

CONFERENCIAS PLENARIAS

Técnicas modernas para el análisis de datos estadísticos

José Ramón Vizmanos

A modo de presentación

L El análisis de datos ha cambiado muchísimo en estos últimos años con la inclusión de ciertas técnicas que, en general, son muy sencillas y poco conocidas, por ejemplo los diagramas de tallos y hojas, o los diagramas de cajas y bigotes, estos últimos ya incorporados en algunas calculadoras gráficas. La recta de Tukey que figura en el currículo de la ESO. para Andalucía, permite ajustar una recta a una nube de puntos en algunos casos en los que el ajuste mínimo-cuadrático produce resultados no muy buenos.

Por ello, me ha parecido que podría tener cierto interés el conocer algunas de estas técnicas desarrolladas por el matemático americano John Wilder Tukey.

¿Quién es Túkey?

John Wilder Tukey nació el 16 de junio de 1915 en New Bedford, Massachusetts. Después de licenciarse y hacer un master en Química por la Universidad de Brown en 1937, participó en el programa de matemáticas de la Universidad de Princeton donde recibió el master en 1938 y el doctorado en 1939,

Después de su graduación, fue nombrado instructor de Matemáticas de Princeton; 10 años más tarde llegó a ser catedrático. En 1965 la Universidad de Princeton creó el departamento de estadística, y Tukey fue su primer director. Además de su posición en Princeton fue miembro de la Dirección Técnica de los Laboratorios de ATT & Bell desde 1945 hasta su jubilación en 1985 como Director Ejecutivo de Investigación en la División de Información Científica.

Tukey está considerado una de las personas que más ha influido en el análisis exploratorio de datos, habiendo sido el creador de diversas

técnicas, tales como los diagramas de tallos y hojas, los diagramas de cajas y bigotes, etc. También ha hecho importantes aportaciones a la teoría de la estimación y al análisis de las series cronológicas. Tukey es autor de numerosos libros y de más de 350 artículos importantes en matemáticas y estadística. A él también se debe el término bit, contracción de las palabras **b**inary **d**igit.

La participación de Tukey tanto en educación como en servicios de asesoramiento técnico al gobierno ha sido imponente. Fue nombrado miembro del Comité Asesor Científico del Presidente Eisenhower en 1965; ayudó al desarrollo del Programa Nacional para el Progreso de la Educación; y fue un miembro destacado de la Oficina del Censo en 1990.

A lo largo de su vida de trabajo ha conseguido todo tipo de honores, tan sólo mencionaré las más importantes, recibió la Medalla Nacional de la Ciencia, la Medalla de Honor de IEEE, la Medalla James Madison de la Universidad de Princeton y miembro extranjero en la Royal Society de Londres, etc.

Tukey hasta hace muy poco ha seguido desempeñando el cargo de Profesor Emérito de Estadística, Profesor Emérito de Ciencia, y Alto Investigador Estadístico en la Universidad de Princeton

TÉCNICAS Y ESTRATEGIAS

Una técnica de recogida de datos: los diagramas de tallos y hojas

EL PROBLEMA

Las puntuaciones obtenidas por 40 alumnos en un test han sido las siguientes:

41 53, 72, 62, 81, 93, 81, 74, 56, 62, 45, 47, 62, 58, 88, 76, 77, 63, 43, 56, 76, 63, 78, 73, 65, 66, 91, 82, 61, 72, 36, 50, 91, 32, 60, 80, 51, 68, 61, 71

Hacer el recuento ordenado de los datos.

RESOLUCIÓN DEL PROBLEMA

Una nueva técnica de recuento de datos es mediante un **diagrama de tallos y hojas**, que se construye del siguiente modo:

Paso 1º

Se observa entre qué valores están las cifras de las decenas de todos los datos y se tiene que van de 3 a 9:

Tallo	
3	
4	
5	
6	
7	
8	
9	

Paso 2º

Se va leyendo uno a uno cada dato, anotando la cifra de las unidades de cada uno en la fila correspondiente:

Primer dato 41 Segundo dato 53

Tallo	Hojas	Tallo	Hojas
3		3	
4	1	4	1
5		5	3
6		6	
7		7	
8		8	
9		9	

Y así se continúa con los restantes datos.

