Correlación, Simetría y Variabilidad

Raúl P. Mentz

Introducción

En muchos problemas de estadística aplicada, las observaciones numéricas disponibles para el análisis forman pares, un par de números asociado con cada individuo. Por ejemplo, podemos tener los pesos de niños antes y después de un tratamiento, calificaciones escolares al comienzo y al final de un experimento educativo, altura y peso de atletas, cantidades compradas y precios unitarios pagados por una persona en varias compras o por diferentes personas, etc. Una técnica útil del análisis de datos para conjuntos bidimensionales de la forma $\{(x_1, y_1),...,(x_n, y_n)\}$, es construir un gráfico de "puntos" $P_i = (x_i, y_i)$ en un sistema común de coordenadas cartesianas ortogonales, con sus correspondientes ejes x o de las abscisas e y de las ordenadas. Esta representación gráfica del conjunto de puntos o pares será llamado un *diagrama de dispersión*.

El diagrama de dispersión es adecuado para estudiar problemas de *correlación*. Más adelante definiremos el concepto de correlación para un diagrama de dispersión, pero previamente trataremos de desarrollar un sentido intuitivo para este concepto y para r, la medida de correlación. Una alternativa es leer la Sección II antes de la I.

I. Ejemplos
Significado de Correlación

Para ilustrar el significado de la correlación, se presenta a menudo un argumento gráfico como el siguiente. En el Gráfico 1 se muestran cinco diagramas de dispersión, cada uno con 100 puntos o pares. Cuando algunos pares coinciden se usa una marca mas gruesa. Los pares en la parte (a) del gráfico son aproximadamente no correlacionados, r es aproximadamente igual a cero. A medida que descendemos en el gráfico, la correlación aumenta, y en la parte (e) tenemos el valor extremo de r=1, los pares están todos en una recta, en este caso una que pasa por el origen del sistema de coordenadas.

El lector puede suponer que en todos los casos, a medida que los puntos se concentran alrededor de una recta, la correlación aumenta y se aproxima a 1. En primer lugar, si la concentración ocurre a lo largo de una recta de pendiente negativa [por ejemplo una que pasa por el origen y por el punto (x, y) = (-1, 1)], r no se aproxima a 1 sino a –1. Más importante aún es que si la concentración ocurre

con relación a una recta paralela al eje x (o coincidente con él), r no se aproxima a 1 o –1 sino que lo hace a 0.

Por lo tanto el signo y el grado de la correlación (y el valor de r) son diferentes según el diagrama de dispersión se concentre con relación a una recta o a otra, y debemos diferenciar las rectas con pendientes no nulas de aquellas con pendiente nula: pendiente 0 de la recta de concentración significa que r = 0. Si denotamos por b la pendiente de la "recta de concentración" (que se definirá con precisión en la Sección II), tenemos que en la parte (a) del Gráfico 1, r = b = 0, en la parte (e) r = + 1 con b > 0, mientras que en las partes intermedias de (b) a (d), r y b son ambos positivos y $0 < r < 1$.

Detalle Técnico. La parte (a) del diagrama de dispersión del Gráfico 1 se construyó con pares de números seudo-aleatorios, distribuidos uniformemente entre 0 y 1. Para cada par (x, y), su proyección a la línea y = x es ((x + y)/2, (x + y)/2), y ellos forman los pares de la parte (e). Los puntos de las rectas que unen ambos pares están dados por (ax + (1-a)(x + y)/2, ay + (1-a)(x + y)/2) para a entre 0 y 1. Por lo tanto las partes (a) hasta (e) corresponden a los valores a = 1, 0.75, 0.5, 0.25 y 0, respectivamente. A estos diagramas de dispersión corresponden los valores r = 0.03, 0.31, 0.62, 0.89 y 1 señalados en el gráfico.

Diagramas de Dispersión Simétricos

Un enfoque que se utiliza a menudo para ilustrar el sentido de la falta de correlación (r = 0) es recurrir a los diagramas simétricos. Los diagramas en el Gráfico 2 tienen r = 0.

Cualquiera de estos diagramas puede alterarse (haciéndolo más disperso) en una o ambas direcciones, y todavía corresponderá a r = 0. Por ejemplo, los diagramas de dispersión en el Gráfico 3 se obtuvieron del diagrama a la izquierda del Gráfico 2, y todos tienen r=0.

