Una interpretación geométrica de los estadísticos uni y multivariados

POR

JOSE MANUEL SERRANO GONZALEZ-TEJERO
MANUEL ATO GARCIA

RESUMEN

En consonancia con el enfoque propuesto por Hays (1980), se presenta aquí una interpretación geométrico-vectorial de los principales estadísticos univariados y bivariados, de carácter paramétrico (medidas de posición, escala, correlación y regresión), cuya comprensión es imprescindible para abordar el estudio de las técnicas de investigación y análisis multivariados.

ABSTRACT

This paper follows the suggestions proposed by Hays (1980) showing a complete geometrical interpretation of the esential statisties from an intervale-scale measurement and starting the basic assumptions to a further analysis of multivariate desing and data analysis.

INTRODUCCION

Dado un conjunto de 'n' pares de valores (X_i, Y_i), existen dos formas de representación gráfica. La primera y más natural, que hace hincapié en la noción de *individuo*, consiste en colocar los 'n' puntos que representan a los 'n' individuos en el espacio bidimensional determinado por las variables X e Y (fig. A). La segunda y más estructural, que hace hincapié en la noción de *variable*, consiste en utilizar un espacio n-dimensional en donde los 'n' valores de cada una de las dos variables representan dos vectores de ese espacio n-dimensional:

$$X_i = (x_1, x_2, \ldots, x_i, \ldots, x_{n-1}, x_n) \qquad \text{(fig. B)}$$
$$Y_i = (y_1, y_2, \ldots, y_i, \ldots, y_{n-1}, y_n) \qquad \text{e}$$

Es fácilmente deducible que, para los estudios de dependencia o correlación, la segunda representación es la más adecuada y, por tanto, la que de ahora en adelante trataremos de desarrollar.

INTERPRETACION GEOMETRICA DE LA MEDIA ARITMETICA Y DE LA DESVIACION STANDARD (O TIPICA)

Consideremos un eje I en el que las coordenadas de cualquiera de sus puntos sean iguales, es decir, un eje tal que para todo vector (X) situado sobre él:

$$X = (x_1, x_2, \ldots, x_i, \ldots, x_{n-1}, x_n)$$

se cumpla que:

$$x_1 = x_2 = \ldots = x_i = \ldots = x_{n-1} = x_n.$$

Este vector representaría una *variable estadística degenerada* (o constante) puesto que todos los individuos presentan los mismos valores en la variable X (fig. C).

Si la distancia del extremo del vector, X_0, a x_1 la representamos por m, valdrá m, también, la distancia de X_0 a x_2, la de X_0 a x_i, etc. y por tanto, la distancia OX_0 (o módulo del vector) vendría dada por

