Capítulo 4:
ANÁLISIS DE LA RELACIÓN ENTRE VARIABLES.

4.1.- INTRODUCCIÓN Y CONCEPTOS.

En la investigación empírica, y especialmente en las Ciencias Sociales, lo habitual no es tanto analizar una variable por separado, tal y como se ha mostrado en los procedimientos analizados en el Capítulo 2, sino estudiar diversas variables conjuntamente, lo que proporciona una visión de conjunto y un mayor acercamiento a la compleja "realidad" bajo estudio, si bien, los procedimientos vistos, univariados, son paso previo para abordar el estudio y el empleo de los que veremos ahora.

A efectos didácticos, en este Capítulo nos ceñiremos únicamente al estudio de la relación entre dos variables, desde la perspectiva del A.E.D., en aras de una mayor simplicidad que facilite su análisis, confiando en que el lector extienda lo expuesto a la relación entre más de dos variables. Comenzaremos con un ejemplo.

Pensemos que queremos estudiar la relación existente entre dos variables, concretamente la altura y la masa en la población de adultos varones. Para ello, se elige una muestra de N sujetos representativos de dicha población. De cada uno de los N sujetos obtendremos tanto su altura como su masa (que, en adelante consideraremos como "peso"), con lo que se poseerá una tabla como la siguiente (supondremos, para simplificar, que N=12), siendo X la variable altura (medida en cm) e Y la variable "peso" (medida en kg)[1]:

Nº sujeto	X	Y
1	170	68
2	181	78
3	166	67
4	188	90
5	183	81
6	177	73
7	166	95
8	182	87
9	172	70
10	177	75
11	178	77
12	185	83

[1] Debido a que este ejemplo es un estudio correlacional, se podría haber asignado igualmente X al peso y dejar Y para la altura.

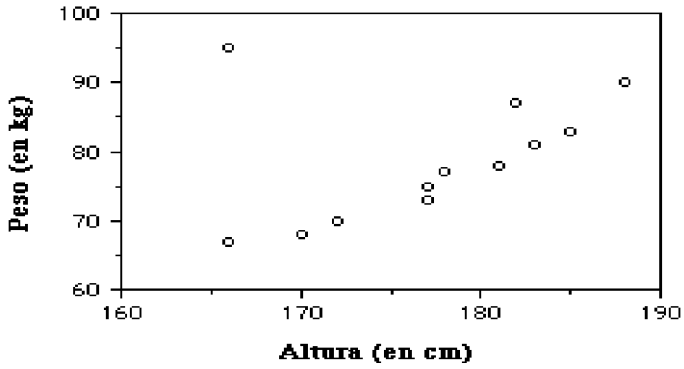
Evidentemente, de esta tabla se deduce, por ejemplo, que el sujeto 9 mide 172 cm y pesa 70 kg.

4.1.1.- REPRESENTACIÓN DE LA RELACIÓN: EL DIAGRAMA DE DISPERSIÓN (*Scatter Plot*).

Para la representación de la relación entre dos variables se suele emplear el llamado diagrama de dispersión (*scatter plot*) que consiste en la representación gráfica, en forma de puntos, de los pares $(X_i;Y_i)$ en un sistema de coordenadas cartesianas, en cuyo eje de abscisas se situará la variable X, mientras que en el eje de ordenadas se situará la variable Y. Así, la primera puntuación del ejemplo es el par (170;68).

De acuerdo con los datos proporcionados en el ejemplo del anterior apartado, el diagrama de dispersión entre las variables X e Y es el siguiente:

GRÁFICA Nº1.



Una vez se ha realizado el diagrama se podrán observar una serie de características del mismo, las más importantes de las cuales, a saber, la fuerza, la dirección y la forma de la relación se hallan expuestas en el próximo apartado.

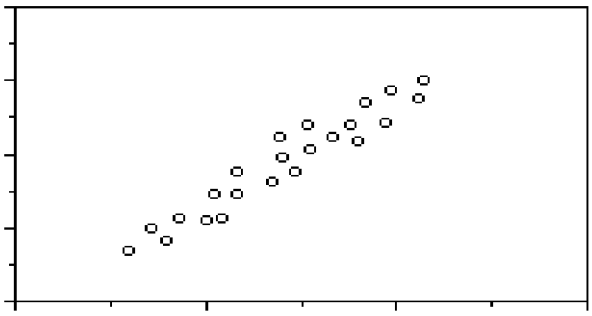
4.1.2.- FUERZA, DIRECCIÓN Y FORMA.

En el Capítulo primero se destacó la importancia que tenían diversos factores para el análisis de la distribución de una variable. En esta línea se señaló la localización, dispersión y la forma de la distribución de los datos.

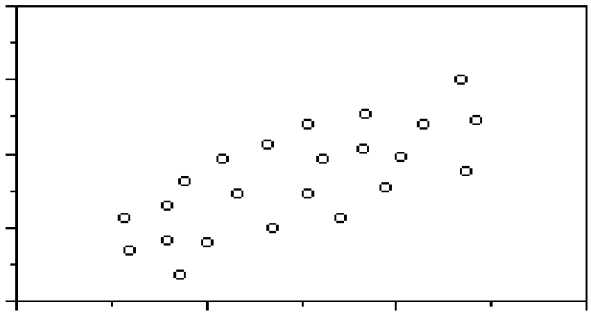
Naturalmente, cuando tenemos dos variables se puede realizar dos veces el procedimiento efectuado en el primer Capítulo, es decir analizar cada variable por separado. Pero, por otra parte, lo más interesante en estos casos suele ser el análisis de la relación entre ambas variables. Tal relación tiene tres aspectos fundamentales (aunque nos referiremos a la relación entre dos variables, tales aspectos se pueden hacer extensivos a la relación de tres ó más variables):

I) *LA FUERZA DE LA RELACIÓN:* Se refiere al grado de ajuste de los datos a la hipotética recta/curva que los enlazaría. En términos del Análisis Exploratorio de Datos sería "la relativa importancia del ajuste respecto al residual" (HARTWIG Y DEARING, 1979, p. 31). En las gráficas siguientes se aprecia, en la primera, un buen ajuste de los datos a la recta que sirve como modelo, es decir, bastante fuerza en la relación. Mientras, en la segunda se observa un menor ajuste a la recta, es decir, una menor fuerza en la relación.

GRÁFICA Nº2.



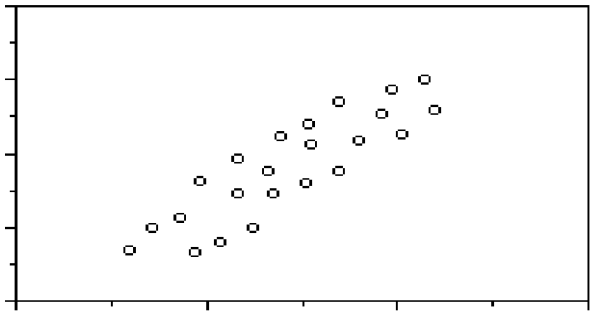
GRÁFICA Nº3.



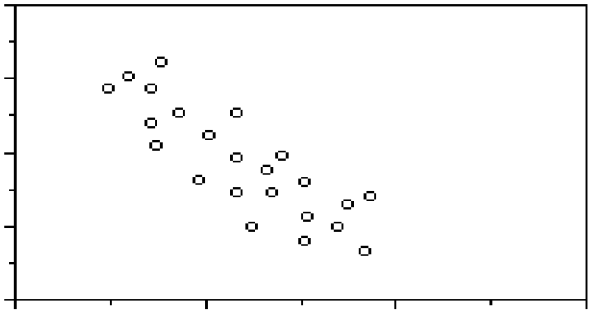
II) ***DIRECCIÓN DE LA RELACIÓN:*** Se refiere, en los casos de **relación lineal** entre variables, al signo de la pendiente de la recta que se ajusta a los datos. Es decir, si a un mayor valor de una variable suele acaecer un mayor valor en la otra variable, la pendiente de la recta será positiva y se dice que la dirección de la relación es positiva o **directa**. Mientras, si a un mayor valor de una variable suele acaecer un menor valor en la otra variable, la pendiente será negativa, con lo que se dice que la dirección es negativa o **inversa**. En el caso de que la relación no sea lineal, podrá haber tanto una dirección positiva como una negativa (piénsese en el caso de una parábola, o véase, por ejemplo, el segundo de los gráficos (nº 7) que se incluyen para ilustrar la "forma de la distribución").

En las dos gráficas que siguen se muestra, en la primera una relación directa y en la segunda una relación inversa.

GRÁFICA Nº4.