Sin embargo, la presencia de simetría no es fácil de detectar. Algunos de los diagramas de dispersión en el Gráfico 4 tienen r=0 mientras que otros tienen r<0 ó r>0. Por lo tanto se justifica analizar el tema de la simetría con más detención. Lo haremos después de introducir algo de teoría y una notación.

II. Notación

Para un conjunto de n pares de números, $\{(x_1, y_1), ..., (x_n, y_n)\}$, definimos el *coeficiente de correlación de Pearson* por

(1) $$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Donde $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \quad \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$.

Es útil representar a las sumas con la letra S. Entonces,

(2) $$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{S_{xy}}{S_x S_y},$$

donde $S_x = \sqrt{s_{xx}}$, $S_y = \sqrt{s_{yy}}$. Si a un conjunto de n pares ajustamos una recta por el llamado *método de los cuadrados mínimos*, la recta tendrá pendiente

(3) $$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

y ordenada al origen

(4) $$a = \bar{y} - b\bar{x}.$$

Mantenemos como supuesto que $S_{xx} \neq 0$ y $S_{yy} \neq 0$, lo que significa que los casos en que todas las x son iguales, o todas las y iguales, o ambos, están descartados. Bajo estas condiciones, está claro que r y b sólo pueden ser iguales a 0 en un caso, esto es cuando $S_{xy}=0$.

Nótese que S_{xx} y S_{yy} (o bien S_x y S_y que están, en valor absoluto, en las mismas unidades de las variables) miden la variabilidad, la *variabilidad marginal*, presente en las respectivas variables, mientras que S_{xy} mide la *variabilidad conjunta* de los

pares (x,y) y tiene como unidades el producto de las dos variables consideradas. Utilizando divisores adecuados, constituyen medidas estadísticas utilizadas frecuentemente: $S_x^2=S_{xx}/(n-1)$ y $S_y^2=S_{yy}/(n-1)$ son las *varianzas muestrales*, sus raíces cuadradas (no-negativas) con las *desviaciones estándares muestrales* y $S_{xy}=S_{xy}/(n-1)$ [o S_{xy}/n] es la *covarianza muestral.*

Propiedades de Invariancia

Es importante analizar cuestiones de *invariancia*. Primero consideremos el efecto de *traslaciones*, que convierten a x en x+p, a y en y+q, o a ambos simultáneamente. Estas traslaciones afectan de la misma manera a los promedios, esto es, la media de los valores trasladados x+p es $\bar{x} + p$, y la de los valores trasladados y+q es $\bar{y} + q$. Por lo tanto, los *desvíos* $x_i - \bar{x}$ *e* $y_i - \bar{y}$ no son afectados por las traslaciones. Dado que r y b fueron definidos en (1) y (3) en términos de sumas de estos desvíos, concluimos que *r y b son invariantes a las traslaciones.*

Sean ahora c y d constantes positivas. Dado que

$$(5) \qquad \sum_{i=1}^{n}(cx_i - c\bar{x})^2 = c^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = c^2 S_{xx},$$

$$\sum_{i=1}^{n}(dy_i - d\bar{y})^2 = d^2\sum_{i=1}^{n}(y_i - \bar{y})^2 = d^2 S_{yy},$$

$$\sum_{i=1}^{n}(cx_i - c\bar{x})(dy_i - d\bar{y}) = cd\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = cdS_{xy},$$

se deduce que el coeficiente de correlación de los pares (cx_i, dy_i) es $cdS_{xy}/(cS_xdS_y)=r$, mientras que la pendiente de la recta computada por el método de cuadrados mínimos se toma el valor $cdS_{xy}/(c^2S_{xx})=(d/c)b$.

La conclusión es que r es *invariante a las transformaciones (positivas) de escala*, mientras que b resulta afectado de la manera indicada. Una manera interesante de enfatizar estos resultados es notar que en el análisis de correlación para diagramas de dispersión como los presentados en la Sección I, no es necesario referir las observaciones a un sistema de coordenadas, pues las traslaciones y los cambios (positivos) de escala no tienen efecto sobre el coeficiente de correlación. Cuando se

estudian las rectas por cuadrados mínimos, se deben considerar o trazar los ejes coordenados, para recordar que las elecciones de escala son importantes.