$$D(OX_0) = \sqrt{m^2 + m^2 + \ldots + m^2 + \ldots + m^2 + m^2}$$

es decir

$$D(OX_0) = \sqrt{n.m^2} = \sqrt{m^2}\,\sqrt{n} = m.\sqrt{n}$$

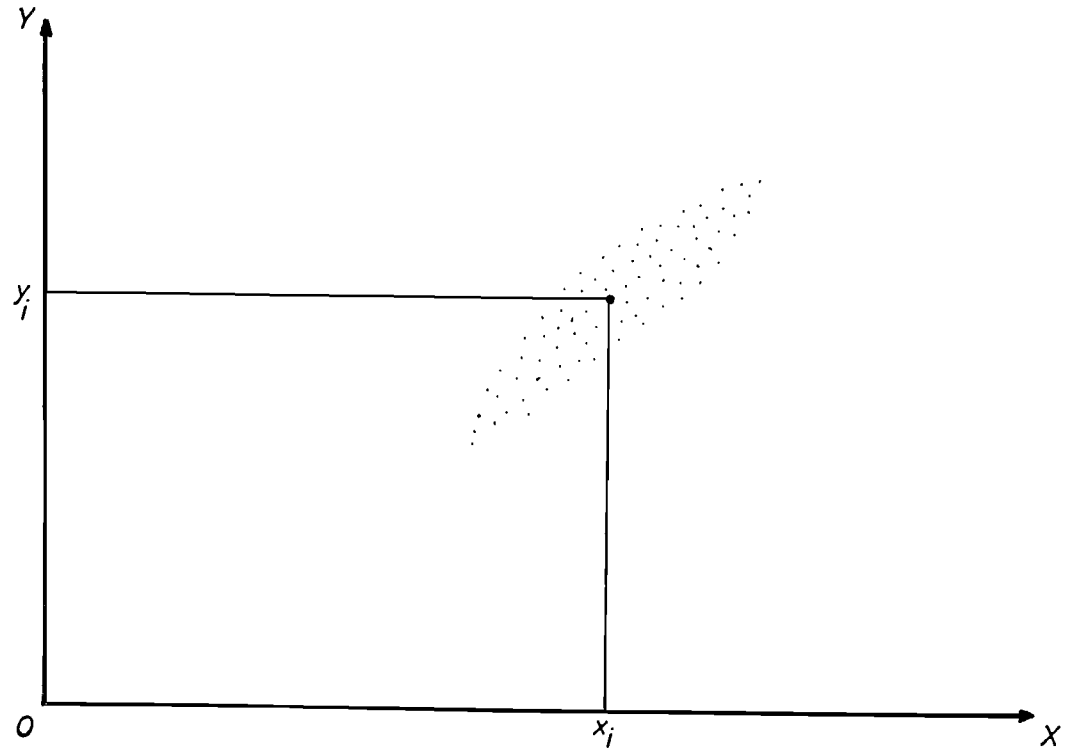


Fig A

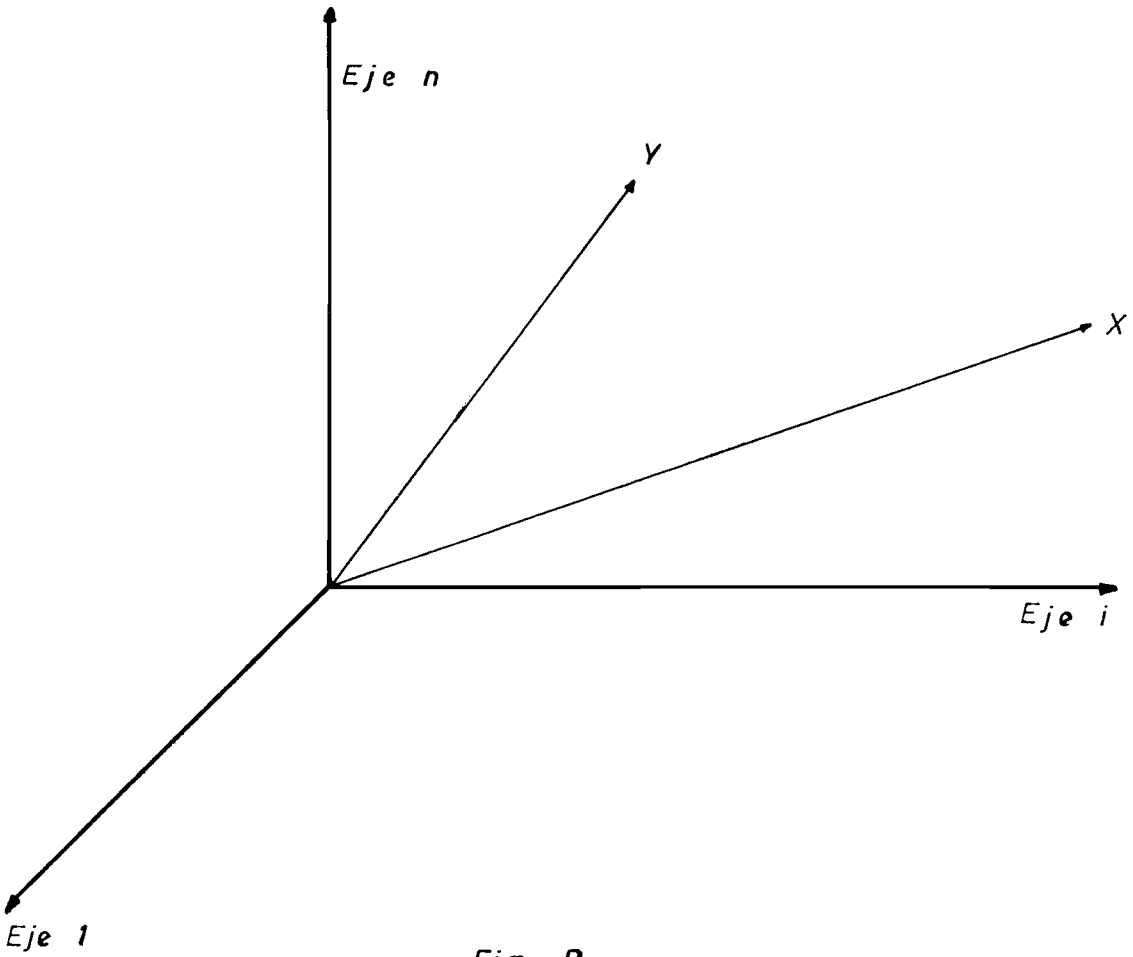


Fig B

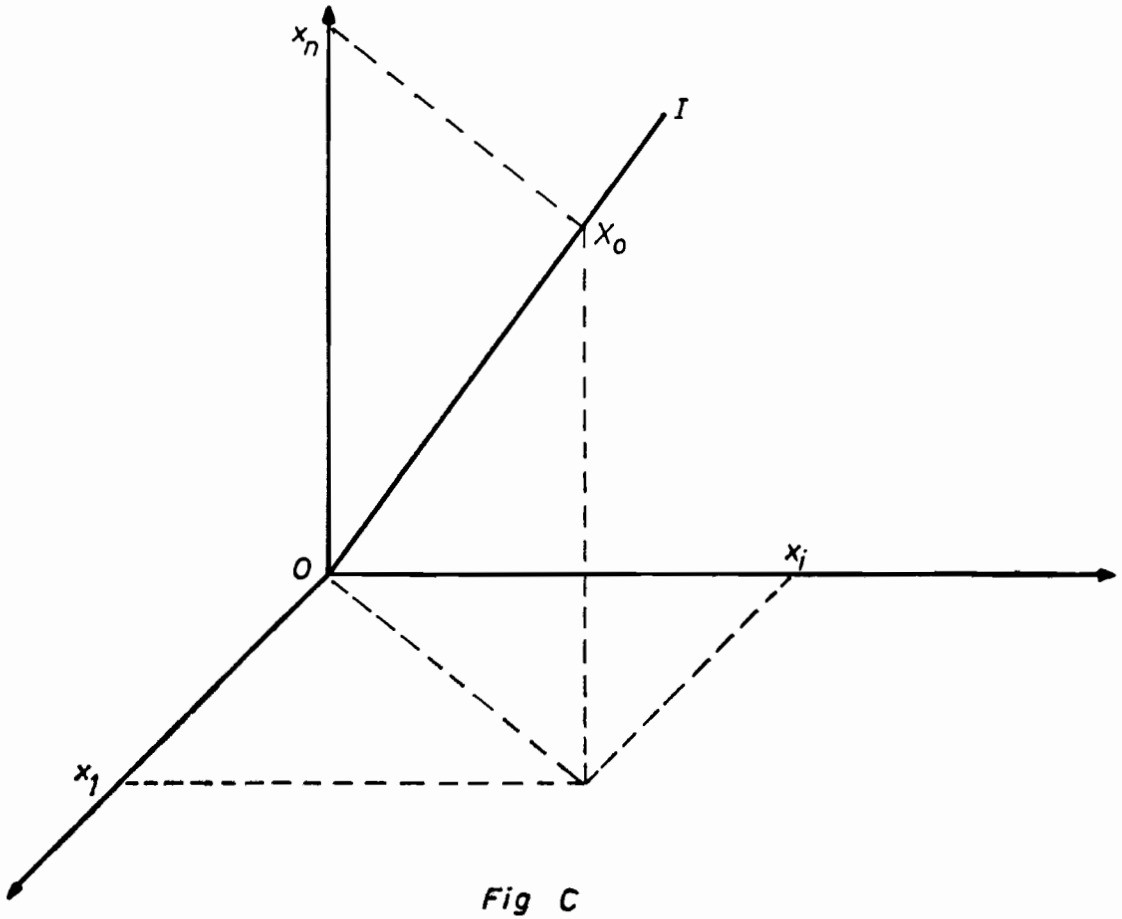
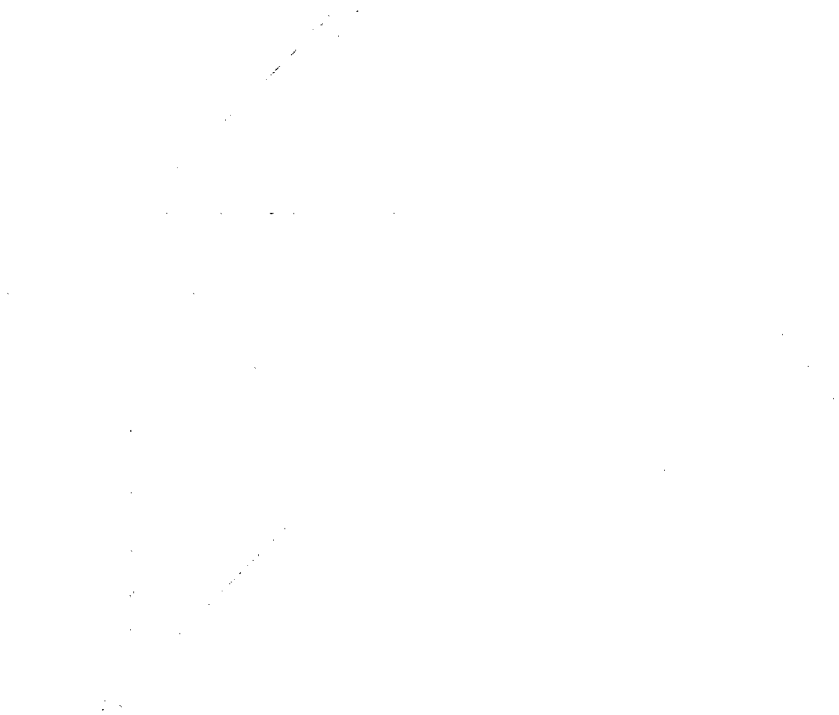


Fig C



Consideremos ahora otro vector X, distinto del anterior, y por tanto, de coordenadas distintas. Si proyectamos este nuevo vector sobre el eje I, obtendríamos un vector X' (OX) de coordenadas iguales en X' y si designamos por k estas coordenadas idénticas, el cuadrado de la distancia euclídea entre los puntos X' y X vendría dada por

$$\left[D(XX')\right]^2 = (x_1 - k)^2 + (x_2 - k)^2 + \ldots + (x_i - k)^2 + \ldots + (x_{n-1} - k)^2 + (x_n - k)^2,$$

luego

$$\left[D(XX')\right]^2 = \sum_{i=1}^{n} (x_i - k)^2.$$

Ahora bien, esta distancia será mínima cuando $k = \bar{X}$, puesto que

$$\sum_{i=1}^{n} (x_i - k)^2 \text{ será mínima}$$

para aquellos valores que anulando la primera derivada hagan positiva la segunda y, en efecto:

Desarrollando el sumatorio tendremos

$$\sum_{i=1}^{n} (x_i^2 - 2.x_i.k + k^2)$$

y teniendo en cuenta que el sumatorio de una suma algebráica es la suma algebráica de los sumatorios, la expresión anterior tomaría la siguiente forma

$$\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2.x_i.k + \sum_{i=1}^{n} k^2 -$$

Como las constantes pueden salir fuera del signo de sumar y el sumatorio de una constante es 'n' veces la constante, la expresión anterior quedaría como sigue

$$\sum_{i=1}^{n} x^2{}_i - 2.k \sum_{i=1}^{n} x_i - n.k^2$$

y derivando la función con respecto a k obtendremos

$$\frac{\delta}{\delta k} = O - 2 \sum_{i=1}^{n} x_i + 2.n.k = -2 \sum_{i=1}^{n} x_i + 2.n.k$$

derivando nuevamente con respecto a k, la segunda derivada sería

$$\frac{\delta^2}{\delta k^2} = O + 2.n = 2.n$$

por lo que al ser la segunda derivada positiva puesto que $n > O$ la función sería mínima para el valor de k que anule la primera derivada. Haciendo, por tanto, la primera derivada igual a O, tendríamos

$$-2 \sum_{i=1}^{n} x_i + 2.n.k = O$$

sacando el 2 factor común en el primer miembro de la igualdad

$$-2 \left(\sum_{i=1}^{n} x_i - n.k \right) = O$$

dividiendo los dos miembros por —2

$$\sum_{n}^{i=1} x_i - n.k) = O$$

transponiendo términos y multiplicando por —1

$$n.k = \sum_{i=1}^{n} x_i$$

despejando k para calcular su valor

$$k = \frac{\sum_{i=1}^{n} x_i}{n}$$

luego la expresión sería mínima para $\bar{k} = \bar{X}$,

Si $k = \bar{X}$, se deduce que el punto X' tiene por coordenadas $(\bar{x}, \bar{x}, \ldots, \bar{x}, \ldots, \bar{x}, \bar{x})$.

La longitud OX' (equivalente a OX_0, calculada anteriormente) valdría, por tanto, $\bar{x}\sqrt{n}$ y la distancia XX' vendría dada por

$$\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Pero sabemos que la desviación típica de una distribución de frecuencias viene dada por la expresión:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

en donde el numerador del segundo miembro es D(XX'), por tanto, se podría expresar esta distancia mediante el producto $s_x \sqrt{n}$ fig. D

Tenemos así una interpretación geométrica de la media y la desviación típica de un conjunto de 'n' observaciones:

"Sobre el coeficiente $\sqrt{n}$, la distancia entre el vector X y el eje I de variables estadísticas degeneradas es la *desviación típica*