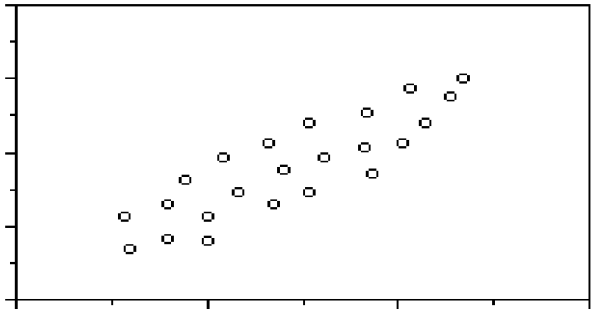


GRÁFICA Nº5.

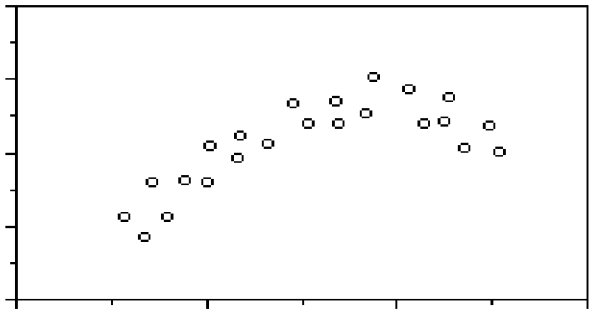


III) *FORMA DE LA DISTRIBUCIÓN:* Se refiere al modo en que ambas variables están relacionadas. Básicamente, la relación entre las variables podrá ser lineal o bien podrá ser curvilínea. En los gráficos siguientes se expone una relación lineal y una curvilínea.

GRÁFICA Nº6.



GRÁFICA Nº7.



De acuerdo con las tres características que acabamos de explicar, de la observación del diagrama de dispersión de nuestro ejemplo (véase Apartado 4.1.1. -gráfica nº 1-) podemos concluir que (a) parece haber bastante fuerza en la relación (salvo la existencia de una puntuación atípica), (b) que la relación es lineal (los puntos parecen acomodarse -o seguir- a una recta), y (c) que dicha relación es directa (a más altura más peso y viceversa).

En los apartados que siguen se analizará únicamente la relación lineal entre variables, reservando la relación no-lineal para los Apartados 4.3.3. y 4.3.6.

4.1.3.- EL MÉTODO DE LOS "MÍNIMOS CUADRADOS".

El procedimiento de Mínimos Cuadrados es un procedimiento "clásico" muy empleado, que permite estudiar la relación lineal entre variables, mediante una ecuación (lineal) o "recta" de regresión.

Supongamos que nos interesa ajustar una recta al diagrama de dispersión mostrado en el Apartado 4.1.1. para así poder obtener una relación matemática simple que indique, con cierta exactitud, el peso de un sujeto (su valor en la variable Y) como función de su altura (su valor en la variable X). En otras palabras, se desea conocer con cierta precisión el peso de una persona que no esté en la muestra (pero que sea de la misma población) a partir de su altura (habiendo comprobado previamente la relación existente entre ambas variables con una muestra representativa de la población a la que se supone pertenece dicha nueva persona). El método de Mínimos Cuadrados sirve para ajustar una recta a una serie de puntos que guarden una relación lineal. El criterio de este método es el siguiente: **La recta se ha de construir de tal manera que la suma de los cuadrados de las desviaciones verticales de todos los puntos respecto a la recta sea mínima.** Tales desviaciones verticales son los ***residuales***, esto es, lo que no se ha podido ajustar con la recta (véase el gráfico nº 8 para comprobar las desviaciones verticales de la recta). En la terminología del A.E.D., los residuales son la "parte residual" o exclusiva, mientras que la recta nos proporciona la "parte ajustada" o tendencia (aspectos a los que se aludió en el Capítulo 1).

La pendiente y la ordenada en el origen de la recta $Y=a+bX$ que estima este procedimiento se ofrecen seguidamente.

La pendiente es,

$$b = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}$$

mientras que la ordenada en el origen es,

$$a = \bar{Y} - b\bar{X}$$

donde $\bar{X}$ representa la Media de X, e $\bar{Y}$ la media de Y.

Sin embargo, este procedimiento es poco resistente, ya que, al tener en cuenta todas las puntuaciones (se puede observar la utilización de las Medias de X y de Y en las fórmulas anteriores), puede resultar alterado fácilmente por la existencia de unas pocas puntuaciones atípicas.

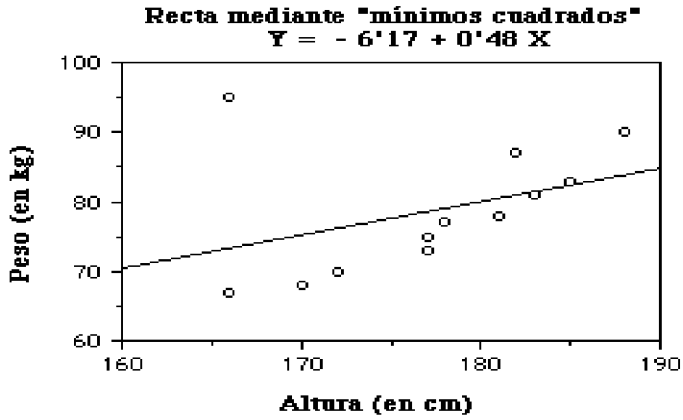
En nuestro ejemplo, se obtienen los datos siguientes (en puntuaciones directas):

La pendiente es b=0'48, mientras que la ordenada en el origen es a=-6'17. Consiguientemente, con estas dos constantes, origen y pendiente, se obtiene la siguiente ecuación $\hat{Y}$=-6'17+0'48X, que nos indica que cuando (por ejemplo, en un sujeto) el valor de X sea "0" el valor estimado[2] de Y será "-6'17". Así, para el valor más bajo de X, "166", el valor estimado correspondiente de Y será -6'17+0'48*166="73'51",....., y al valor más alto de X, "188", corresponderá un valor de Y estimado de -6'17+0'48*188="84'07"[3]. La ecuación anterior obtenida, trasladada al diagrama de dispersión nos dará la representación de la recta de Regresión de *Y sobre X* por el criterio de Mínimos Cuadrados, que se ofrece en la gráfica siguiente, en la que el lector podrá comprobar los valores estimados de Y para cada valor de X (entre ellos los dos valores de Y que acaban de ser obtenidos mediante este procedimiento de estimación):

[2] En la ecuación anterior , se ha utilizado para la ordenada el símbolo $\hat{Y}$ en lugar de Y para señalar que se trata de una estimación (no el valor real de Y). En Psicología también es frecuente utilizar Y' para referirse a la puntuación estimada.

[3] Puede comprobarse la diferencia entre los valores de Y predichos, mediante este procedimiento (o ecuación), y los valores de Y reales, para los correspondiente valores de X, que figuran al comienzo del apartado 4.1., y en la gráfica siguiente representados por puntos. Especialmente en el caso del valor de X "166" que aparece dos veces con diferentes valores de Y asociados ("67" y "95") uno de ellos atípico ("95"), resultando para el valor "166" de X (en ambos casos) un valor estimado para Y de "73'51". Los residuales, que se calculan como las diferencias entre Y (valor medido) y el valor estimado para Y ($\hat{Y}$), serían para el ejemplo previo, de -6'51 y de 21'49 respectivamente.

GRÁFICA Nº8.



En el gráfico anterior, en el que se ha representado la ecuación obtenida, se observará que dicha línea no describe bien la supuesta recta que siguen la mayoría de los puntos, sino que se halla muy afectada por la existencia del par atípico (166;95). Ello se debe al empleo de un estadístico poco resistente como la Media, como se aprecia, por ejemplo, en la fórmula de la pendiente en este método. En consecuencia, el método de Mínimos Cuadrados es poco resistente.

4.1.4.- EL MÉTODO DE LA "LÍNEA DE TUKEY".

El método denominado "Línea de Tukey", también denominado "método de la línea resistente de tres grupos" (EMERSON Y HOAGLIN, 1983b), es un procedimiento resistente para observar la relación lineal entre dos variables. Para ello, emplea un estadístico resistente como la Mediana que, como ya ha sido analizado en los dos primeros capítulos, es un estadístico que se halla poco afectado por las puntuaciones extremas, ya que no tiene en

cuenta todas la puntuaciones, sino solamente las centrales, a diferencia de la Media Aritmética[4]. Por ello, el método de la Línea de Tukey es asimismo resistente. En los siguientes apartados se describe su cálculo.

4.2.- PROCEDIMIENTOS DE CÁLCULO DE LA "LÍNEA DE TUKEY".

La Línea de Tukey puede ser calculada bien a través de un procedimiento gráfico, sobre el diagrama de dispersión, bien a través de un procedimiento numérico, aspectos de los que nos ocupamos en los dos siguientes apartados.

