EL COEFICIENTE DE CORRELACION Y LAS RECTAS DE REGRESION: su aplicación en Geografía

por ORLANDO PEÑA ALVAREZ

La Estadística ofrece a la Geografía una serie de instrumentos de trabajo útiles en cualquiera investigación que se emprenda. Entre ellos consideraremos particularmente el cálculo y el uso del coeficiente o factor de correlación.

En Geografía diversos fenómenos están ligados unos con otros. La no-independencia de dichos valores puede ser expresada matemática y gráficamente sobre la base de los principios siguientes:

1. La correlación no se expresa bajo la forma de una función, toda vez que es imposible establecer una progresión exactamente paralela entre dos series de datos que no tienen un vínculo específico. La correlación ilustra sobre las chances, más o menos fuertes, de que dos variables sigan un camino relativamente semejante.

2. El cálculo del coeficiente de correlación (R) requiere, como toda operación estadística, el uso de abundante información. Es conocido, por ejemplo, que un "promedio" en Climatología tiene el valor de "normal" sólo si ha sido construido sobre la base de datos correspondientes a una serie de —por lo menos— treinta años sucesivos.

En el caso que nos interesa se comienza por una ordenación de los datos sobre una tabla de frecuencias. Se calcula la "desviación" de cada término con respecto al promedio del período escogido, se les eleva al cuadrado y se les suma: tratándose de dos columnas (de x y de y), en ambas se procede de la misma manera, obteniéndose al final dos cifras que corresponden a $\Sigma\ (x-\bar{x})^2$ y a $\Sigma\ (y-\bar{y})^2$. Para simplificar las operaciones se ha convenido en reemplazar los términos $(x_i-\bar{x})^2$ y $(y_i-\bar{y})^2$ por u^2 y v^2, respectivamente.

Las desviaciones de cada serie se multiplican entre sí, por parejas, construyéndose con los resultados una nueva columna de datos que permite llegar, por suma, a $\Sigma\ (x_i-\bar{x})\ (y_i-\bar{y})$, o bien, $\Sigma u_i\, v_i$. Es conveniente señalar que este resultado puede ser positivo o negativo, según exista un signo u otro en los datos que se usan para la operación. Al multiplicar $(x_i-\bar{x})$ por $(y_i-\bar{y})$ se dan dos posibili-

dades: o ambos valores tienen el mismo signo (+ o —) y en ese caso el producto es positivo, o tienen signos opuestos, razón por la cual el producto es negativo.

Una vez obtenidos los datos indicados más arriba se aplica la fórmula siguiente:

$$R = \frac{\Sigma (x_i - \overline{x})\ (y_i - \overline{y})}{\sqrt{\Sigma (x_i - \overline{x})^2\ \Sigma (y_i - \overline{y})^2}} \qquad \text{o sea, } R = \frac{\Sigma u_i\, v_i}{\sqrt{\Sigma u_i^2\, \Sigma v_i^2}}$$

3. El coeficiente de correlación puede ser representado por una cifra positiva o negativa que oscila entre +1 y —1. Mientras más cercano está de la unidad el coeficiente muestra una correlación más estrecha; si se aproxima a +1 significa que la variación de uno de los términos responde a la variación del otro en el mismo sentido; si —por el contrario— se acerca a —1, quiere decir que la variación se efectúa también pero en sentido inverso. En el primer caso se habla de correlación positiva o directa y en el otro, de correlación negativa o inversa. En cambio, si el índice se acerca a 0 estamos en presencia de una correlación muy débil o, más aún, de una correlación nula ($R = 0$), lo que implica que cada término varía de manera totalmente independiente del otro.

A través de un ejemplo, reproduciremos el proceso indicado. Los datos corresponden a las precipitaciones medidas en Valparaíso y Santiago, anualmente, durante el período 1951-1965. Se trata de determinar si las variaciones pluviométricas en alguna de estas estaciones se reproducen en la otra; si eso ocurre, importa saber asimismo la magnitud de esa correlación.

LAS PRECIPITACIONES DEL PERIODO 1951-1965 EN VALPARAISO Y SANTIAGO
(VALORES ANUALES)

Años	*Valparaíso*			*Santiago*			*Covarianza*
	x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	y_i	$y_i - \overline{y}$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})\ (y_i - \overline{y})$
1951	429,8	51,6	2.662,56	322,9	12,7	161,29	655,34
1952	350,1	—28,1	789,61	334,4	24,2	585,64	—680,02
1953	488,9	110,7	12.254,49	538,0	227,8	51.892,84	25.217,46
1954	392,8	14,6	213,16	316,2	6,0	36,00	87,60
1955	255,6	—122,6	15.030,76	193,9	—116,3	13.525,69	14.258,38
1956	281,4	—96,8	9.370,24	246,0	—64,2	4.121,64	6.214,56
1957	461,1	82,9	6.872,41	309,4	—0,8	0,64	—66,32
1958	394,8	16,6	275,56	335,8	25,6	655,36	424,96
1959	234,1	—144,1	20.764,81	319,7	9,5	90,25	—1.368,95
1960	208,6	—169,6	28.764,16	193,9	—116,3	13.525,69	19.724,48
1961	442,0	63,8	4.070,44	260,9	—49,3	2.430,49	—3.145,34
1962	228,3	—149,9	22.470,01	226,6	—83,6	6.988,96	12.531,64
1963	452,9	74,7	5.580,09	455,5	145,3	21.112,09	10.853,91
1964	241,5	—136,7	18.685,89	186,5	—123,7	15.301,69	16.909,79
1965	810,5	432,3	186.883,29	413,6	103,4	10.691,56	44.699,82
$\overline{x} =$	378,2		334.688,48	310,2	$= \overline{y}$	141.119,83	146.317,31