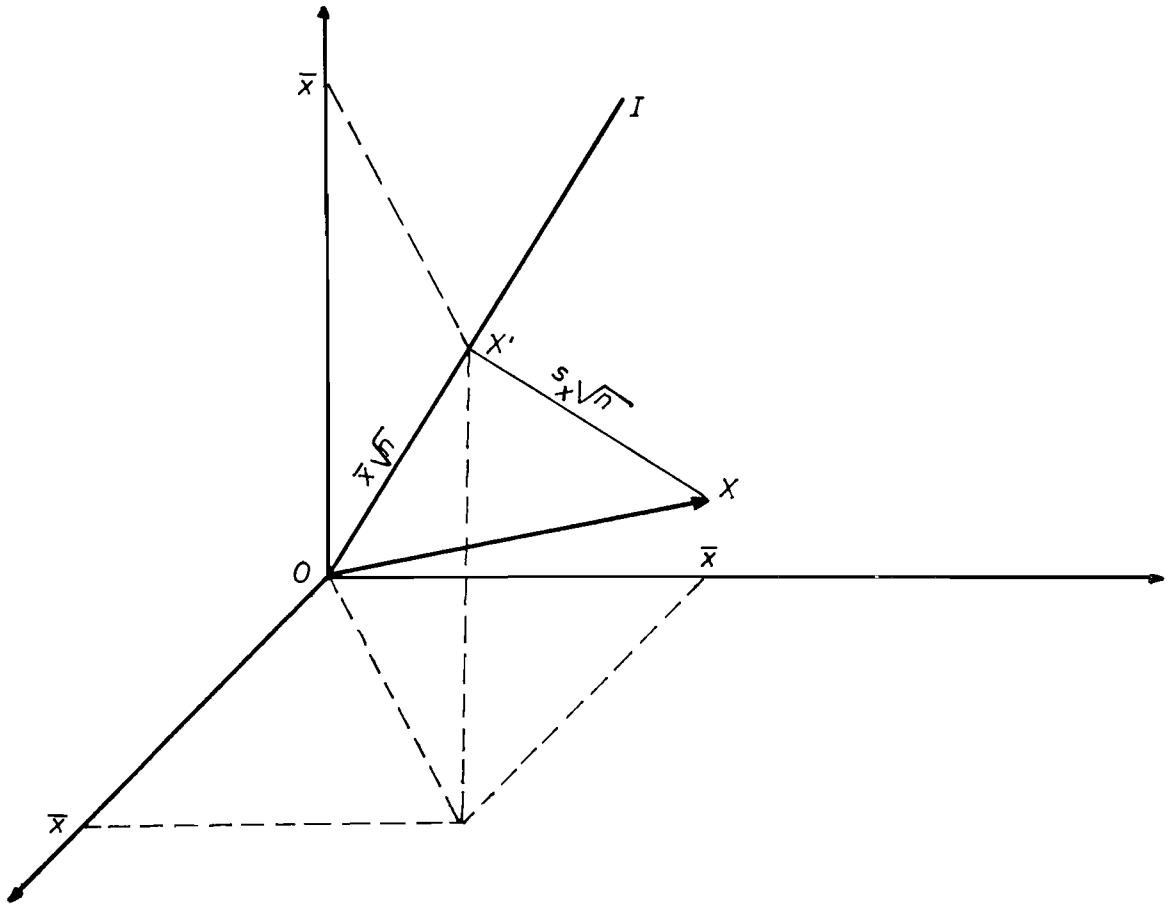


Fig D

y la distancia OX' es la *media* de la distribución de frecuencias considerada".

TEOREMA DE KONIG

Si consideramos un punto M del eje I cuyas coordenadas son iguales a k, la distancia OM es igual a

$$k\sqrt{n}$$

y la distancia MX' es igual a

$$|k - \bar{x}|\sqrt{n} \qquad \text{fig. E y F}$$

Aplicando el teorema de Pitágoras al triángulo MXX' de la fig. F, tenemos:

$$\overline{MX}^2 = \overline{XX'}^2 + \overline{MX'}^2$$

es decir,

$$\sum_{i=1}^{n} (x_i - k)^2 = n.s^2_x + n\,(k - \bar{x})^2$$

y dividiendo por 'n' los dos miembros de esta ecuación tendremos la expresión del teorema de König:

$$_km_2 = s^2_x + (k - \bar{x})^2$$

cuando k = X, el momento de orden dos respecto a k (la media) es la varianza.

"Sobre el coeficiente $\sqrt{n}$, el cuadrado construido sobre la distancia entre el vector X y el eje I de variables estadísticas degeneradas es la *varianza* de la distribución de frecuencias considerada".

INTERPRETACION GEOMETRICA DE LA COVARIANZA

Supongamos dos ejes OX y OY sobre los que se sitúan respectivamente, las desviaciones típicas de dos variables X e Y. Sea

$$OA = s_x \text{ y } OB = s_y \qquad \text{fig. 1}$$

siendo φ el ángulo que forman los mencionados ejes. El valor máximo de la suma de los vectores s_x y s_y vendría dado, exclusiva-