En ambos casos, se precisan estos pasos previos:

I) Ordenar los N pares de valores en función de la variable X, en orden ascendente. En el caso del ejemplo, tendremos:

Antes de la ordenación		
Nº sujeto	X	Y
1	170	68
2	181	78
3	166	67
4	188	90
5	183	81
6	177	73
7	166	95
8	182	87
9	172	70
10	177	75
11	178	77
12	185	83

Después de la ordenación			
Nº sujeto	X	Y	Orden
3	166	67	(1)
7	166	95	(2)
1	170	68	(3)
9	172	70	(4)
6	177	73	(5)
10	177	75	(6)
11	178	77	(7)
2	181	78	(8)
8	182	87	(9)
5	183	81	(10)
12	185	83	(11)
4	188	90	(12)

Consiguientemente, $X_{(12)}=188$, ó $Y_{(8)}=78$.

[4] Aquí no tomaremos la segunda acepción de resistencia, referida a los problemas de agrupamiento o redondeo, sino solamente la primera, referida a la resistencia a las puntuaciones extremas.

II) Dividir las N observaciones ordenadas de X en tres grupos, con sus respectivos valores en Y, cada uno de los grupos con, aproximadamente, un tercio de las observaciones. De este modo, se obtendrá un grupo inferior (el primer tercio de puntuaciones de X), un grupo medio (el segundo tercio de puntuaciones) y un grupo superior (con las N/3 puntuaciones mayores en X).

Cabe hablar de tres posibilidades:

a) En el caso de que N=3k (donde k es un número entero), es decir que N sea múltiplo de 3 (de una forma práctica, si la división N/3 no da decimales), cada grupo podrá tener el mismo número de observaciones (k observaciones en cada grupo, siendo k=N/3).

b) En el caso de que N=3k+1 (o si la división N/3 da como resultado decimal: k'$\hat{3}$), habrá k observaciones en los grupos inferior y superior, y k+1 observaciones en el grupo medio.

c) En el caso de que N=3k+2 (o si la división N/3 da como resultado decimal: k'$\hat{6}$), habrá k observaciones en el grupo medio, mientras que habrá k+1 observaciones en los grupos inferior y superior.

Sin embargo, este paso puede plantear algunos problemas. Concretamente, nos referiremos a la existencia de una restricción que señala que dos puntuaciones de X con el mismo valor, es decir, los empates, deben estar en el mismo grupo. Por ejemplo, en el conjunto ordenado de puntuaciones de la variable X: 3,3,3,4,4,4,4,5,5; habrá tres grupos, el primero formado por los tres primeros, un segundo formado por los cuatro siguientes (y no sólo por los tres siguientes, para no separar las puntuaciones iguales), mientras que el grupo superior estará formado por las dos últimas puntuaciones. Por ello, líneas arriba se empleó el término de *"aproximadamente" un tercio de las puntuaciones*, no sólo por el hecho de que la muestra no sea múltiplo de 3, sino también porque los casos de empate pueden plantear problemas. En cualquier caso, no nos extenderemos en este punto que debe estar ya solucionado en los programas informáticos que calculen la Línea de Tukey.

Con los datos del ejemplo anterior, se obtienen los siguientes grupos, cada uno de ellos con cuatro pares de puntuaciones (en nuestro caso, no ha habido problemas de empates en las puntuaciones limítrofes inter-grupos):

	X	Y
Grupo inferior o primero	166	67
	166	95
	170	68
	172	70
Grupo medio o segundo	177	73
	177	75
	178	77
	181	78
Grupo superior o tercero	182	87
	183	81
	185	83
	188	90

III) Averiguar la Mediana de la variable X en el primer grupo (que denotaremos MdX_1), así como la de Y también en el primer grupo (MdY_1). En el ejemplo,

$MdX_1=168$
$MdY_1=69$

IV) Averiguar la Mediana de la variable X (MdX_2), así como también la de Y (MdY_2), en el segundo grupo. En el ejemplo,

$MdX_2=177'5$
$MdY_2=76$

V) Averiguar la Mediana de la variable X (MdX_3), así como también la de Y (MdY_3), en el tercer grupo. En el ejemplo,

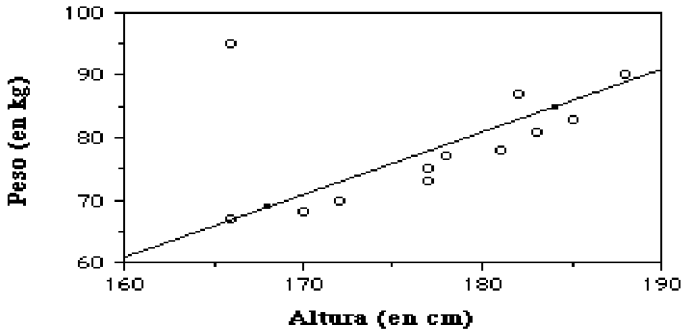
$MdX_3=184$
$MdY_3=85$

4.2.1.- PROCEDIMIENTO GRÁFICO.

Para efectuar el procedimiento gráfico, hay que realizar primeramente los pasos efectuados en el anterior apartado, así como, también, un diagrama de dispersión entre ambas variables (junto a la observación de que los puntos sigan una relación aproximadamente lineal). Una vez realizados, se efectúan los siguientes pasos, siguiendo la recomendación de Hartwig y Dearing (1979):

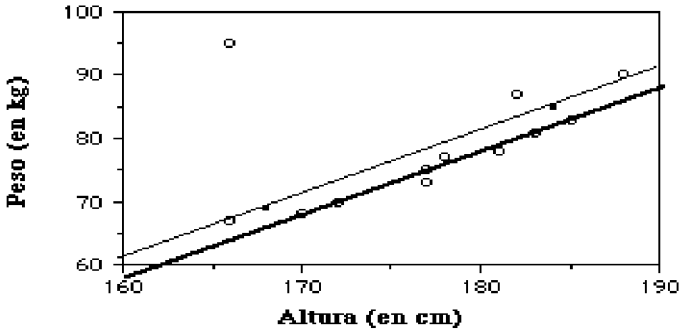
III) Unir en el diagrama de dispersión, mediante una recta a modo de línea "imaginaria", los dos puntos indicados por las coordenadas (MdX_1 ; MdY_1) y (MdX_3 ; MdY_3). Esta línea dividirá en dos a los pares de puntuaciones o casos: unos casos estarán sobre tal línea, mientras que otros casos estarán bajo tal línea. En nuestro ejemplo, las coordenadas serán (168;69) y (184;85), que son los puntos marcados en negro en la gráfica siguiente, por los que se ha trazado, o pasa, la línea ("imaginaria") que se verá.

GRÁFICA Nº9.



IV) Trazar una línea paralela a la anterior hasta que queden la mitad de los casos por encima de la misma y la otra mitad por debajo. Tal línea es la denominada Línea de Tukey. En el ejemplo, esta línea, más destacada en negro, resultará como sigue,

GRÁFICA Nº10.



Naturalmente, en el caso de que el número de observaciones sea elevado, resulta muy complejo realizar el anterior proceso manualmente, por lo que en tal caso se recomienda efectuarlo a partir de las fórmulas matemáticas. Además, el proceso puede dar lugar a diversas líneas paralelas (es decir, todas con la misma pendiente) que cumplan el paso IV, es decir que dejen por encima y por debajo el 50% de los casos (de forma análoga a la existencia de muchos valores que cumplen la definición de Mediana en diversas distribuciones), con lo que habrá ciertas diferencias respecto al método matemático, que a continuación se describe, en el término que corresponde a la ordenada en el origen.

4.2.2.- PROCEDIMIENTO MATEMÁTICO.

Como fase previa, se han de efectuar los pasos indicados en el Apartado 4.2. Seguidamente, al igual que en el procedimiento de Mínimos Cuadrados, como lo que hay que calcular es una recta, el procedimiento matemático calcula la pendiente y la ordenada en el origen. Es decir, si la ecuación de una recta se define por $Y=a+bX$, este procedimiento calcula tanto a (la ordenada en el origen) como b (la pendiente).

El cálculo de la **pendiente** en la Línea de Tukey se muestra seguidamente (que coincide con la pendiente resultante del método gráfico):

$$b = \frac{MdY_3 - MdY_1}{MdX_3 - MdX_1}$$

mientras, el valor de la ordenada en el **origen** es el siguiente[5],

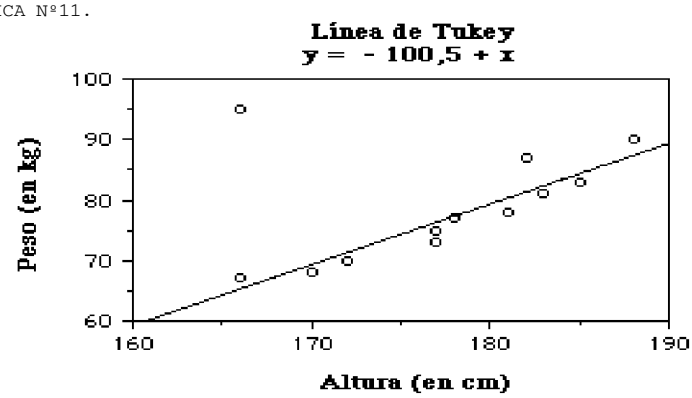
$$a = \frac{\sum_{i=1}^{3}(MdY_i - b*MdX_i)}{3}$$

En el ejemplo, se obtienen los siguientes resultados:

b = (85 - 69)/(184 - 168) = 1
a = (1/3) (69 - 1*168 + 76 - 1*177'5 + 83 - 1*184) = -100'5

En consecuencia, la Línea de Tukey expresada numéricamente será $\hat{Y} = X - 100'5$, muy parecida a la obtenida en el método gráfico (aunque varía, como ya se advirtió, la ordenada en el origen), como se observa en la siguiente gráfica:

GRÁFICA Nº11.