Las restricciones c>0 y d>0 son importantes: si una de estas constantes es positiva y la otra negativa, los signos de r y b cambian, mientras que si ambas son positivas o ambas negativas, r y b no cambian de valor ni de signo.

III. Cómputos

Para analizar con más detalle los ejemplos de la Sección I, presentamos a continuación ejemplos numéricos. El diagrama a la izquierda del Gráfico 4 (o del Gráfico 5) puede considerarse generado por los pares (z_i, y_i) de la Tabla 1.

Tabla 1. Ejemplo numérico con 5 puntos bidimensionales.

1 Abscisas z_i	2 Ordenadas Y_i	3 Abscisas Centradas $x_i=z_i - \bar{z}$	4 Productos x_iy_i	5 Cuadrados x^2_i
1	1	-2.5	-2.5	6.25
2	-1	-1.5	1.5	2.25
3	1	0	0	0
4	-1	1.5	-1.5	2.25
5	1	2.5	2.5	6.25
Suma = 15	1	0	0	17.00

Nótese que la última fila contiene las sumas de las columnas.

Teniendo en cuenta la invariancia a las traslaciones, operamos con $x_i = z_i - \bar{z}$ en vez de hacerlo con z_i. Como la suma de los x es igual a 0, también lo es su promedio. Dado que los y son muy simples operamos sin modificarlos, siendo su promedio 1/5. Tenemos que

$$(6) \qquad S_{xy} = \sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{5}(x_i - \bar{x})y_i = \sum_{i=1}^{5}x_i y_i = 0,$$

la segunda igualdad se debe a que $\sum(x_i - \bar{x})\bar{y} = \bar{y}\sum(x_i - \bar{x}) = 0$, y la

tercera a que $\bar{x} = 0$, lo que conduce a que $S_{xy}=\sum x_i y_i$. También obtenemos $S_{xx}=17$. En conclusión,

$$(7) \qquad r = \frac{S_{xy}}{S_x S_y} = 0, \quad b = \frac{S_{xy}}{S_{xx}} = 0,$$

como se dijo. Sin embargo consideremos ahora el diagrama a la derecha del Gráfico 5 (o el segundo contando desde la izquierda en el Gráfico 4); los datos aparecen en la Tabla 2.

Tabla 2. Ejemplo numérico con 6 puntos bidimensionales

1 Abscisas z_i	2 Ordenadas y_i	3 Abscisas Centradas $x_i=z_i - \bar{z}$	4 Productos $x_i y_i$	5 Cuadrados x^2_i
1	1	-2.5	-2.5	6.25
2	-1	-1.5	1.5	2.25
3	1	-0.5	-0.5	0.25
4	-1	0.5	-0.5	0.25
5	1	1.5	1.5	2.25
6	-1	2.5	-2.5	6.25
Suma 21	0	0	-3.0	17.50

Ahora calculamos

$$(8) \qquad r = \frac{S_{xy}}{S_x S_y} = \frac{-3}{\sqrt{17.50 x 6}} = -0.29, \qquad b = \frac{S_{xy}}{S_{xx}} = -0.17,$$

mientras que $a = \bar{y} - b\bar{x} = 0$. Diagramas correspondientes a las tablas 1 y 2 forman el Gráfico 5.

A medida que aumenta la cantidad de pares de datos del tipo que estamos analizando, disminuyen b y r. Por ejemplo, podemos controlar que para conjuntos de pares como los de las tablas 1 y 2, se cumple lo siguiente:

n=4 pares, b=-0.4, r=-0.44,
n=6 pares, b=-0.17, r=-0.29 (como se vio en (8))
n=50 pares, b=-0.0024, r=-0.00346.

IV. Diagramas de Dispersión Simétricos

Definición Operativa. Un diagrama de dispersión con n puntos o pares bidimensionales (x_i , y_i) (donde $n \geq 1$) se dice que es *simétrico* si: (a) contiene n_1 puntos ($0 \leq n_1 \leq n$) para los cuales $(x_i - \bar{x})(y_i - \bar{y}) = 0$; y (b) los restantes $2n_2$ puntos (donde $0 \leq n_2 \leq n/2$) aparecen en pares, (x_j, y_j) y (x_k, y_k) en los que $(x_j - \bar{x})(y_j - \bar{y}) = -(x_k - \bar{x})(y_k - \bar{y}) \neq 0$.

