CORRELACION, REGRESION, ERROR TIPICO

Por ANTONIO ALONSO NUÑEZ
(*Estadístico Técnico del I. N. E.*)

OO.—Con pretensiones puramente didácticas, se indica a continuación el método seguido para enseñar los conceptos citados en el título a un grupo de estudiantes del Curso Preuniversitario (Instituto de Pontevedra). La exposición necesitó cuatro sesiones.

O.—El día anterior al comienzo se les encarga que repasen la media aritmética, la desviación típica y el ajuste de una recta por mínimos cuadrados (se trata de explicar tan sólo la correlación rectilínea).

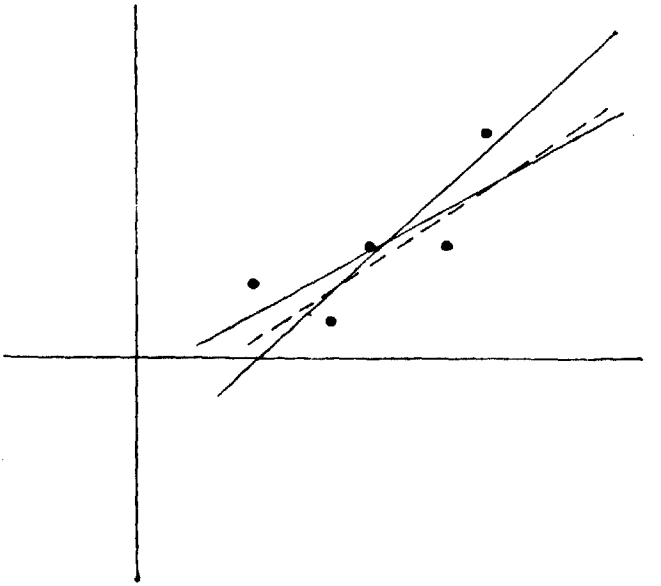

Figura 1

1. Supongamos un colectivo, de cada uno de cuyos elementos vamos a estudiar dos características: obtendremos una variable estadística bidimensional (tabulada en el cuadro núm. 1).

Cuadro número 1

X	Y
X_1	Y_1
X_2	Y_2
.	.
.	.
.	.
X_n	Y_n

Cuadro número 2

X	Y
3	2
5	1
6	3
8	3
9	6

Ejemplo: Se preguntó a cinco alumnos el número de veces que habían pedido dinero a sus padres en un mes, y el número de veces que en ese mes fueron al cine. Variable X: número de veces que pidieron dinero; variable Y: número de veces que fueron al cine. (Cuadro núm. 2).

Llevados estos pares de valores a unos ejes coordenados, determinarán una nube de puntos. La figura 1 corresponde al cuadro núm. 2. Cada alumno debe hacer un dibujo de la misma, a escala bastante grande. (En la figura 1 se han añadido las tres rectas que se citan en el apartado siguiente).

2. Se pide a los alumnos que adapten a la nube una recta que, pasando entre los puntos, indique a su juicio la tendencia del fenómeno. (La recta de trazos de la figura 1 fué hecha realmente por un alumno). Es ahora el momento de indicarles las diversas rectas de tendencia posibles, y que hay casos en que ésta se representa mejor por una curva. Los dos valores "3" que la variable "Y" presenta en los alumnos 3.º y 4.º proporcionan base para explicar la diferencia entre dependencia funcional y dependencia estadística.

3. Se les pide ahora la adaptación de una recta por mínimos cuadrados, explicándoles por qué es preferible este procedimiento a la anterior adaptación "a ojo". (Conviene dejar que sigan su tendencia a calcular la regresión de Y sobre X.)

El método que se va a seguir significa que, respetando las X reales, las Y reales serán sustituídas por unos valores teóricos ("normales"), obtenidos por medio de la ecuación $Y = a + bX$. (Y, valores normales.) Véase cuadro núm. 3.