[5] Diversos autores (por ejemplo, HARTWIG y DEARING, 1979) señalan la utilización para la ordenada en el origen no de la Media Aritmética entre los tres valores que se expresan en la fórmula de la ordenada de origen, sino de la Mediana.

Lo más importante es destacar que el par atípico (166;95) no ha influido sobre la relación entre X e Y, esto es, la Línea de Tukey se ha mostrado resistente a la existencia de tal par atípico.

4.2.3.- COMPARACIÓN DEL MÉTODO DE "MÍNIMOS CUADRADOS" CON EL DE LA "LÍNEA DE TUKEY".

Se puede apreciar a través del ejemplo mostrado que la Línea de Tukey, al emplear la Mediana en lugar de la Media, se halla menos afectada por las puntuaciones atípicas que la recta calculada mediante el procedimiento de Mínimos Cuadrados, es decir, aquélla es más resistente que ésta.

En cualquier caso, hay que recordar que si la muestra es amplia y no hay puntuaciones atípicas, el procedimiento de Mínimos Cuadrados es generalmente adecuado. Sin embargo, en los casos en que tales supuestos no ocurran, algo bastante frecuente, es mejor emplear un procedimiento más resistente, como lo es un procedimiento basado en las Medianas, como es la Línea de Tukey. Un procedimiento alternativo podría ser la eliminación de las puntuaciones extremas de acuerdo con algún criterio y, seguidamente, aplicar el método de Mínimos Cuadrados. En el ejemplo anterior, si se eliminase el par atípico, la ecuación resultante de aplicar el método de Mínimos Cuadrados es:

$$\hat{Y} = 1'07\ X - 112'88$$

que es bastante parecida a la obtenida con la Línea de Tukey.

4.3.- OTROS MÉTODOS.

En los siguientes apartados se mostrará, por una parte, un método resistente alternativo al de la Línea de Tukey, desarrollado sobre éste (Apartado 4.3.1.), así como algunos procedimientos para el *análisis de residuos* (Apartado 4.3.2.); y, por otra parte, un método de Líneas de Regresión sobre Medianas y Cuartos (Apartado 4.3.3.) útil en el caso de que haya pocos valores distintos en la variable X, y que resulta válido incluso para relaciones no-lineales (curvilíneas) entre variables.

Seguidamente, en los Apartados 4.3.4. y 4.3.5. se mostrarán procedimientos ya de uso muy común en el ámbito del análisis de series temporales y retomados por los estudiosos del enfoque del Análisis Exploratorio de Datos aunque, a diferencia de otros métodos "novedosos" del A.E.D. (como la Línea de Tukey, el diagrama de Tallo-y-Hojas, entre otros), aquéllos ya han tenido una amplia difusión previamente a su tratamiento por los autores de este enfoque, si bien han sido retomados por ellos. Son procedimientos que proporcionan una suavización de los datos, disminuyendo la influencia de los valores atípicos, por lo que se justifica su adopción por este enfoque y su tratamiento en este texto. Finalmente, el Apartado 4.3.6. señala los procedimientos de transformación de datos para el tratamiento de las relaciones no lineales.

4.3.1.- VARIACIONES SOBRE EL MÉTODO DE LA LÍNEA DE TUKEY: EL MÉTODO ITERATIVO.

Habitualmente, a la hora de expresar una recta se suele utilizar tanto la pendiente como la ordenada en el origen. Sin embargo, en mucho casos, lo importante no es tanto la ordenada en el origen cuanto la pendiente, ya que, retomando a Emerson y Hoaglin (1983b), en nuestro ejemplo, ¿qué significa la ordenada

en el origen para alguien cuya altura fuera "0"?. Por ello, se han determinado ligeras variaciones respecto al método de la Línea de Tukey a través de un procedimiento de tipo iterativo en el que a partir de la pendiente de la recta y de un denominado "valor central" se llega a una expresión más realista de la ecuación de la recta. En cualquier caso, los resultados son similares a los que se obtienen con la Línea de Tukey. La principal diferencia estriba en el tipo de análisis que se hace, ya que en el procedimiento iterativo se dedica una mayor atención a los residuales.

Antes de continuar, señalaremos que un método o procedimiento iterativo es aquel que a través de una serie de pasos repetitivos converge hacia una solución. Por ejemplo, si deseamos conocer la raíz cuadrada de un número, un proceso iterativo es el siguiente:

$$X_i = \frac{X_{i-1} + \frac{a}{X_{i-1}}}{2} \qquad i=1,2,...$$

donde **a** es el número del cual se desea averiguar su raíz cuadrada y X_i es la solución aproximada de la i-ésima iteración. (El valor inicial de i es 1, luego para dicho valor de i, i-1 será 0, es decir $X_{i-1} = X_0$).

Se comienza (valor inicial de i igual a 1) por asignar a X_0 un valor más o menos aproximado a la solución que se prevé.

Para ilustrarlo, pensemos, por ejemplo, que queremos calcular la raíz cuadrada de 5, y que una primera aproximación que proponemos es 2 (que será, pues, X_0). Entonces, sustituyendo en la fórmula general, $X_i = \frac{X_{i-1} + \frac{a}{X_{i-1}}}{2}$ i=1,2,...; para **a**=5 y comenzando con X_0=2, la primera iteración (X_1) da como resultado: $X_1 = \frac{2 + \frac{5}{2}}{2} = 2'25$,

mientras, las siguientes,

$X_2 = (X_1 + (5/X_1))/2 = (2'25 + (5/2'25))/2 =$
$2'361111$,

$X_3 = (X_2 + (5/X_2))/2 = (2'361111 + (5/2'361111))/2 = 2'2360679$,

$X_4 = (X_3 + (5/X_3))/2 = (2'2360679 + (5/2'2360679))/2 = 2'2360679$,...

Debido a que de la iteración tercera a la cuarta no ha habido prácticamente ningún cambio, y dado que la fórmula es convergente, se puede indicar que 2'2360679 es ya una buena aproximación a la raíz cuadrada de 5. De hecho, ese es el número que nos proporciona cualquier calculadora de 8 dígitos al efectuar tal operación directamente.

Retomando nuestro tema, Emerson y Hoaglin (1983b) destacan la utilidad del siguiente procedimiento iterativo para obtener la recta entre X e Y, que exponemos a continuación.

La línea inicialmente tendrá la forma:

$$\hat{Y} = a_0^* + b_0 (X - MdX_2),$$

donde, al igual que en la línea de Tukey,

$$b_0 = \frac{MdY_3 - MdY_1}{MdX_3 - MdX_1}$$

Mientras, a_0^* , que se llama en este procedimiento "valor central" o "nivel" (*level*), se calcula según la siguiente fórmula,

$$a_0^* = \frac{\left(MdY_1 - b_0 (MdX_1 - MdX_2)\right) + MdY_2 + \left(MdY_3 - b_0 (MdX_3 - MdX_2)\right)}{3}$$

El cálculo de los residuales (rs)[6] para cada punto se efectúa del modo siguiente:

$$rs_i = Y_i - \left(a^* + b(X_i - MdX_2)\right) \qquad i=1,2,...,N$$

donde: Y_i representa el i-ésimo valor de Y

[6] Se empleará "rs" como abreviatura de residual, para distinguirlo del Coeficiente de Correlación de Pearson.

X_i representa el i-ésimo valor de X.

Si la recta que se ajusta a la distribución de los residuales (las distancias verticales de la recta ajustada hasta la ordenada de cada par de datos) de Y sobre X tiene una pendiente de 0 (ó muy cercano), entonces el ajuste es adecuado. Sin embargo, lo habitual es que el ajuste tras la primera iteración no sea del todo perfecto. Consiguientemente, con los primeros residuales de Y sobre X tenemos,

$$rs_i^{(0)} = Y_i - \left(a_0^* + b_0 (X_i - MdX_2)\right) \quad i=1,2,...,N$$

donde: Y_i representa el i-ésimo valor de Y
X_i representa el i-ésimo valor de X.
en $rs_i^{(0)}$, el número entre paréntesis indica que es la primera iteración

Mientras, tras la segunda iteración se ajustarán tanto a_0^* como b_0, de modo que el nivel ajustado será ahora de $a_1^*=a_0^*+\gamma_1$, mientras que la pendiente ajustada será de $b_1=b_0+\delta_1$, donde δ_1 es la pendiente de la recta de los residuales de Y sobre X, mientras que γ_1 es el valor central (nivel) de dicha recta. Es decir, en lugar de analizar la relación de Y sobre X, ahora se toma la relación de los residuales de Y sobre X.