Los n_1 puntos cuyos productos cruzados son iguales a 0 pertenecen a las líneas $x = \bar{x}$ *o* $y = \bar{y}$, o a ambas si ellos coinciden con el par de promedios $(\bar{x}, \bar{y})$. Estas líneas son paralelas a los ejes coordenados, o coinciden con ellos.

Una consecuencia de la definición operativa es la siguiente:

Proposición. Un diagrama de dispersión simétrico tiene $S_{xy}=0$, y por lo tanto r=b=0.

Esto proviene directamente de la definición de S_{xy} dada, por ejemplo, en (6).

Proposición. Existen diagramas de dispersión que no satisfacen la condición de simetría de la definición operativa y sin embargo tienen $S_{xy}=0$.

Ejemplo. Para n=3, el conjunto de puntos {(-2, 2), (-1, -1), (3, 1)} tiene esta propiedad.

Generación de Diagramas de Dispersión Simétricos.

En esta sección analizamos cómo generar diagramas de dispersión simétricos. De

acuerdo con nuestras observaciones sobre invariancia, en los siguientes diagramas omitimos los ejes coordenados. Considere los diagramas siguientes:

(9)

```
X      X X     X X     X X X     X X X X
X       X       X X       X
X       X
X
```

Ellos cumplen con las condiciones de la definición operativa y son por lo tanto diagramas de dispersión simétricos. Corresponden a las *particiones* de 4 con los enteros 1, 2, 3 y 4,

(10) 1+1+1+1 2+1+1 2+2 3+1 4

Euler demostró que el número p(n) de maneras de representar un entero positivo n como la suma de enteros positivos (sin considerar el orden) es igual al coeficiente de q^n en la expansión en serie de potencias del producto infinito

$$(11) \qquad \prod_{k=1}^{\infty} \frac{1}{1-q^k} = 1 + q + 2q^2 + 3q^3 + 5q^4 + 7q^5 + \ldots,$$

(Bressoud and Propp, 1999). Estos autores usan lo que llaman *diagramas de Young* para representar las particiones. En nuestro caso ellos son,

(12)

De los diagramas de dispersión simétricos presentados en (10), podemos generar otros por permutación, con lo que obtenemos lo siguiente:

(13)

```
X      X X     X      X      X X   X X X    X     X X XX
X       X     X X     X      X X     X    X X X
X       X      X     X X
X
```

1+1+1+1 2+1+1 1+2+1 1+1+2 2+2 3+1 1+3 4

Proposición. La cantidad s(n) de diagramas de dispersión simétricos generados por las permutaciones de las particiones es igual a 2^{n-1}.

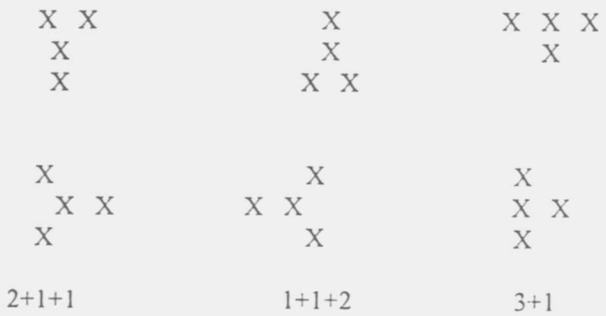
Esto es válido por ser s(n) la cantidad de soluciones de las ecuaciones

(14) $$x_1+x_2+...+x_k = n$$

para k=1,2,...,n, cuando las soluciones son enteros de 1 a n. Para cada k la cantidad de soluciones es $\binom{n-1}{k-1}$ (Niven, 1965, Capítulo 5). Por lo tanto

(15) $$s(n) = \sum_{k=1}^{n} \binom{n-1}{k-1} = \sum_{k=0}^{n-1} \binom{n-1}{k} = 2^{n-1}.$$

Observamos que algunos (pero no todos) los diagramas de dispersión de (13) pueden escribirse con orientaciones distintas, y todavía retener la propiedad de simetría. En efecto, tenemos lo siguiente:



Sin embargo no podemos hacer algo semejante con 1+2+1 o con 2+2.