Cuadro número 3

X	Y	$Y = a + bX$
X_1	Y_1	$\underline{Y}_1$
X_2	Y_2	$\underline{Y}_2$
.	.	.
.	.	.
.	.	.
X_n	Y_n	$\underline{Y}_n$

Cálculos:

De $\Sigma (Y - \underline{Y})^2 = \text{mínimo}$, obtenemos el siguiente sistema:

$\Sigma Y = na + b \Sigma X$
$\Sigma XY = a \Sigma X + b \Sigma X^2$

Ver cuadro núm. 4 (en el que se incluye ΣY^2, necesario más tarde).

Cuadro número 4

X	Y	X_2	Y_2	XY
3	2	9	4	6
5	1	25	1	5
6	3	36	9	18
8	3	64	9	24
9	6	81	36	54
31	15	215	59	107

$$\left.\begin{array}{l} 15 = 5a + 31b \\ 107 = 31a + 215b \end{array}\right\} \begin{array}{l} a = -0'81 \\ b = 0'61 \end{array}$$

La ecuación resultante es, pues, $\bar{Y} = -0'81 + 0'61X$. (Figura 1: conviene que para dibujarla calculen los alumnos todos los valores de $\bar{Y}$, y no tan sólo los dos suficientes, con objeto de que comprendan mejor el significado de la recta adaptada; además, estos valores tendrán aplicación para el cálculo del error típico. (Véase cuadro núm. 5.)

Comparada la recta adaptada a ojo y la calculada, se observa que ésta tiende a acercarse más a los puntos en sentido vertical que aquélla: se les presenta ahora la idea de otra posible línea de tendencia que se acerque más a los puntos en sentido horizontal; desde luego, nada de esto podría darse si los puntos se adaptasen exactamente a la recta (de nuevo diferencia entre dependencia funcional y dependencia estadística).

El sistema normal para el cálculo de la recta de regresión de X sobre Y (datos del cuadro núm. 4), es:

$$\left.\begin{array}{l} 31 = 5c + 15d \\ 107 = 15c + 59d \end{array}\right\} \begin{array}{l} c = 3'2. \\ d = 1. \end{array}$$

La nueva recta de regresión es $\bar{X} = 3{,}2 + Y$ (figura 1: Tal como se hizo anteriormente, conviene ahora que calculen todos los $\bar{X}$; ver cuadro núm. 6. Para mejor comprensión, conviene que construyan tablas con los valores X, Y, $\bar{Y}$ y X, $\bar{X}$, Y. Como preparación para el párrafo siguiente, y para acabar de afianzar el concepto de rectas de regresión, sería también interesante hacer un cuadro a 6 columnas, en el que figurasen $\Sigma|Y - \bar{Y}|$ y $\Sigma|X - \bar{X}|$ para cada una de las tres rectas de la figura 1, pero no por cálculo, sino simplemente medidas con el doble decímetro: aparecería patente la tendencia de las rectas de regresión a adaptarse a la nube de puntos en sentido horizontal o vertical).

4. Los sistemas normales se obtuvieron a partir de $\Sigma (Y - \bar{Y})^2 =$ mínimo y $\Sigma (X - \bar{X})^2 =$ mínimo: cuanto menor sea este mínimo, tanto mejor representarán las rectas de *regresión* al fenómeno; el caso ideal es cuando este mínimo valga cero. Por tanto, este mínimo puede servir de medida de la bondad de las rectas de regresión, etc.: se da ahora la fórmula del error típico, hacien-

do un paralelo entre media aritmética-recta de regresión y desviación típica-error típico.