Consecuentemente, en la iteración siguiente, los nuevos residuales son,

$$rs_i^{(1)} = rs_i^{(0)} - \left(\gamma_1 + \delta_1 (X_i - MdX_2)\right) \quad i=1,2,...,N$$

Mientras, en las siguiente iteraciones, habrá que reajustar la pendiente y la ordenada en el origen de modo que:

$$b_2 = b_1 + \delta_2, \quad b_3 = b_2 + \delta_3, \ldots.$$

y, análogamente,

$$a_2^* = a_1^* + \gamma_2, \quad a_3^* = a_2^* + \gamma_3, \ldots.$$

Se podrá seguir haciendo más iteraciones hasta que el cociente entre el valor de la pendiente ajustada en la j-ésima iteración y el de la pendiente inicial (b_0) sea menor que una

determinada cantidad, tal como por ejemplo, el 1% ó de manera más conservadora, el 0'01% (EMERSON Y HOAGLIN, 1983b), con lo que el proceso habrá concluido. Naturalmente, no es aconsejable realizar este proceso a mano, debido a lo tedioso del proceso, sino mediante un programa de ordenador.

Veamos este desarrollo a través del ejemplo que ha venido exponiéndose a lo largo del capítulo, que podrá aclarar mejor tal proceso.

Sabemos que,

$MdX_1=168$
$MdY_1=69$

$MdX_2=177'5$
$MdY_2=76$

$MdX_3=184$
$MdY_3=85$

La pendiente inicial es:

$$b_0 = (85-69)/(184-168) = 1$$

y el valor central (o nivel) inicial es:

$$a_0^* = \frac{69+(1\ (168-177'5))+76+85+(1\ (187-177'5))}{3} = 75'667$$

Con lo que, primeramente, $\hat{Y} = 75'667 + (X-177'5)$

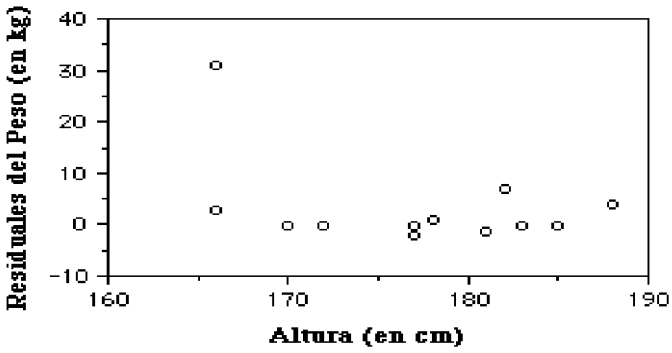
Veamos ahora cómo quedan los residuales, en los que se aprecia el elevado valor residual del par (166;95), es decir, del par atípico (que, concretamente, es de 30'833, como se puede apreciar con la lectura de la siguiente tabla):

X	Y	rs = Y - (75'667 + (X - 177'5))
166	67	2'833
166	95	30'833
170	68	-0'167
172	70	-0'167
177	73	-2'167
177	75	-0'167
178	77	0'833

181	78	-1'167
182	87	6'833
183	81	-0'167
185	83	-0'167
188	90	3'833

Mientras, la gráfica de los residuales de Y sobre X (obsérvese cómo las puntuaciones están centradas sobre el valor "0" en las ordenadas, salvo el residual correspondiente al par atípico que había en el conjunto de datos) es la siguiente,

GRÁFICA Nº12.



Calculando la primera iteración, en la que se hace la regresión de los residuales de Y (tercera columna) sobre X, tenemos las siguientes medianas en los tres grupos,

MdX_1=168
$Mdrs_1$=1'333

MdX_2=177'5
$Mdrs_2$=-0'667

MdX_3=184
$Mdrs_3$=1'833

como pendiente, resultará,

$$\delta_1 = \frac{1'833 - 1'333}{184 - 168} = 0'03125$$

y como valor central,

$$\gamma_1 = \frac{1'33 + 0'03(168 - 177'5) + (-0'67) + 1'83 + 0'03(184 - 177'5)}{3} = 0'8$$

Este proceso habrá de repetirse hasta que la razón entre la j-ésima iteración y la pendiente inicial sea menor de, por ejemplo, el 1%, es decir, se haya producido la convergencia.

En nuestro ejemplo, se aprecia la poca magnitud de la pendiente de los residuales sobre X (0'03), cuya razón respecto a la pendiente inicial (1) es 0'03/1, que en porcentaje es del 3%, por lo que aún se podrá acercar más a 0 en próximas iteraciones. Para ello, se ha de realizar alguna nueva iteración entre los residuales de la anterior iteración y X.

En cualquier caso, para simplificar y no alargar excesivamente el apartado, supongamos que con la anterior iteración baste, entonces el valor de la pendiente y de la ordenada de la ecuación ajustada tras la iteración son los siguientes:

Sabiendo que

$$b_1=b_0+\delta_1$$

resulta que (empleando dos decimales),

$$b_1=1+0'03=1'03$$

Mientras, la ordenada en el origen responde a la fórmula,

$$a_1^*=a_0^*+\gamma_1$$

con lo que,

$$a_1^*=75'67+0'8=76'47$$

En consecuencia, la recta ajustada tendrá por ecuación,

$$\hat{Y}=76'47+1'03(X-177'5)$$

que es la línea ajustada de la relación entre X e Y mediante el proceso iterativo. Si expresamos la recta en la nomenclatura habitual, en función de la pendiente y de la ordenada en el origen resulta:

$$\hat{Y} = -106'36 + 1'03\ X$$

que es bastante similar a la obtenida con el procedimiento de la Línea de Tukey analizado en el Apartado 4.2.2.

Finalmente, hay que indicar que, como Emerson y Hoaglin (1983b) señalan, este método no es aplicable en diversos casos, en los que no se produce la convergencia.

4.3.2.- ANÁLISIS DE RESIDUALES.

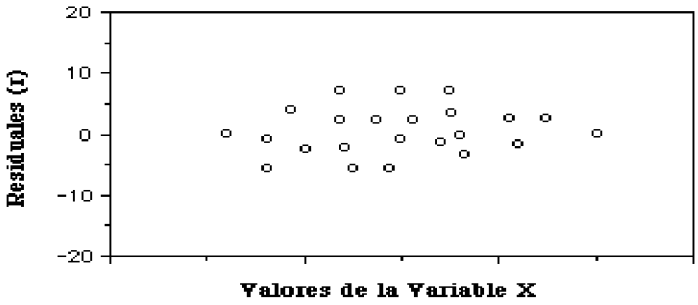
Es preciso indicar que el análisis de los residuales no es redundante, sino que tiene gran importancia para los análisis estadísticos. Concretamente, los residuos, tanto positivos como negativos deberían estar representados aleatoriamente entre los puntos $(X_i;rs_i)$ en forma de nube de puntos. Si los residuos positivos rs_i tienden a concentrarse en los valores menores de X_i ó en los valores mayores de X_i, entonces la suposición de que la relación entre X e Y es una función lineal ó la suposición de que las observaciones $Y_1,Y_2,...,Y_N$ son independientes puede haber sido violada. De hecho, si la gráfica de los puntos $(X_i;rs_i)$ muestra algún tipo de patrón regular, es que pueden haber sido violadas las hipótesis mencionadas anteriormente (EMERSON Y HOAGLIN, 1983b).

Además, como Goodall (1983a) indica, es recomendable el análisis de la distribución de residuos para así poder apreciar diversos aspectos tales como la asimetría de la distribución, la relación de los valores extremos respecto al grueso de los datos, observar los valores extremos, así como los diversos agrupamientos que puedan ocurrir. Para ello, se pueden efectuar las técnicas gráficas estudiadas en el Capítulo tercero, tal como un diagrama de "Tallo-y-Hojas" sobre la distribución de residuos. Mientras, si lo que se desea es no sólo realizar un análisis sobre los residuos por sí solos, sino un análisis más completo, típico de una actitud "exploratoria" por parte del analista, se pueden realizar los análisis que se describen a continuación.