V. Conclusiones

Hemos analizado algunas propiedades de la correlación. El elemento básico de nuestro enfoque es el *diagrama de dispersión* de un conjunto de pares o puntos $P_i=(x_i, y_i)$ en el espacio bidimensional.

La correlación se mide con relación a rectas: si los pares están en una curva, aún siendo la relación (no lineal) aparente, su correlación (lineal) puede ser baja, incluso igual a 0, cuando utilizamos el coeficiente de correlación de Pearson definido en (1). Por lo tanto la *correlación es lineal* aún cuando omitimos el calificativo.

Cuando medimos el signo y el grado de la correlación con el coeficiente r de Pearson, como se hace habitualmente en la práctica, las traslaciones y los cambios de escala son irrelevantes, excepto que si multiplicamos a los x con c y a los y con d, y se satisface que $cd<0$, ocurre un cambio de signos. La conclusión es que los diagramas de dispersión, cuando se analiza la correlación, pueden dibujarse sin los ejes coordenados. Desde este punto de vista la correlación no está relacionada con la variabilidad marginal de x o de y por separado, sino que sólo está relacionada con la variabilidad conjunta medida por S_{xy}, la suma de productos de las desviaciones con relación a los promedios. El coeficiente r es S_{xy} estandarizado por el producto de las desviaciones estándares, de manera que sus valores posibles están entre -1 y $+1$.

En consecuencia hemos analizado un conjunto de conexiones entre la correlación y su medida r, con la variabilidad que presenta el diagrama de dispersión. Hemos enfatizado que ciertas aseveraciones con respecto a r deben relacionarse con el valor de la pendiente b.

A continuación relacionamos correlación con simetría. Encontramos que los diagramas de dispersión simétricos (según una "definición operativa") tienen $r=0$, pero que la recíproca no es cierta, podemos tener $r=0$ en diagramas de dispersión no simétricos. También analizamos ejemplos de diagramas de dispersión que se alejan levemente de la simetría.

Finalmente desarrollamos un procedimiento simple para generar diagramas de dispersión simétricos. Este procedimiento revela una relación interesante con la teoría matemática de los números (enteros): los diagramas de dispersión simétricos están relacionados con particiones de enteros como sumas de enteros positivos, y a soluciones en enteros de ciertas ecuaciones. Ilustramos el procedimiento sistemático para generar 2^{n-1} diagramas de dispersión simétricos de n puntos. Dos de ellos, correspondientes a la partición de un entero n como $1+1+...+1$ o bien como n, tienen $S_{xx}=0$ y $S_{yy}=0$, respectivamente, de manera que r no está definido para ellos; todos los restantes tienen $S_{xy}= r =0$.

Bibliografía

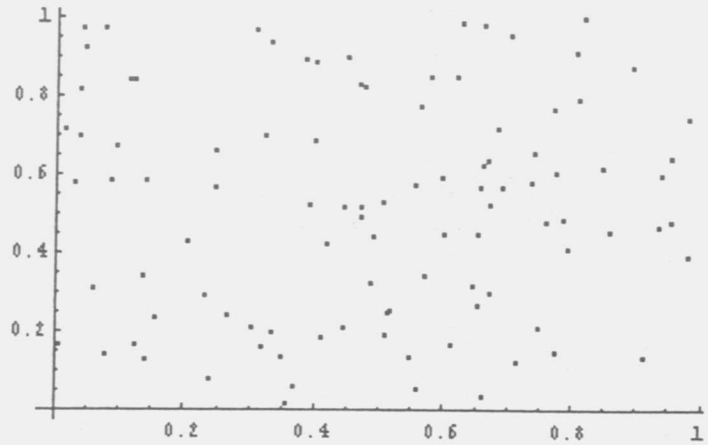
Bressoud, David and James Propp (1999), How the alternating sign matrix

Conjecture was solved, *Notices of the American Mathematical Society*, Vol. 46, No. 6, pp. 637-646.

