 1. Mediana (50%), borde inferior (25%), borde superior (75%), frontera inferior (mínimo) y frontera superior, en los padres, toman los valores, respectivamente, 44, 41, 50, 35 y 59.

Además, existe un valor *anómalo*, 78. Para las madres, las mismas medidas toman los valores 41, 39, 47, 35 y 55; quizás, un valor a considerar también como anómalo es 56.

Para ambos diagramas, padres y madres, se obtiene un valor para la frontera inferior más pequeño que el mínimo valor que toma la variable; por lo tanto, se dibuja el *bigote* sólo hasta el respectivo valor mínimo.

Al comparar las dos cajas se observa que: la edad mediana de los padres es 3 años mayor que la de las madres; el borde inferior (25%) de los padres coincide con la mediana (50%) de las madres; el gráfico de los datos de los padres presenta mayor simetría que el de las madres; los datos están mucho más concentrados en las zonas inferiores de los diagramas, especialmente en el de las madres; entre 35 y 41 están el 50% de las edades de las madres, el otro 50% presenta una amplitud que es más del doble.

El diagrama de caja (box-plots) admite varias versiones que se pueden adaptar al nivel y conocimientos de los alumnos con los que se va a trabajar. El objetivo es analizar la concentración y variación de los datos, así como contrastar conjuntos de datos diferentes.

EJEMPLO 2. En la tabla 2 se recogen los datos de 27 alumnos de 2º de B.U.P. de un Instituto de Bachillerato de Santa Cruz de Tenerife, obtenidos al preguntarles el nº de hermanos del padre, de la madre y del alumno (padre, madre y alumno incluidos).

Nº DE HERMANOS	ALUMNOS	PADRES	MADRES
1	2	2	2
2	9	7	4
3	9	4	6
4	3	2	7
5	3	6	3
6	1	1	1
7	–	2	2
8	–	1	1
11	–	2	–
13	–	–	1

Tabla 2

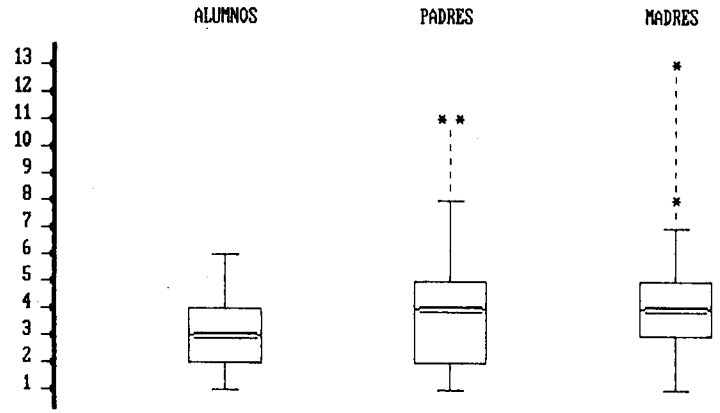

Fig. 4

En la fig.4 se observa el cambio del número de hijos de una generación a otra. Precisamente, una de las ventajas de los diagramas de cajas es poder presentarlas por secuencias que permitan observar algunos cambios sociales.

IV. DENDROGRAMA

Un *dendrograma* o *"árbol de clasificación"* es un modelo gráfico que permite la observación de grupos de individuos o variables, teniendo en cuenta la semejanza de los datos disponibles.

Para la construcción de un dendrograma se ordenan los diferentes valores que toma la variable. Se examinan por parejas los valores adyacentes. Los valores que presentan la menor distancia se unen dando lugar a las ramas del último nivel en el árbol de clasificaciones. Los demás niveles del árbol se obtienen buscando de nuevo los valores que presentan menor distancia en los grupos anteriores. Se repite el proceso hasta llegar a la raíz, esto es, cuando sólo quede un grupo.

En cualquier caso, la distancia entre dos grupos es la distancia entre sus valores más próximos.

La fig. 5 corresponde a dendrogramas construidos con los datos de la tabla 1.

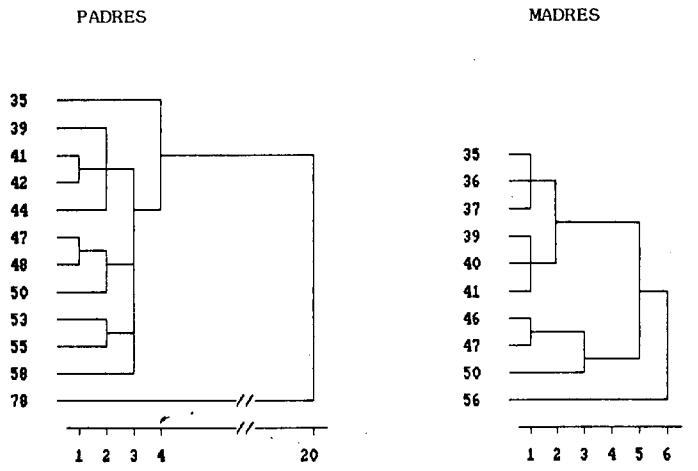

Fig. 5

El objetivo es formar grupos con características similares. En nuestro caso, grupos de padres y madres con edades similares.

El dendrograma de la edad de los padres en la fig. 5 presenta 10 grupos con similaridad 1; 6 grupos con similaridad 2; 3 grupos con similaridad 3, etc. En el de las madres hay sólo 5 grupos con similaridad 1; 4 grupos con similaridad 2, etc. Aún despreciando el valor anómalo 78, se observa que las edades de las madres son mas homogéneas que la de los padres, aunque con similitud 3 hay el mismo número de grupos en los padres y las madres.

V. CONCLUSIONES

La aparición del ordenador ha cambiado la forma tradicional de analizar datos. Se puede iniciar a los alumnos en técnicas que no necesitan de grandes conocimientos matemáticos. Las tres técnicas propuestas se suelen utilizar para un análisis exploratorio de los datos, tarea que tradicionalmente ha correspondido a la Estadística Descriptiva.

Estas técnicas están implementadas en algunos programas de ordenador, primordialmente para dar solución parcial a otros problemas de análisis de datos más complejos. Son paquetes que aún resultan poco manejables por el alumno, sin embargo es posible darlas a conocer y aplicarlas al análisis de datos no muy numerosos.

El diagrama *tallo y hoja* permite observar la distribución de los datos así como prever el valor de las medidas de centralización.

El diagrama de *cajas* resalta la mediana y la dispersión de los datos. Y al señalar los valores anómalos, da una información más veraz que si se utiliza la media y la desviación típica. Su mayor ventaja consiste en que permite comparar medidas de conjuntos de datos diferentes, por medio de una secuencia de cajas, que utilicen la misma escala.

El *dendrograma* prepara para los procesos de clasificación de grupos homogéneos. Permite iniciar al alumno en el "análisis cluster", una de las técnicas fundamentales para buscar grupos similares de individuos en: Biología, Medicina, Sociología, Geografía, Antropología, etc.

BIBLIOGRAFÍA

ANDERBERG, M.(1973) "Cluster Analysis for Applications". Academic Press.

MORRIS, M. (1989) "Estudies in mathematics education. The teaching of statistics". UNESCO.

NCTM (1981) "Teaching stadistics and probability" NCTM.

NCTM (1988) "Data Analysis". Material and software. NCTM.

SPSS/PC (1986) Manual SPSS/PC.