En relación a la distribución de los residuales de los valores de Y sobre la variable X, Goodall (1983a) señala la existencia de cuatro patrones principales que se pueden observar tras la realización del diagrama de dispersión:

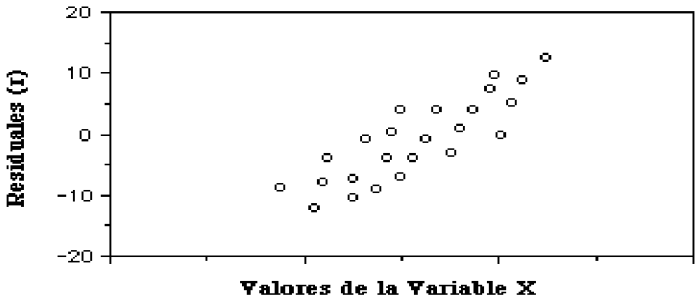
I) **Nube de puntos**. En la que no se aprecia ningún tipo de relación entre los residuales de Y sobre X, es decir, los residuales no presentan ningún patrón de tipo sistemático. Véase, para ello, la gráfica siguiente,

GRÁFICA Nº13.



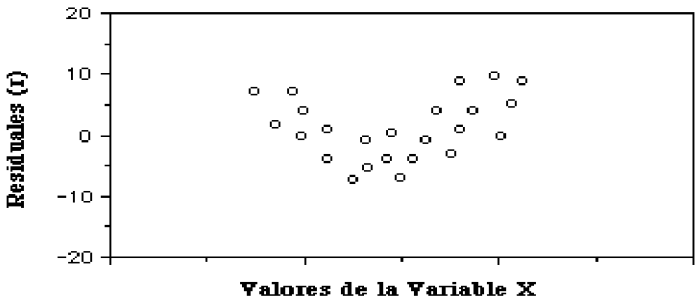
II) **"Franja o banda diagonal"**. En este caso se muestra una relación lineal entre X y los residuales de Y. Es decir, cabe asociar una recta a la franja que une los residuos de Y sobre X. Véase, como ilustración, la gráfica siguiente,

GRÁFICA Nº14.



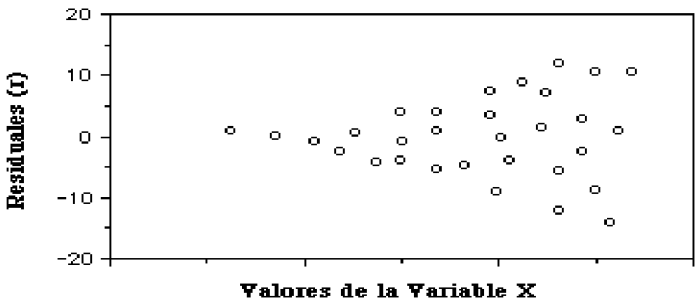
III) **"Franja o Banda Curva"**. En la que se revela una relación de tipo no lineal de los residuos de Y sobre X. En este caso, cabría hacer una transformación de los datos para ver si se puede estudiar linealmente. Véase la gráfica siguiente,

GRÁFICA Nº15.



IV) **"Cuña"**. En este caso, la dispersión de los residuos de Y se relaciona con los diferentes valores de X del modo siguiente: La densidad de los puntos varía con X, de modo que la variabilidad se incrementa (o decrementa) cuando se incrementan los valores de X. Véase, como ejemplo, la gráfica siguiente,

GRÁFICA Nº16.



En consecuencia, el análisis de residuales en el Análisis Exploratorio de Datos, como se indicó en el primer Capítulo, nos ayudará a observar la existencia de diversos patrones inesperados de los datos, lo que nos podrá servir para buscar explicaciones alternativas o analizar las razones por las que han ocurrido tales resultados inesperados. Si se desean explicaciones complementarias, el lector puede consultar el texto de Goodall (1983a)[7].

4.3.3.- LÍNEA DE REGRESIÓN DE MEDIANAS Y CUARTOS.

El método de la Línea de Regresión de Medianas y Cuartos resulta de utilidad especialmente cuando la variable que se sitúa en la abscisa, esto es, la variable X o variable independiente, sólo adopta unos pocos valores (o bien, se halla agrupada en intervalos de igual longitud). Por otra parte, se permite no sólo el análisis de la relación lineal entre X e Y, como el procedimiento de la Línea de Tukey, sino también de tipo curvilíneo.

Al igual que ocurría respecto a la relación lineal entre variables donde el procedimiento "clásico" de Mínimos Cuadrados empleaba la Media Aritmética, mientras que el procedimiento "exploratorio" de la Línea de Tukey empleaba la Mediana, el procedimiento que se va a analizar en este Apartado es análogo al que se realiza con la Línea de Regresión de Medias que se suele emplear en la denominada "Razón de Correlación" de la Estadística "clásica", pero empleando la Mediana en lugar de la Media.

El proceso consiste en dividir la muestra en tantos grupos como valores (o intervalos) de la variable X existan. Dentro de cada grupo, pues, habrá variaciones únicamente de la variable Y. Seguidamente, se calculan la Mediana y los dos Cuartos de la

[7] GOODALL, C. (1983a): Examining residuals. En HOAGLIN, D.C., MOSTELLER, F. Y TUKEY, J.W. (Eds.). *Understanding robust and exploratory data analysis*. New York: Wiley & Sons.

variable Y para cada grupo. Finalmente, se efectúa la representación gráfica, utilizando una línea continua (o gruesa) para unir los puntos del tipo (X_i;MdY_i), donde X_i representa a la puntuación del grupo i-ésimo (i=1,2,...,número de grupos), mientras que MdY_i representa la Mediana en la variable Y del grupo i-ésimo, con lo que, al unirlos, obtenemos la Línea de Regresión de Medianas. Igualmente, se emplean dos líneas discontinuas (o de poco grosor) para unir, por una parte, los puntos del tipo (X_i;C_{Si}), es decir, utilizando el Cuarto superior en lugar de la Mediana, y por otra parte, los puntos del tipo (X_i;C_{Ii}), es decir, empleando el Cuarto inferior. Entre ambas líneas discontinuas o de poco grosor (Líneas de Regresión de Cuartos) se hallará, aproximadamente, el 50% de las puntuaciones.

Veamos un ejemplo. Pensemos que se está analizando el nivel de conductividad de la piel (la variable Y, que se mide a través de un aparato que puntúa en una escala continua de 0 a 10), para ver la evolución en los primeros meses de vida de la misma, utilizando una muestra de niños desde 1 a 5 meses (el número de meses será, pues, la variable X o independiente). Supongamos que la muestra está compuesta por 35 niños, de manera que haya siete niños que tengan el mismo mes. Las puntuaciones (ficticias) que consideraremos, para el ejemplo, han sido las siguientes:

X	Y
1	2'4
1	2'1
5	8'5
3	3'2
2	2'4
1	2'5
1	3'4
2	2'8
3	3'1
2	4'2
2	2'8
2	3'9
1	3'4
3	4'0
4	4'9
5	2'9
2	3'4
3	3'4
3	4'1
5	4'0
4	5'3
5	3'1
4	3'2
5	6'7
2	3'1
4	3'6

5	9'2
5	8'3
4	6'2
1	4'2
4	7'3
1	2'3
3	4'1
3	3'2
4	3'9

En primer lugar, los datos se ordenan, de menor a mayor, conforme a la variable X y se forman los grupos. En nuestro caso, al haber cinco valores de la variable X, se formarán cinco grupos.

En el ejemplo, resultarán los siguientes valores de la variable Y:

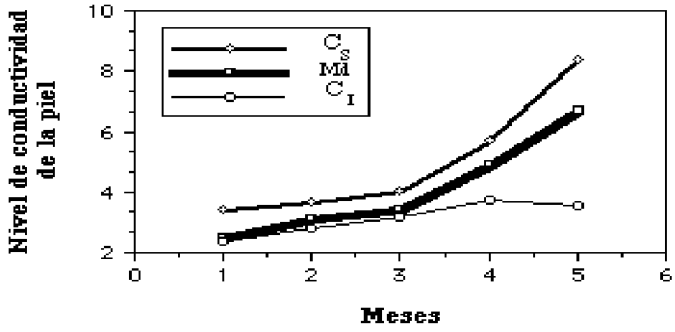
X=1	X=2	X=3	X=4	X=5
2'1	2'4	3'1	3'2	2'9
2'3	2'8	3'2	3'6	3'1
2'4	2'8	3'2	3'9	4'0
2'5	3'1	3'4	4'9	6'7
3'4	3'4	4'0	5'3	8'3
3'4	3'9	4'1	6'2	8'5
4'2	4'2	4'1	7'3	9'2

Seguidamente, hay que obtener la Mediana y los Cuartos (superior e inferior) en la variable Y para cada uno de los grupos. El resultado será el siguiente:

	X=1	X=2	X=3	X=4	X=5
Md	2'5	3'1	3'4	4'9	6'7
C_I	2'35	2'80	3'20	3'75	3'55
C_S	3'40	3'65	4'05	5'75	8'40

Una vez obtenidos estos índices, se elabora, como paso final, la representación gráfica correspondiente. Por ejemplo, para elaborar la línea continua o gruesa (la de las Medianas) se une primeramente el punto $(X_1 ; MdY_1)$ con el $(X_2 ; MdY_2)$, y así sucesivamente. El proceso es análogo para los Cuartos superiores e inferiores (utilizando líneas de menor grosor). El resultado definitivo es el siguiente:

GRÁFICA Nº17.