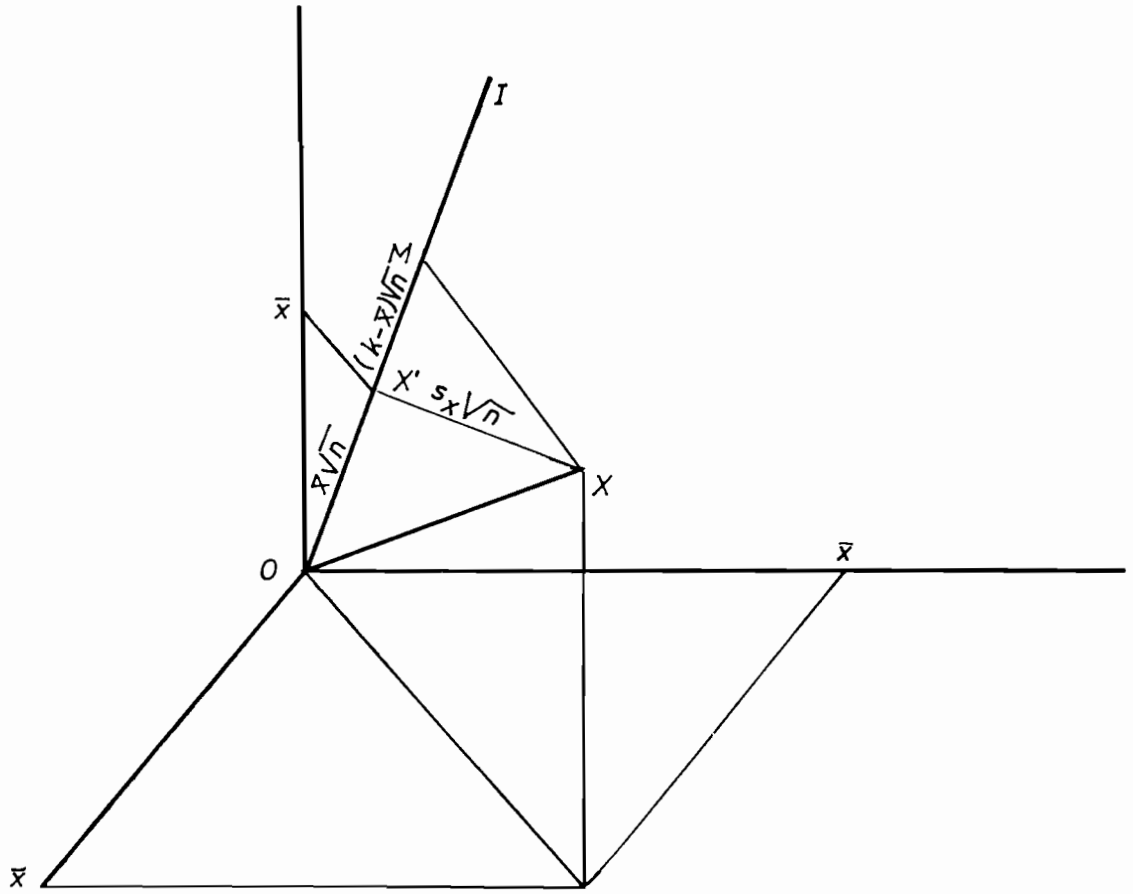


Fig E

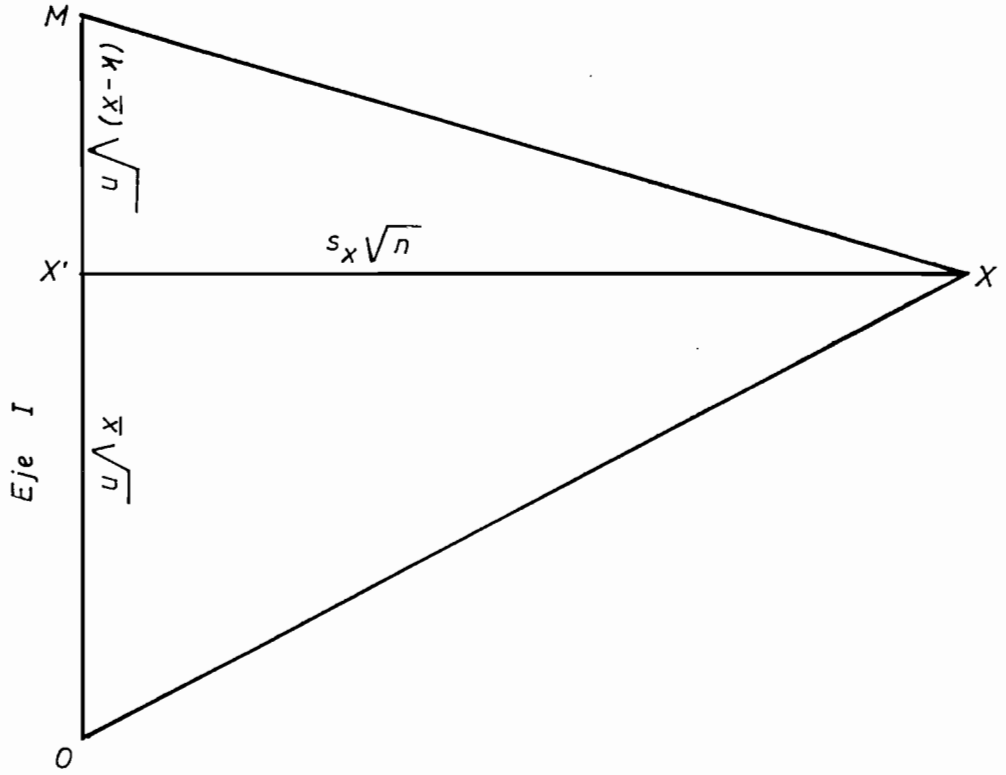


Fig F

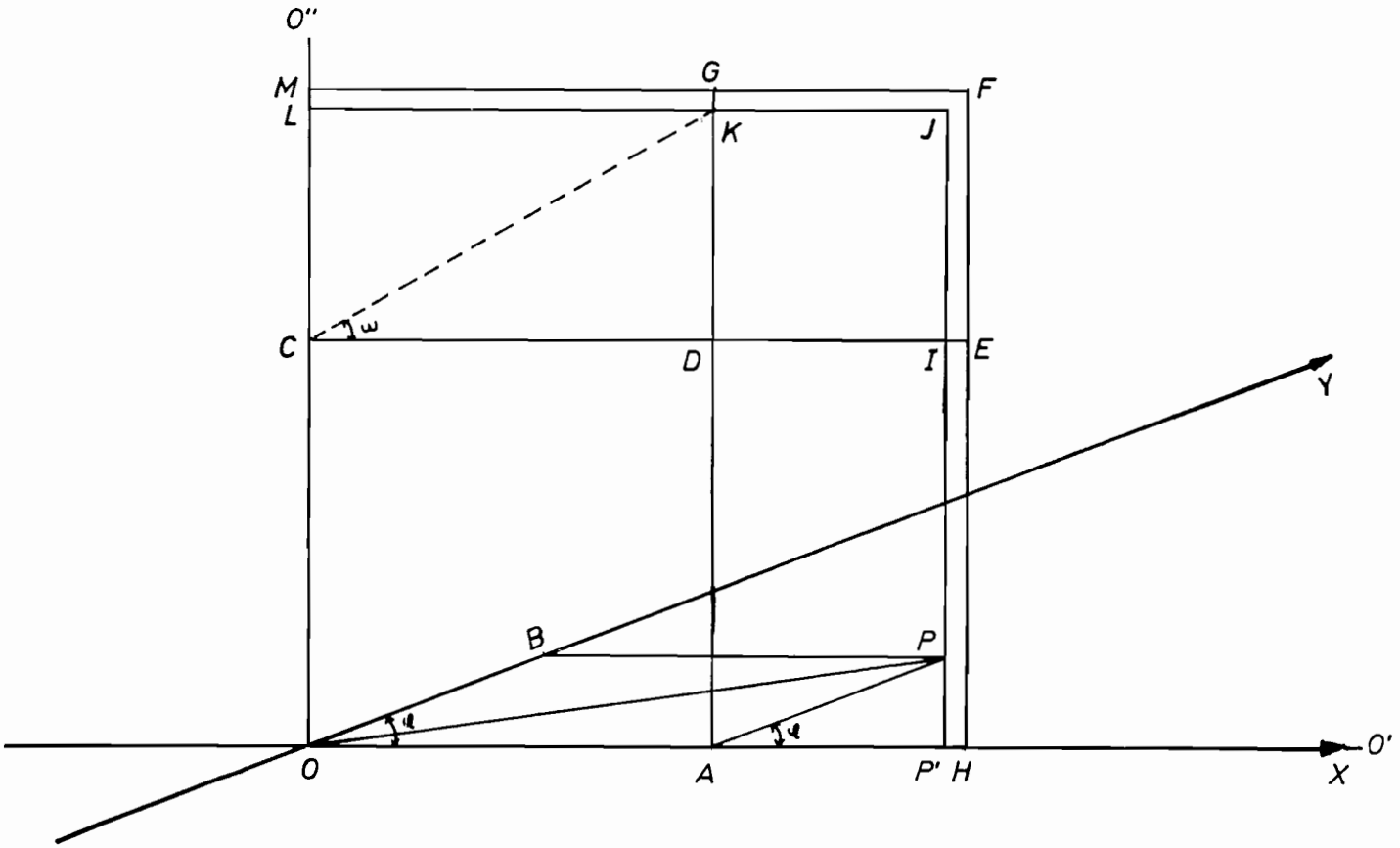


Fig 1

mente, en el caso en que ambos ejes tuvieran la misma dirección. La suma vectorial se realizaría llevando, por ejemplo, el origen de OB sobre el extremo de OA, con lo que tendríamos OA + OB = OH, en donde AH sería igual a OB.

En nuestro caso, las direcciones de los ejes OX y OY no son coincidentes, puesto que, hemos partido del hecho (o supuesto) que $\varphi \neq 0$. La suma OA + OB no sería, por tanto, la suma máxima, y si pensamos que una desviación refleja en cierto modo un desplazamiento de las puntuaciones con respecto al centro de gravedad de una distribución de frecuencias, no tendremos ningún obstáculo en asimilar la desviación típica a un vector, puesto que éstos están totalmente representados mediante un desplazamiento. Observando nuestra figura, podemos darnos perfecta cuenta que desplazarse OA y a continuación OB (o viceversa) supone desplazarse OP, por lo que tendremos:

$$OA + OB = OP$$

Imaginemos ahora, unos ejes de coordenadas rectangulares (ortogonales), OO' para las abscisas y OO" para las ordenadas. Hagamos coincidir uno de los ejes donde se presentan las desviaciones típicas de las variables, el eje X por ejemplo, sobre el eje de abscisas OO', proyectando sobre él, tanto la suma máxima como la real. La proyección de OH por situarse sobre el propio eje de proyección sería OH y la proyección de OP vendría dada por OP' puesto que O se encuentra sobre el eje de proyección (recordemos que la proyección de un segmento sobre un eje es el segmento determinado por la proyección de sus puntos extremos).

Al operar de esta forma, sobre el eje OO' encontramos dos desviaciones-suma, una, la máxima, representada por OH y otra, la real en nuestro caso, representada por OP'; a ambas desviaciones-suma les corresponderán otras tantas varianzas-suma, que estarían determinadas por las áreas de los cuadrados construidos sobre ellas, es decir, las varianza-suma máxima vendría determinada por el área del cuadrado OHFM y la varianza-suma real vendría determinada por el área del cuadrado OLJP'.