$$R = \frac{146.317,31}{\sqrt{334.688,48 \times 141.119,83}} = \frac{146.317,35}{217.327,35} = 0,67$$

La cifra de 0,67 corresponde a una correlación relativamente mediocre que, de todas maneras, debe ser probada con la ayuda de una serie más larga de valores. Es evidente que ninguna ley puede esbozarse con este único índice, pero el principio es válido y sólo falta ampliar la base de información a emplearse. Pédelaborde señala con razón que "una correlación no es más que una herramienta empírica, es decir imperfecta y provisoria". En todo caso, como instrumento de trabajo para el geógrafo puede proporcionar apoyo a interesantes comprobaciones y a puntos de vista novedosos.

4. El cálculo del coeficiente de correlación puede también realizarse determinando por anticipado la "varianza" de cada serie. Esta "varianza" o "fluctuación" muestra la dispersión de cada uno de los caracteres considerados, en un caso con relación a $\overline{x}$ y en el otro con respecto a $\overline{y}$. De su cálculo deriva la fijación de la "desviación tipo" correspondiente a cada columna.

En el caso del ejemplo anterior las varianzas y las desviaciones tipos son los siguientes:

i) *Precipitaciones en Valparaíso (1951-1965):*

$$\text{varianza } (\sigma x^2) \ : \ \frac{334688{,}48}{14} = 23906{,}32 \ (1)$$

$$\text{écart tipo } (\sigma x) \ : \ \sqrt{23906{,}32} \ = \ 154{,}61 \ (2)$$

ii) *Precipitaciones en Santiago (1951-1965):*

$$\text{varianza } (\sigma y^2) \ : \ \frac{141119{,}83}{14} = 10079{,}99$$

$$\text{écart tipo } (\sigma y) \ : \ \sqrt{10079{,}99} = \ 100{,}39$$

Paralelamente se debe establecer el valor de la "covarianza" (k^2) que para el caso anterior es igual a:

$$\frac{146317{,}31}{14} = 10451{,}23 \ (3)$$

---

(1) $\sigma_x^2 = \frac{1}{n-1} \Sigma u_i^2$ (La división se realiza teniendo como denominador el total de términos de la serie menos uno).

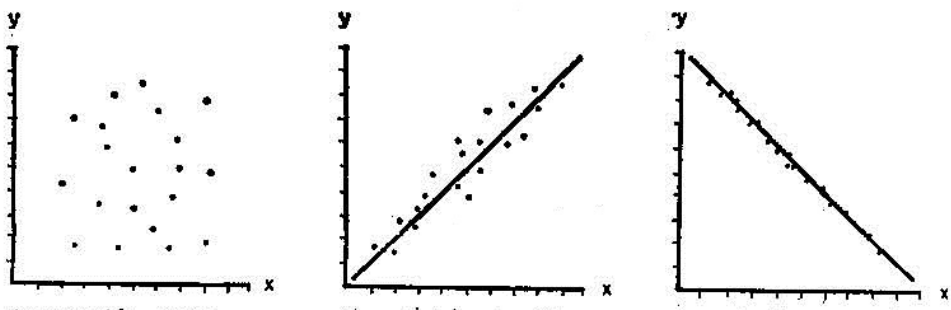
(2) $\sigma_x = \sqrt{\frac{1}{n-1} \Sigma u_i^2}$ (La desviación tipo -σ- es la raíz cuadrada de la varianza)

(3) covarianza (k^2) : $\frac{1}{n-1} \Sigma u_i v_i$

Con estos datos se puede llegar finalmente a la solución de la fórmula siguiente destinada a la determinación del coeficiente de correlación: $R = \frac{k^2}{\sigma x \, \sigma y}$; o sea:

$$R = \frac{10451{,}23}{154{,}61 \times 100{,}39} = \frac{10451{,}23}{15521{,}30} = 0{,}67$$

5. Gráficamente, estos cálculos permiten trazar las rectas de regresión que se inscriben sobre un fondo en el que se representan los datos utilizados mediante puntos. Dicha representación puede mostrar algunas variantes: o bien corresponde a una nube de puntos totalmente esparcidos (correlación nula), o bien los puntos se ordenan en torno a un eje central, de manera más o menos concentrada (fuerte o débil correlación).