En el gráfico se puede observar, por una parte, el aumento acelerado de la variable Y a lo largo de los meses (variable X), y la asimetría negativa de los primeros meses, así como el aumento también de la variabilidad entre los propios sujetos, como lo denota la amplia separación de las líneas entre Cuartos (que engloban, aproximadamente, al 50% de las puntuaciones) en los dos últimos periodos (meses 4º y 5º).

4.3.4.- MÉTODO DE LAS MEDIANAS MÓVILES (*Running Medians*): PROCEDIMIENTO Y CARACTERÍSTICAS.

Las Medianas Móviles son un procedimiento de suavización que ya ha sido estudiado en el segundo Capítulo, al analizar los estadísticos resistentes. Por ello, antes de continuar la lectura de este apartado, se recomienda que se tenga ya un adecuado nivel de conocimiento del mismo.

La utilidad de este método tiene lugar, básicamente, al analizar series temporales, es decir, fenómenos que varían a lo largo del tiempo. A la hora de efectuar representaciones

gráficas de estos fenómenos, en el eje de abscisas se situarán los valores temporales, mientras que en el de ordenadas se situarán los de la variable concreta estudiada. Como ejemplos de utilización de este procedimiento, se puede señalar, la evolución del éxito de un determinado tratamiento a lo largo de 15 semanas, con lo que tendríamos 15 datos referidos al éxito del tratamiento, las preferencias de voto de determinado partido a lo largo de diversos periodos electorales, o las preferencias de elección, en la conducta del consumidor, por una determinada forma de presentación de un mismo producto.

Ya que el procedimiento de cálculo ha sido previamente analizado (Apartado 2.4), indicaremos únicamente un ejemplo de cálculo del mismo. Simplemente, cabe añadir (HARTWIG Y DEARING, 1979) que este procedimiento suele ser realizado iterativamente de modo que una vez se obtenga la columna con las Medianas Móviles se ha de volver a realizar el mismo procedimiento sobre la columna transformada. Si esa nueva columna no muestra ningún cambio respecto a la anterior, el proceso habrá finalizado. Pero si esa columna varía, habrá que realizar una nueva transformación, y así sucesivamente, hasta que no se aprecien cambios.

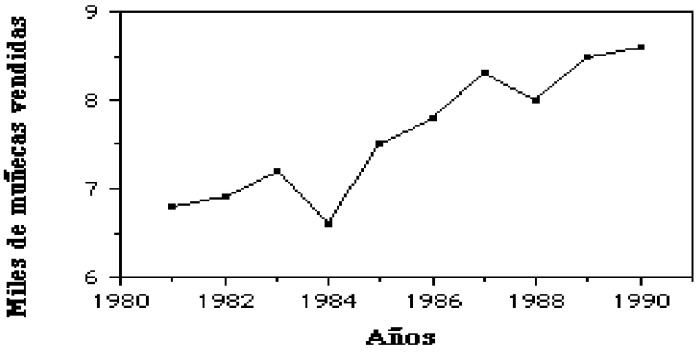
Pensemos que un investigador desea averiguar la evolución de las ventas de las muñecas marca "ACME" en Benetússer[8], durante el periodo 1981-1990. Los valores son los siguientes (el número de muñecas se expresa en miles de unidades vendidas):

Año	Nº muñecas vendidas
1981	6'8
1982	6'9
1983	7'2
1984	6'6
1985	7'5
1986	7'8
1987	8'3
1988	8'0
1989	8'5
1990	8'5

El gráfico resultante de unir estos valores resulta como sigue:

[8] Gran pueblo de la comarca de l'Horta (Sud) del País Valenciano.

GRÁFICA Nº18.



Se pueden apreciar las diversas alteraciones que hay en el gráfico sobre una secuencia ascendente de datos. Supongamos, igualmente, que el analista desea suavizar la gráfica empleando en la investigación Medianas Móviles de orden 3. En consecuencia, y utilizando el procedimiento que se expresa en el Apartado 2.4 del presente texto se obtiene una nueva columna con las Medianas Móviles de orden 3:

Año	Nº muñecas vendidas	MM Orden 3
1981	6'8	(6'8)
1982	6'9	6'9
1983	7'2	6'9
1984	6'6	7'2
1985	7'5	7'5
1986	7'8	7'8
1987	8'3	8'0
1988	8'0	8'3
1989	8'5	8'5
1990	8'6	(8'6)

Ahora, cabe repetir el proceso (es decir obtener Medianas Móviles de orden 3) aplicado sobre la columna ya transformada, para ver si no hay cambios. Caso de no haberlos el proceso ya habría acabado. En nuestro ejemplo, resulta como sigue (se añade

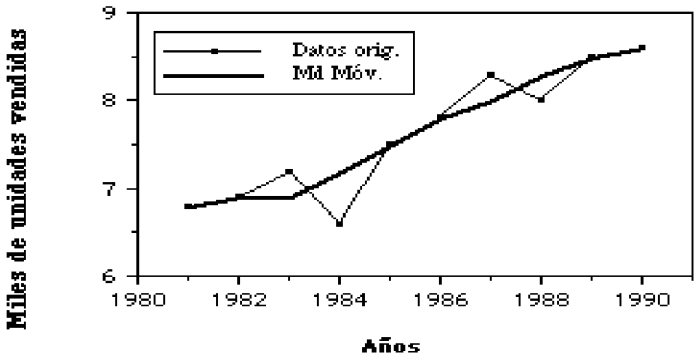
una R a la última columna para indicar que es una repetición del mismo proceso),

Año	Nº muñecas vendidas	MM Orden 3	MM Orden 3 R
1981	6'8	(6'8)	(6'8)
1982	6'9	6'9	6'9
1983	7'2	6'9	6'9
1984	6'6	7'2	7'2
1985	7'5	7'5	7'5
1986	7'8	7'8	7'8
1987	8'3	8'0	8'0
1988	8'0	8'3	8'3
1989	8'5	8'5	8'5
1990	8'6	(8'6)	(8'6)

Se aprecia que no ha habido cambios entre las dos últimas columnas, por lo que no hacen falta más iteraciones.

Si ahora realizamos el gráfico utilizando las Medianas Móviles de orden 3 y uniendo los puntos se obtiene la siguiente representación:

GRÁFICA Nº19.



Se aprecia la diferencia respecto al gráfico que utilizaba los datos originales, de manera que en este segundo hay una gran

suavización de los datos, con lo que se puede apreciar mejor la tendencia general de los datos (o *smooth* en este caso), eliminando buena parte de los aspectos residuales de los datos originales (es decir, parte exclusiva o *rough* en este caso), por ejemplo, los datos correspondientes a 1984 y 1988 (en el que hay una bajada de las ventas), concentrándose en la parte ajustada o suavizada de los datos. Un aspecto de interés para el investigador es tratar de explicar la razón por la que ha habido tales bajadas en el número de ventas, o bien de la tendencia actual de las ventas.

4.3.5.- MÉTODO "HANNING": CARACTERÍSTICAS.

El denominado método "Hanning" es, al igual que el procedimiento de Medianas Móviles, un método de suavización, también aplicable preferentemente al análisis de series temporales. De hecho, tal y como se podrá apreciar es bastante parecido al anterior. Simplemente, este procedimiento realiza dos veces el procedimiento visto de las Medianas Móviles pero a través de pares de puntuaciones consecutivas (es decir, a través de Medianas Móviles de orden 2). Debido a que la Mediana de un par de puntuaciones es igual que la Media Aritmética entre ellas, en este caso, es lo mismo indicar que se calcula la Media que la Mediana (véase Apartado 2.4). Es decir, cabría decir, igualmente, que se realiza dos veces un proceso de cálculo de Medias Móviles de orden 2.

El procedimiento **"HANNING"** consta de dos pasos:

I) Una vez se tiene una serie de datos a suavizar dispuestos en una columna de acuerdo con una determinada variable, se crea una nueva columna, en la que se calculará la Media (o la Mediana, por el razonamiento expuesto antes) entre los valores consecutivos de aquella columna. Tal proceso podrá realizarse bien sobre una columna de datos originales, bien

sobre una columna de datos resultantes de una suavización previa. Tomando este segundo caso como más completamente ilustrativo, veremos primero un ejemplo de aplicación del método "Hanning" sobre datos previamente suavizados por otro procedimiento (como el de Medianas Móviles).