Cálculo de los dos errores típicos:

Cuadro número 5

X	Y	$\underline{Y}$	$Y - \underline{Y}$	$(Y - \underline{Y})^2$
3	2	1'02	0'98	0'9604
5	1	2'24	−1'24	1'5376
6	3	2'85	0'15	0'0225
8	3	4'07	−1'07	1'1449
9	6	4'68	1'32	1'7424
				5'4078

$$S_y = \sqrt{\frac{5{,}4078}{5}} = 1{,}04$$

Cuadro número 6

X	$\underline{X}$	Y	$X - \underline{X}$	$(X - \underline{X})^2$
3	5'2	2	−2'2	4'84
5	4'2	1	0'8	0'64
6	6'2	3	−0'2	0'04
8	6'2	3	1'8	3'24
9	9'2	6	− 0'2	0'04
				8'80

$$S_x = \sqrt{\frac{8{,}80}{5}} = 1{,}3.$$

A las pendientes *b* y *d* de las rectas de regresión se les llama coeficientes de regresión, y suelen representarse, respectivamente, b_{yx} y b_{xy} (regresión de Y sobre X, regresión de X sobre Y). Es de señalar que son pendientes respecto a distintos ejes. Por tanto, caso de que las dos rectas sean una misma (que es el caso de máxima *correlación* posible), $b_{yx} \cdot b_{xy} = 1$. Este valor nos lleva a investigar la variabilidad posible del producto de ambos factores; no sería difícil demostrar que ambos son siempre del mismo signo, el valor máximo del producto es 1 (en cuyo caso los dos errores típicos son cero) y el mínimo es 0 (en cuyo caso los dos errores típicos adquieren el valor máximo posible). Pero como esto se podrá hacer más fácilmente en el capítulo siguiente, lo damos por demostrado y definimos el "coeficiente de correlación":

$$r^2 = b_{yx} \cdot b_{xy} \qquad r = \sqrt{b_{yx} b_{xy}}$$

que indica la mayor o menor tendencia que presentan las dos características a asociarse (es decir, a dejarse representar conjuntamente por las rectas de regresión). Al coeficiente de correlación se le pone el signo + cuando los dos coeficientes de regresión sean positivos, y el signo — cuando negativos (correlación directa e inversa, respectivamente).

Con esto termina la primera sesión.

5. Con objeto de demostrar lo indicado al final de la primera sesión, y, al mismo tiempo, encontrar métodos sencillos de cálculo de las diversas medidas, es conveniente hacer una traslación paralela de ejes al baricentro de la distribución (es decir, al punto cuya abscisa es la media aritmética de las abscisas y cuya ordenada es la media aritmética de las ordenadas). (Figura 2.)

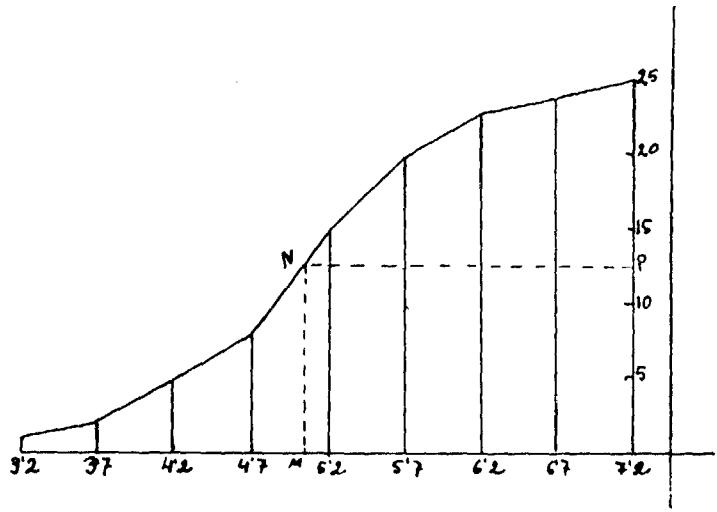

Figura 2

El cuadro original de valores queda sustituído por las desviaciones de éstos respecto a sus medias aritméticas respectivas:

Cuadro número 7

$x = X - \overline{X}$	$y = Y - \overline{Y}$
x_1	y_1
x_2	y_2
.	.
.	.
.	.
x_n	y_n