Paso 3º

Cuando todos los datos han sido anotados se obtiene una tabla como ésta:

Tallo	Hojas
3	6 2
4	1 5 7 3
5	3 6 8 6 0 1
6	2 2 2 3 3 5 6 1 0 8 1
7	2 4 6 7 6 8 3 2 1
8	1 1 8 2 0
9	3 1 1

Paso 4º

Se vuelve a escribir la tabla ordenando de menor a mayor las unidades dentro de cada fila:

Tallo	Hojas
3	2 6
4	1 3 5 7
5	0 1 3 6 6 8
6	0 1 1 2 2 2 3 3 5 6 8
7	1 2 2 3 4 6 6 7 8
8	0 1 1 2 8
9	1 1 3

Esto es un diagrama de tallos y hojas.

Los diagramas de tallos y hojas son, en sí mismos, diagramas de frecuencias, pues si unimos los últimos números de cada fila obtenemos el polígono de frecuencias.

Podemos ver que hay 2 alumnos con puntuaciones entre 30 y 39, 4 alumnos con puntuaciones entre 40 y 49,

y así sucesivamente. Hay más alumnos con puntuaciones entre 70 y 79 que entre 50 y 59¹.

La clase con mayor frecuencia es la que tiene extremos 60-69.

¹ Esta es una reproducción de la página 255 del libro MATEMÁTICAS de 3º de E.S.O. de Ediciones S.M. Autores: J.R. Vizmanos Y M. Anzola.

ALGO MAS SOBRE DIAGRAMAS DE TALLOS Y HOJAS

Si se desea comparar dos series de datos que tengan los tallos comunes se pueden *situar enfrentados del siguiente modo*:

Las tal/as, en centímetros, de un grupo de 20 alumnos de un curso son las siguientes: 146, 152, 160, 172, 180, 175, 181, 160, 167, 190, 163, 153, 158, 162, 159, 155, 177, 167, 161, 158.

Las tallas, en centímetros de un grupo de 17 alumnas de un curso son las siguientes: 147, 140, 155, 150, 142, 165, 159, 152, 156, 163, 171, 160, 163, 150, 158, 141, 143.

Representar el correspondiente diagrama de tallos y hojas de ambas distribuciones

alumnas (n=17)alumnos (n=20)

73210146

986520015235889

53301600123477

117157

1801

190

La simple visión del gráfico nos da una idea de como se comportan ambas distribuciones, además es posible comparar algunos parámetros, como las medianas, los cuartiles etc. sin necesidad de realizar pesados cálculos, aprovechando el hecho de que los datos se encuentran ordenados



Esta foto de Tukey a sus 30 años apareció en el *Beil Labs News* en 1 985, con el siguiente pie:



En esta foto aparece **Tukey** junto a su amigo y colaborador de toda la vida **Frederick Mosteller** (1916-) y la esposa de este, **Virginia Girloy Mosteller**.

UTILIZA LA TECNOLOGIA ACTUAL

Hoy en día existen calculadoras gráficas de bolsillo que representan los diagramas de caías y bigotes sólo con introducir los datos de la distribución

LOS DIAGRAMAS DE CAJAS Y BIGOTES²

Los **diagramas de cajas y bigotes** —también se llaman *boxplots* o *box and whiskers*— son representaciones gráficas de una distribución estadística unidimensional en las que se reflejan cinco parámetros: límite inferior, primer cuartil, mediana, tercer cuartil y límite superior. A partir de estos cinco parámetros se pueden obtener fácilmente otros dos parámetros: el rango y el rango intercuartílico. Además,

² Esta es una reproducción de la página 355 del nuevo libro ALGORITMO MATEMÁTICAS II DE COU de Ediciones SM. Autores: J.R. Vizmanos y M. Anzola.

también dan una medida de la simetría o asimetría de la distribución, del sesgo y de la dispersión.

Veamos un ejemplo.