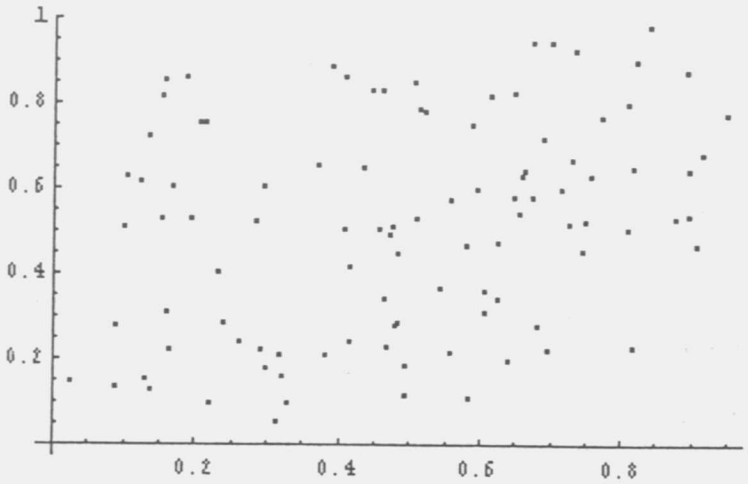
Niven, Ivan (1965), *Mathematics of Choice*. New York: Random House, the L. W. Singer Company.

Grafico 1. Los diagramas de dispersión tienen 100 puntos o pares. El diagrama a) tiene r aproximadamente igual a 0, el diagrama e) tiene r exactamente igual a 1. Los puntos del diagrama a) se proyectan a la recta y = x; b), c), y d) representan pasos sucesivos de la aproximación de cada par (x,y) a su proyección ((x+y)/2,(x+y)/2).