Esquema Nº I

Las dos rectas de regresión (regresión de x con relación a y — regresión de y con relación a x) tendrán una disposición diferente en cada caso, siendo más abiertos los ángulos cuando la correlación sea más débil. Como la anota M. Péguy "es precisamente la dualidad de las rectas de regresión la que subraya el carácter no funcional de la relación que puede existir entre las dos variables" (4). Solamente en el caso en que las dos rectas de regresión se confunden en una sola línea (el coeficiente de correlación es igual a la unidad), esta recta común puede marcar una relación lineal de carácter funcional entre las dos variables.

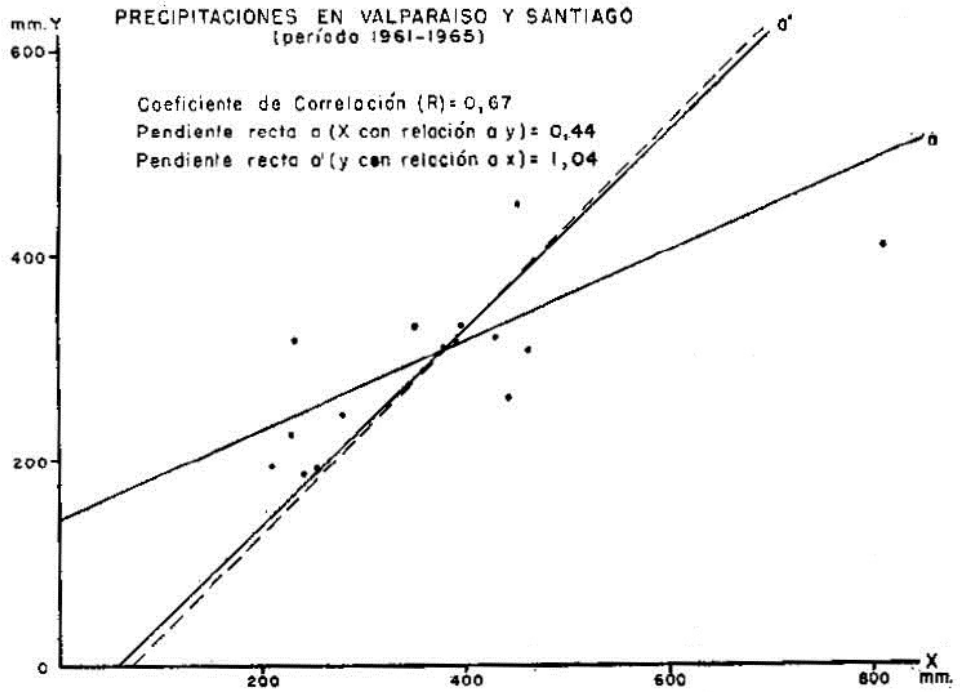
La pendiente de las rectas de regresión se obtiene aplicando:

$$a = \frac{\Sigma uv}{\Sigma u^2} \qquad a' = \frac{\Sigma uv}{\Sigma v^2}$$

En el ejemplo utilizado, los cálculos dan los resultados siguientes:

$$a = \frac{146317{,}31}{334688{,}48} = 0{,}44 \qquad a' = \frac{146317{,}31}{141119{,}83} = 1{,}04$$

(4) Charles-Pierre Péguy: *Elements de Statistique apliquée aux Sciences Géographiques;* Centro de Documentación Universitario, París, 1958.



Esquema Nº 2

El punto en que se cortan las dos rectas corresponde a la intersección de los promedios de las series utilizadas. Por él se puede hacer pasar también una recta de pendiente 1 que correspondería a la hipótesis de un gradiente pluviométrico constante, susceptible de ser expresado mediante la función $p_1 = f(p_2)$.

En el gráfico obtenido (esquema 2) el eje de las abscisas (o de las x) es el de la estación más baja (Valparaíso) y el eje de las ordenadas (o de las y) corresponde a la estación que se encuentra a mayor altura (Santiago).

BIBLIOGRAFIA

1. GERBE, ERIC. *Critique de la notion de gradient thermique moyen appliquée à une coupe des Monts de Tarare;* Facultad de Letras y Ciencias Humanas de Lyon (Instituto de Geografía), dactilografiado, sin fecha.
2. LIBAULT, ANDRÉ. *L'interpretation des valeurs numériques dans la recherche géographique;* Annales de Géographie Nº 320; mayo-junio, 1951.
3. PÉDELABORDE, PIERRE. *Remarques sur l'emploi de deux notions classiques en Climatologie: Les moyennes, les correlations;* Revue Géographique des Pyrénées et du Sud-Ouest; t. XXVIII, marzo 1957.
4. PÉGUY, CHARLES-PIERRE. *Introduction à l'emploi des méthodes stistiques en géographie physique;* Revue de Géographie Alpine; t. XXXVI, 1948.
5. PÉGUY, CHARLES-PIERRE. *Elements de Statistique appliquée aux Sciences Géographiques;* Centro de Documentación Universitaria, París, 1958.