Cuadro número 8

$x = X - 6{,}2$	$y = Y - 3$
—3,2	—1
—1,2	—2
—0,2	0
1 8	0
2,8	3
0,0	0,0

Los sistemas normales para el cálculo de las rectas de regresión son ahora:

$$\left.\begin{array}{l} \Sigma y = na' + b' \Sigma x \\ \Sigma xy = a' \Sigma x + b' \Sigma x^2 \end{array}\right\} \qquad \left.\begin{array}{l} \Sigma x = nc' + d' \Sigma y \\ \Sigma yx = c' \Sigma y + d' \Sigma y^2 \end{array}\right\}$$

Como $\Sigma x = \Sigma y = 0$, tenemos $a' = c' = 0$: las rectas de regresión pasan por el baricentro. Como se trata de traslación paralela, $b' = b$ y $d' = d$. Por tanto, los coeficientes de regresión pueden calcularse por las siguientes expresiones, sin necesidad de plantear sistemas de ecuaciones:

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} \qquad b_{xy} = \frac{\Sigma xy}{\Sigma y^2}$$

Como comprobación, calculemos los coeficientes de regresión por este procedimiento:

Cuadro número 9

x^2	y^2	xy
10,24	1	3,2
1,44	4	2,4
0,04	0	0,0
3,24	0	0,0
7,84	9	8,4
22,80	14	14,0

$$b_{yx} = \frac{14}{22\,8} = 0{,}61.$$

$$b_{xy} = \frac{14}{14} = 1.$$

Los coeficientes de regresión sólo pueden valer 0 en caso de que lo valga Σxy, y, por tanto, lo serían ambos simultáneamente. En este caso decimos que no existe correlación entre ambas variables, ya que las mejores adaptaciones serían dos rectas perpendiculares entre sí (que forman el mayor ángulo "agudo" posible), y con las que sustituirían todos y cada uno de los valores por su media aritmética.

Ambos coeficientes tienen el mismo signo, que es el de Σxy.

Los errores típicos pueden calcularse a partir de los cuadros 7 y 8 igual que se hizo a partir de los cuadros anteriores. En efecto, para todo el valor de Y (y análogamente para todo valor de X), tenemos $Y - \bar{Y} = y - \bar{y}$. Esto se ve claramente en la figura 2, en que $\overline{PA} = \overline{PC} - \overline{AC} = \overline{PB} - \overline{AB}$. Cuando $b_{yx} = b_{xy} = 0$, los errores típicos son máximos, porque $\Sigma(y - \bar{y})^2 = \Sigma y^2$.

El coeficiente de correlación se obtiene ahora a partir de la siguiente expresión:

$$r^2 = \frac{(\Sigma xy)^2}{\Sigma x^2 \, \Sigma y^2}.$$

Esta cantidad es esencialmente positiva. Vale 0 cuando lo valgan los coeficientes de regresión, y ya queda dicho que entonces entre ambas variables no hay correlación. Para demostrar que su valor máximo es 1, desarrollémoslo:

$$r^2 = \frac{(x_1 y_1 + x_2 y_2 + \ldots)^2}{(x_1^2 + x_2^2 + \ldots)(y_1^2 + y_2^2 + \ldots)} = \frac{x_1^2 y_1^2 + x_2^2 y_2^2 + \ldots + 2x_1^2 y_1^2 x_2^2 y_2^2 + \ldots}{x_1^2 y_1^2 + x_2^2 y_2^2 + \ldots + x_1^2 y_2^2 + x_2^2 y_1^2 + \ldots}$$

Los dos primeros términos escritos de numerador y denominador son iguales; el último del numerador y los dos últimos del denominador son los tres

términos del cuadrado de un binomio. Según una conocida propiedad (1), el término del numerador es, como máximo, igual a la suma de los otros dos: de ahí deducimos que el límite superior de r^2 es 1. Para que este límite se alcance es preciso que todo $x_p y_q = x_q y_p$, es decir $y_p/x_p = y_q/x_q =$ constante: todos los puntos pertenecen a una recta que pasa por el origen (baricentro) y los errores típicos valen 0 por ser nulas todas y cada una de las desviaciones. Cuanto más se dispersen en torno a la recta, tanto más se alejará r^2 del valor límite 1 y el mínimo es 0 (en cuyo caso los dos errores típicos adquieren su valor máximo: $Sx = \sigma x$, $Sy = \sigma y$).