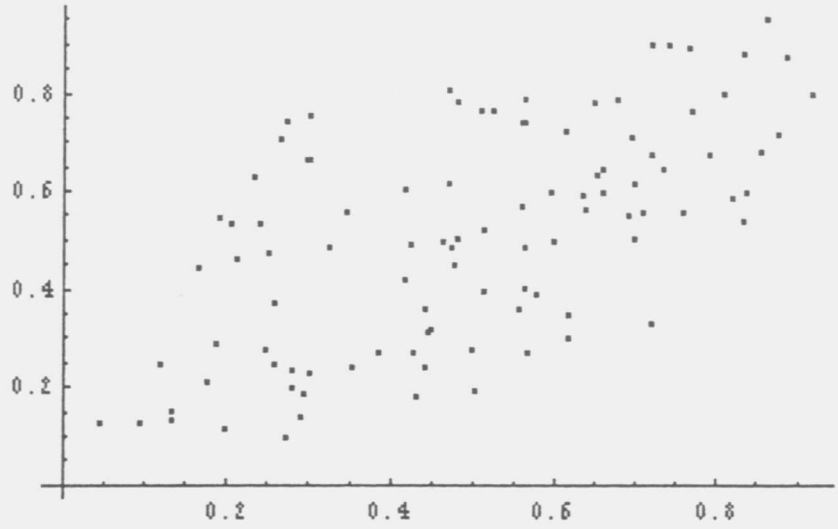
a) r = 0,03



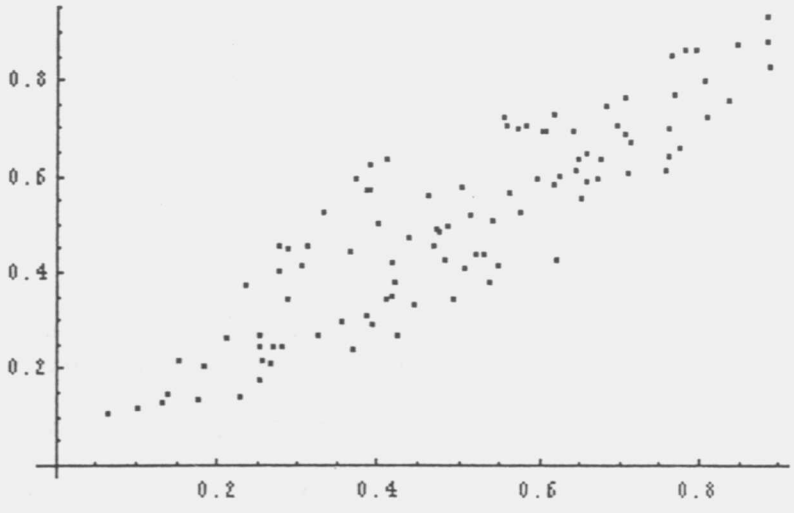
b) r = 0,31



c) r = 0,62



d) r = 0,89



e) r = 1,00

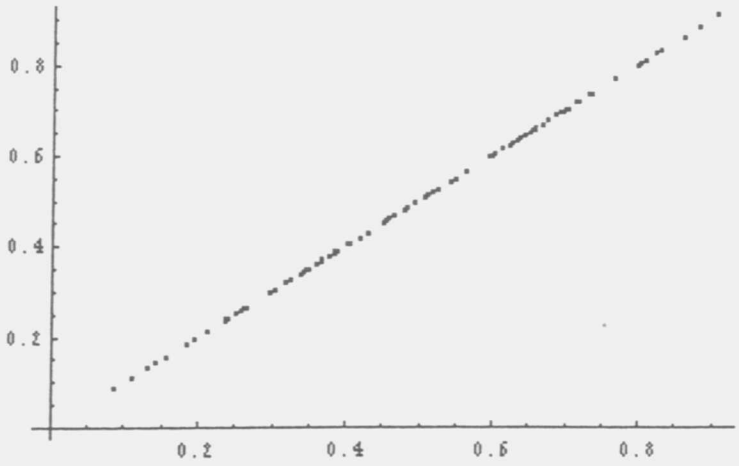


Gráfico 2. Diagramas de dispersión que tienen r=0.

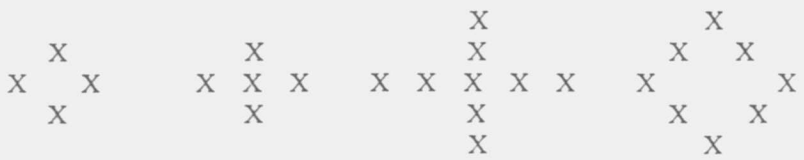


Gráfico 3. Diagramas de dispersión con r=0 deducidos del primer diagrama del Gráfico 1.

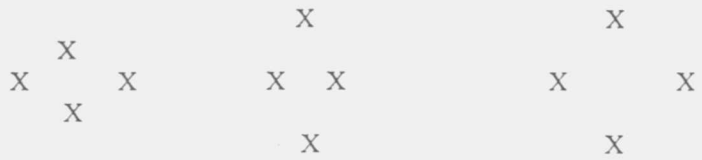


Gráfico 4. Diagramas de dispersión que muestran que un cambio de un solo punto puede afectar el signo y el valor de r, en particular, que sea igual a 0 o que no lo sea.

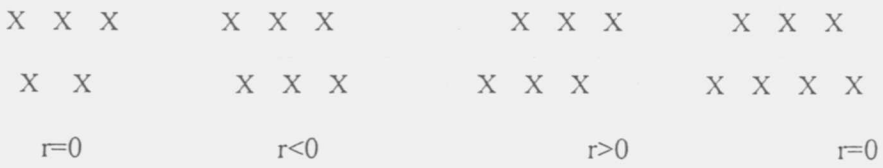
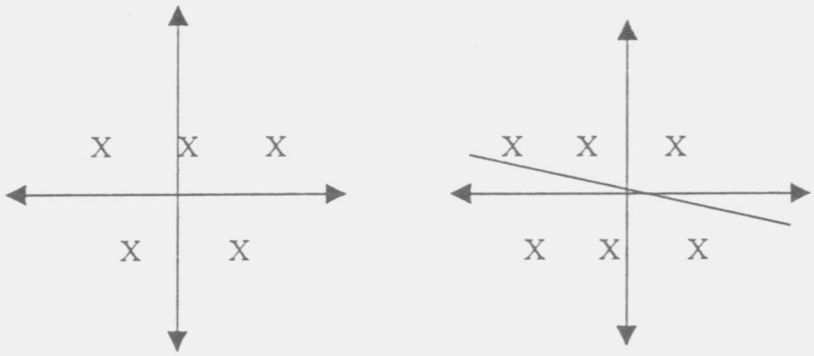


Gráfico 5. Izquierda: Pares de la Tabla 1, r=b=0, la recta por cuadrados mínimos es el eje de las abscisas. Derecha: Pares de la Tabla 2, r=-0.29, b=-0.17 con la recta calculada.



Universidad Nacional deTucumán y CONICET.
Casilla de Correo 209. Tucumán (4000) – Argentina.
Fax 54 (0381) 436-4105.
Email: rmentz@herrera.unt.edu.ar