Regresando a nuestro ejemplo, supondremos que vamos a suavizar aún más los datos tras la aplicación del procedimiento de las Medianas Móviles visto en el Apartado anterior. Primeramente se calcula la Media Aritmética entre 6'8 y 6'9 y se coloca entre los años 1981 y 1982, después se hace lo mismo para los pares consecutivos, hasta llegar al par 8'5 y 8'6, que se situará entre 1989 y 1990. En consecuencia, se obtienen los siguientes valores,

Año	Nº muñecas vendidas	M.M. Orden 3	Hanning (Fase 1)
1981	6'8	(6'8)	
			6'85
1982	6'9	6'9	
			6'90
1983	7'2	6'9	
			7'05
1984	6'6	7'2	
			7'35
1985	7'5	7'5	
			7'65
1986	7'8	7'8	
			7'90
1987	8'3	8'0	
			8'15
1988	8'0	8'3	
			8'40
1989	8'5	8'5	
			8'55
1990	8'6	(8'6)	

II) En el segundo paso se vuelve a efectuar una Media Aritmética entre cada dos valores consecutivos, en este caso sobre los valores obtenidos en la columna "Hanning" de la primera fase, con lo que se vuelven a calcular los datos suavizados para un año (y no como en la primer fase que se hallaban entre dos años). Mientras, los valores referidos al primer y último año se calculan teniendo en cuenta el primer y último valor de la columna sobre la que se efectúa la transformación que, en el ejemplo, es la columna de las Medianas

Móviles, concretamente los valores entre paréntesis. Seguidamente, se puede efectuar la representación gráfica de tales valores.

En nuestro ejemplo, el primer valor, el correspondiente a 1981 se calcula hallando la Media (o la Mediana) entre el primer valor entre paréntesis de la columna de las Medianas Móviles y el primer valor de la columna del primer paso del método "Hanning" (es decir, 6'8 y 6'85). Seguidamente, para el año 1982 se calcula entre los dos valores siguientes de la columna "Hanning" (es decir, 6'85 y 6'90), y así hasta llegar al año 1990 en que se vuelve a utilizar el valor entre paréntesis (el último, 8'6) de la columna de las Medianas Móviles para hacer Media con el último valor de la columna "Hanning" (esto es, 8'55).

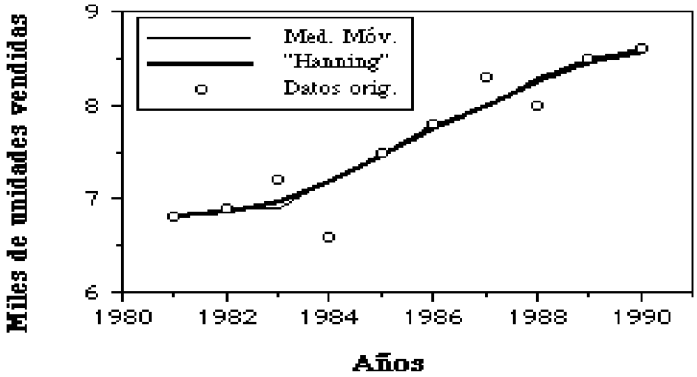
Por tanto, se obtiene la tabla siguiente:

Año	Nº muñecas vendidas	MM Orden 3	Hanning (1º)	Hanning (2º)
1981	6'8	(6'8)		6'825
			6'85	
1982	6'9	6'9		6'875
			6'90	
1983	7'2	6'9		6'975
			7'05	
1984	6'6	7'2		7'200
			7'35	
1985	7'5	7'5		7'500
			7'65	
1986	7'8	7'8		7'775
			7'90	
1987	8'3	8'0		8'025
			8'15	
1988	8'0	8'3		8'275
			8'40	
1989	8'5	8'5		8'475
			8'55	
1990	8'6	(8'6)		8'575

Los valores de la columna de la segunda fase del método "Hanning" muestran la suavización de los datos, mayor que en el

procedimiento de las Medianas Móviles (ya que supone una suavización de los datos previos). En nuestro ejemplo, sin embargo, las diferencias entre el método de las Medianas Móviles respecto al método "Hanning" han sido solamente ligeras (habría de haber una mayor variación en los datos para que se apreciara mejor el ajuste de uno y otro método), tal y como se puede observar en el siguiente gráfico,

GRÁFICA Nº20.



Viendo el gráfico de puede observar el creciente aumento de las ventas de las muñecas "ACME", aunque que el crecimiento de las ventas haya aumentado poco en los dos últimos años.

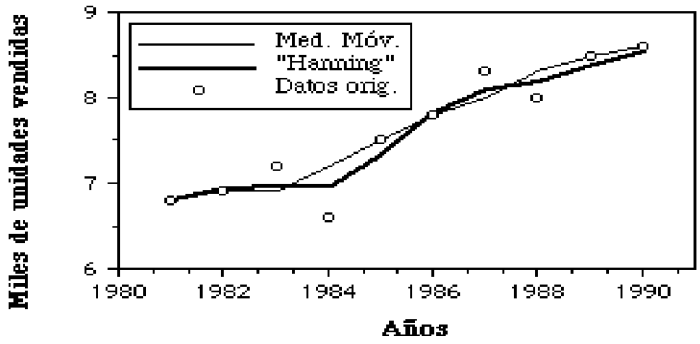
Por otra parte, como se ha indicado previamente, también se puede efectuar el método "Hanning" directamente sobre los datos originales. En nuestro ejemplo resultará la tabla final siguiente:

Año	Nº muñecas vendidas	Hanning (Fase 1)	Hanning (Fase 2)
1981	6'8		6'825
		6'85	
1982	6'9		6'950
		7'05	
1983	7'2		6'975
		6'90	
1984	6'6		6'975
		7'05	
1985	7'5		7'350
		7'65	
1986	7'8		7'850
		8'05	

1987	8'3		8'100
		8'15	
1988	8'0		8'200
		8'25	
1989	8'5		8'400
		8'55	
1990	8'6		8'575

La gráfica comparativa de la aplicación del método "Hanning" sobre datos originales y del método de las Medianas Móviles de orden 3 se expone a continuación:

GRÁFICA Nº21.



Se pueden observar las ligeras diferencias entre ambos resultados. En cualquier caso, como se ha sugerido, el procedimiento "Hanning" se suele aplicar, normalmente, sobre datos ya suavizados, previamente por otro método (HARTWIG Y DEARING, 1979).

En referencia a la Psicología, tales análisis tienen su utilidad no tanto en los estudios de laboratorio sino, principalmente, en áreas más aplicadas tales como la Psicología Clínica. Un ejemplo sugerente de aplicación del método "Hanning" en este área sería el conteo del número de cigarrillos fumados cada día a lo largo de un periodo de tiempo en el caso de una persona que se halle bajo tratamiento para dejar la adicción al tabaco, con lo que el psicólogo podrá tener una clara idea de la evolución temporal de la persona bajo tratamiento respecto a la, por ejemplo, frecuencia de cigarrillos consumidos por día

(minimizando la "engañosa" variabilidad de cada uno de los días, o de determinados días).

4.3.6.- RELACIÓN NO-LINEAL: PROCEDIMIENTOS DE TRANSFORMACIÓN DE DATOS.

Si una vez realizado el diagrama de dispersión se observa que la forma de la relación entre dos variables no es lineal, cabe recurrir a los procedimientos de "re-expresión" o transformación de los datos, ya sea de la variable X, de la variable Y, o de ambas. Por ejemplo, se puede transformar la escala original de los datos de la variable X efectuando su logaritmo (o alguna operación empleando potencias) y seguidamente observar si la relación entre los datos transformados de X e Y es ya lineal a través del correspondiente diagrama de dispersión.

En el caso de que tras la transformación de los datos, el diagrama de dispersión muestre una forma de la relación entre las variables X e Y aproximadamente lineal ya se podrán emplear los procedimientos lineales estudiados en este Capítulo. En el caso de que la relación, aun con la transformación de los datos aplicada, siga siendo no-lineal, habrán de emplearse otros métodos estadísticos que, salvo el procedimiento de la Línea de Regresión de Medianas y Cuartos (analizado en el Apartado 4.3.3), no son tratados en el presente texto.

Tampoco, en este texto, se detallarán los procedimientos concretos para la realización de las adecuadas transformaciones de datos (ni las propiamente destinadas a lograr la simetría de una serie de datos, o las destinadas a lograr linealidad entre dos variables), ya que exceden los objetivos propuestos de realizar una primera introducción al tema. Para un detallado análisis de los procedimientos de transformación de datos en el caso de relaciones no lineales entre variables remitimos al

lector al texto de Emerson (1983), pero cuyo nivel es ya relativamente alto[9].

[9] EMERSON, J.D. (1983): Mathematical Aspects of Transformation. En HOAGLIN, D.C., MOSTELLER, F. Y TUKEY, J.W. (Eds.). *Understanding robust and exploratory data analysis*. New York: Wiley & Sons.