Termina así la segunda sesión.

6. La tercera y cuarta sesión se dedican a la obtención puramente algebraica de las diversas variantes de las fórmulas, con agrupación en frecuencias, métodos prácticos de cálculo y tratamiento de uno o dos ejemplos serios.

GRAFICOS ACUMULATIVOS

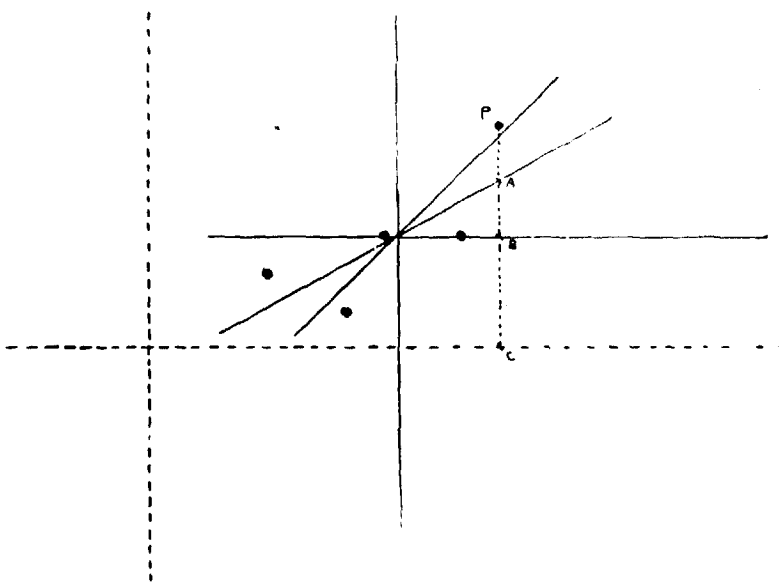

Figura 3

Es error frecuente en los alumnos al construir estos gráficos (ojiva, papel probabilístico normal, etc.), levantar las ordenadas sobre las marcas de clase, en lugar de hacerlo en los extremos sobre los límites superior o inferior de

(1) $(a+a)^2 = [(a+x) + (a-x)]^2 = (a+x)^2 + (a-x)^2 + 2(a+x)(a-x) = 2(a^2+x^2) + 2(a^2-x^2)$.

El primero de estos dos últimos términos es tanto mayor que el segundo, cuando mayor sea X. Sólo son iguales para X = 0.

los intervalos (según se trate de ojiva ascendente o descendente). Indudablemente, incurren en este error por aplicar a la acumulativa la regla adecuada para la construcción del histograma. Las medidas hechas sobre tal gráfico (media, mediana, cuartiles, etc.), se obtienen con un error, en menos o en más, de un semiperíodo.

La regla adecuada no siempre viene suficientemente explícita en los textos, y aún hay alguno que incurre en el mismo defecto. Elijo un libro ya algo antiguo como ejemplo:

Intervalos	Frecuencias	Sumas
3,0—3,4	1	1
3,5—3,9	1	2
4,0—4,4	3	5
4,5—4,9	3	8
5,0—5,4	7	15
5,5—5,9	5	20
6,0—6,4	3	23
6,5—6,9	1	24
7,0—7,4	1	25

$$\text{Mediana} = 5 + \frac{12{,}5 - 8}{7} \cdot 0{,}5 = 5{,}32$$

Gráficamente dice (V. figura 3).

P = punto medio de la ordenada final. Segmento $\overline{PN}$, paralelo al eje X; segmento $\overline{NM}$ perpendicular al mismo eje. Mediana = abscisa del punto M = 5,05.

Si se le añade un semiintervalo, obtenemos un resultado mucho mejor: 5,30, quedando compensado el error gráfico.