Las calificaciones finales de una clase han sido las siguientes:

Calificaciones	3	4	5	6	7	8	9
Nº de alumnos	1	4	7	3	3	0	2

El límite inferior es: $L_1 = 3$

El límite superior es: $L_s = 9$

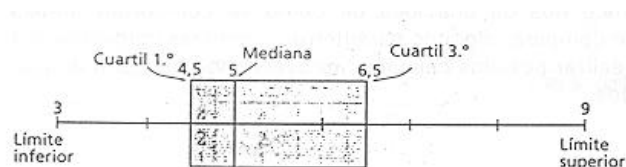
Si efectuásemos los cálculos obtendríamos que:

El cuartil primero es: $Q_1 = 4,5$

La mediana es: $M = 5$

El cuartil tercero es: $Q_3 = 6,5$

Por tanto el diagrama de cajas y bigotes será:



CON LA CALCULADORA GRAFICA

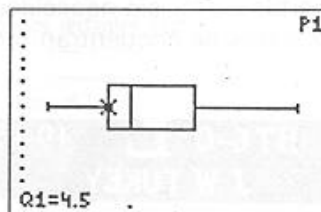
Introducimos los datos de la variable y la frecuencia absoluta en las listas L y L2.

L1	L2	L3
3	1	-----
4	4	-
5	7	
6	3	
7	3	
8	0	
9	2	
L1(1)=3		

A la vista del diagrama

a de cajas y bigotes, se observa que:

1. El bigote de la izquierda es algo más corto que el bigote derecha, lo que indica que las calificaciones de la parte más baja de la clase están algo más concentradas que las calificaciones de la cuarta parte calificaciones más altas.
2. También se observa que la parte izquierda de la caja, corresponde a los alumnos que han obtenido calificaciones entre el 25% y el 50%, es menor que la derecha, lo que indica que las calificaciones de estos últimos están mas dispersas.
3. Es fácil ver que el rango es $L_s - L_1 = 9 - 3 = 6$ y que el rango intercuartílico es: $Q_3 - Q_1 = 6,5 - 4,5 = 2$. También se observa que la distribución es asimétrica y ligeramente sesgada a la derecha.



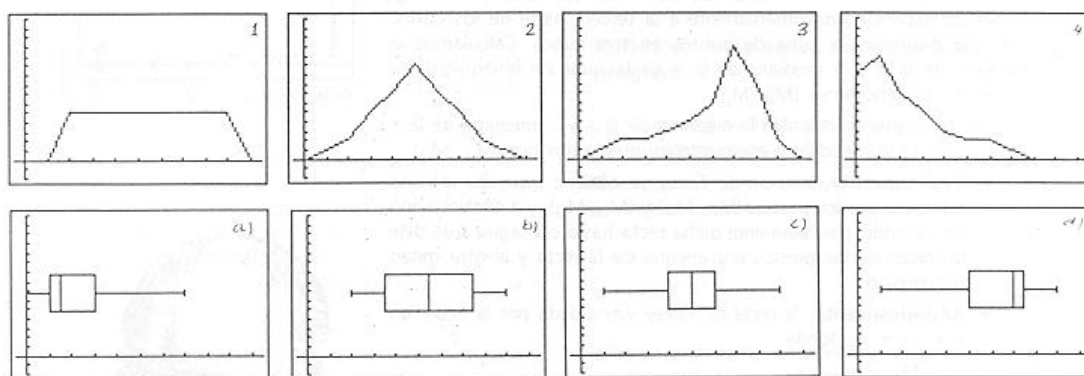
de la cuarta con que de la

ALGO MAS SOBRE LOS DIAGRAMAS DE CAJAS Y BIGOTES

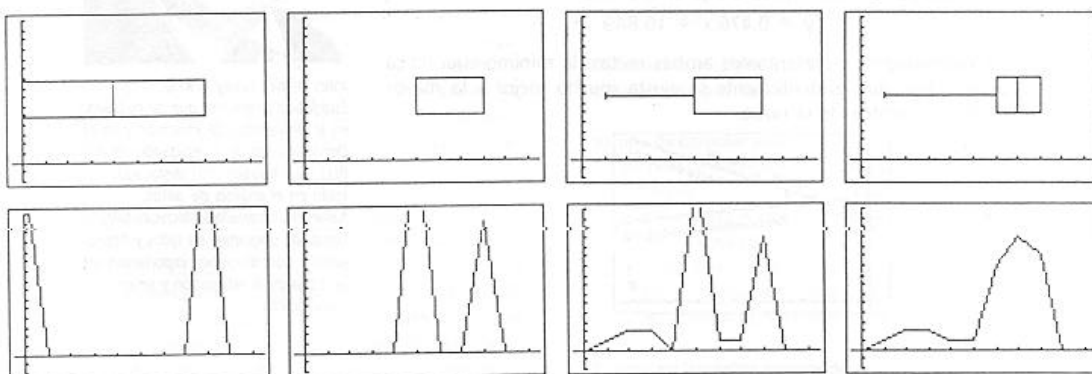
Las calificaciones dadas por 4 profesores a un grupo de 40 han sido las siguientes:

Calificaciones	0	1	2	3	4	5	6	7	8	9	10
Nº alumn. Prof. A			5	5	5	5	5	5	5	5	
Nº alumn. Prof. B		1	2	5	7	10	7	5	2	1	
Nº alumn. Prof. C		1	2	2	3	4	6	12	8	2	
Nº alumn. Prof. D	9	11	7	5	3	2	2	1			

A continuación se han representado los polígonos de frecuencias y los diagramas de cajas y bigotes de estas distribuciones. Indíquese la correspondencia entre unos y otros.



Por último representamos algunos diagramas de cajas y bigotes con alguna patología y debajo los polígonos



de frecuencias asociados. _____

LA RECTA DE TUKEY³

Consideremos la distribución bidimensional dada por la tabla del margen. Con una calculadora con tratamiento bidimensional obtenemos la ecuación de la recta en regresión de Y sobre X, que es la siguiente:

$$y = -0,161x + 13,569$$

En el margen hemos representado la nube de puntos y la recta de regresión. Es obvio que debido a dos puntos patológicos la recta obtenida se ajusta muy mal a la nube.

La regresión que hemos estudiado está basada en las medias, y éstas son muy sensibles a los valores muy extremos. En cambio, la mediana es un parámetro mucho más inerte a estos valores extremos.

El estadístico americano **John Wilder Tukey** ha desarrollado un procedimiento que se basa en esta idea eliminando en gran medida la influencia de los valores muy extremos.

Se divide el eje X en tres partes de modo que cada una de las partes corresponde aproximadamente a la tercera parte de los datos. Con ello dividimos la nube de puntos en tres nubes. Calculamos la mediana de la X y la mediana de la Y de la nube de la izquierda; a ese punto lo denotamos (M_{x1} , M_{y1}).

Análogamente calculamos la mediana de la X y la mediana de la Y de la nube de la derecha; a ese punto lo denotamos por (M_{x3} , M_{y3}).

• **Gráficamente**, la recta de Tukey se obtiene trazando la recta que pasa por los puntos (M_{x1} , M_{y1}) y (M_{x3} , M_{y3}), y a continuación trasladando paralelamente dicha recta hasta conseguir que deje la mitad de los puntos por encima de la recta y la otra mitad por debajo.

• **Analíticamente**, la recta de Tukey viene dada por la ecuación $y = ax + b$, donde

$$a = \frac{M_{y3} - M_{y1}}{M_{x3} - M_{x1}}, \quad b = \text{mediana de los valores } (y_i - ax).$$

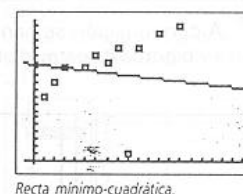
Hoy en día existen calculadoras gráficas que ya tienen incorporado un procedimiento para obtener la ecuación de la recta de Tukey, que también se llama *recta mediana-mediana* o *recta resistente*.

La recta así obtenida tiene la siguiente ecuación:

$$y = 0,476x + 10,849$$

A continuación representamos ambas rectas: la mínimo-cuadrática y la de Tukey, que evidentemente se ajusta mucho mejor a la mayor

x	y
1	9
2	11
3	13
5	13
6	15
7	14
8	16
9	1
10	16
12	18
14	19
21	2



John Wilder Tukey (1915). Estadístico americano que trabajó en Universidad de Prince y en el Departamento Investigación de ABEU. Sus trabajos más destacados están en análisis de datos habiendo desarrollado técnicas tales como diagramas de tallos y hojas, junto con contribuciones importantes en la técnica