El cuadrado construido sobre OH = OA + AH tendrá por área

$$OH^2 = (OA + AH)^2 = OA^2 + 2.OA.AH + AH^2 \qquad (1)$$

y si observamos la figura 1, veremos perfectamente delimitados los miembros del trinomio-producto:

- OA^2 es el área del cuadrado OADC,
- AH^2 es el área del cuadrado DEFG, ya que AH = DE por ser ambos segmentos partes de rectas paralelas comprendidos entre rectas paralelas; por último,
- OA.AH es el área de cualquiera de los rectángulos CDGM o AHED, puesto que para el primero de los rectángulos, su base CD es igual a OA, por ser partes de paralelas comprendidas entre paralelas y su altura GD es igual a DE, por ser ambos lados de un cuadrado, en donde el último, como hemos visto anteriormente, es igual a AH.

Como para el segundo de los rectángulos podemos seguir este mismo razonamiento, nos encontramos ante dos rectángulos cuyas bases son equivalentes a OA y cuyas alturas son equivalentes a AH y, como el área de un rectángulo es el producto de sus dos dimensiones, tendremos unas áreas equivalentes a

$$2.OA.AH$$

con lo que, finalmente, obtenemos el tercer término del trinomio.

Ahora bien, por definición, OA es igual a s_x y AH es igual a OB y, por tanto, a s_y, luego nos encontramos ante un cuadrado cuya área total se encuentra descompuesta en cuatro áreas bien definidas:

a) un cuadrado que es la varianza de X (s^2_x),
b) otro cuadrado que es la varianza de Y (s^2_y), y
c) dos rectángulos de área $s_x.s_y$,

es decir:

$$(s_x + s_y)^2 = s^2_x + 2.s_x.s_y + s^2_y \qquad (2)$$

Por otra parte, el segmento $s_x + s_y$ se puede considerar como el lado que se opone al ángulo formado por los vectores s_x y s_y, y sabemos que el cuadrado del lado opuesto a un ángulo es igual a la suma de los cuadrados construidos sobre los lados que forman el ángulo, más el doble del producto de uno de ellos por la proyección del otro sobre él, y como el ángulo formado por s_x y s_y es de 0° puesto

que ambas desviaciones típicas están situadas en ejes coincidentes, tendremos:

$$(\vec{s_x} + \vec{s_y})^2 = s^2_x + 2.s_x.s_y\cos O^\circ + s^2_y \qquad (3)$$

Si comparamos (2) con (3) veremos que ambas expresiones son idénticas puesto que el coseno de O° es igual a la unidad.

Ahora bien, por definición, r_{xy} es el coseno del ángulo que forman los ejes donde se sitúan las variables y, por tanto, la ecuación (3) puede tomar la forma

$$(\vec{s_x} + \vec{s_y})^2 = s^2_x + 2.s_x.s_y.r_{xy} + s^2_y \qquad (4)$$

y como el coeficiente de correlación de Pearson es definido por

$$r_{xy} = \frac{s_{xy}}{s_x.s_y}$$

tendremos que

$$s_{xy} = s_x.s_y.r_{xy}$$

y sustituyendo en (4)

$$(s_x + s_y)^2 = s^2_x + 2.s_{xy} + s^2_y \qquad (5)$$

lo que nos permite definir cada una de las figuras AHED y CDGM, como los rectángulos cuyas áreas expresan, indistintamente, la *covarianza máxima entre dos variables.*

Si nos remitimos ahora a nuestro caso concreto donde las variables se sitúan en unos ejes regulares con con $\varphi \neq O^\circ$ tendremos igualmente la expresión (4) con r_{xy} menor de uno y bajo cualquier concepto, tendremos la ecuación (5) en donde s_{xy} es igual a

$$s_x.s_y.\cos \varphi$$

o lo que es lo mismo

$$s_{xy} = (s_x) \ . \ \text{Proy}(s_y)$$

en donde la proyección de s_y vendría determinada por el valor de AP', puesto que $s_y = OB = AP$.

Desde el punto de vista físico, podríamos definir la covarianza de X e Y, de acuerdo con lo dicho anteriormente, como EL PRODUCTO ESCOLAR DE LAS DESVIACIONES TIPICAS DE AMBAS VARIABLES.

La covarianza de (X, Y) ha quedado pues determinada por el área

de cualquiera de los rectángulos ADIP' o CDKL, en donde el valor de uno de los lados viene dado por una de las desviaciones típicas, s_x en nuestro caso, y el valor del otro lado viene dado por la proyección de la otra desviación típica, en nuestro caso s_y, sobre la primera o sobre su eje direccional, en nuestro ejemplo s_x u OO', respectivamente.

Si observamos las figuras 2 y 3, vemos que, al abrirse los ejes ($\varphi < \varphi' < \varphi''$) permanece constante una de las dimensiones de los rectángulos —(s_x)— pero va disminuyendo la otra —Proy(s_y)— de tal forma que por haber definido la covarianza como el área de cualquiera de esos rectángulos, su valor va disminuyendo conforme aumenta el valor de φ .puesto que, al permanecer constante una de las dimensiones, siendo la otra función del coseno, su producto (área) disminuirá, ya que esta función disminuye de 1 a 0 cuando el ángulo aumenta de 0° a 90°. Recordemos nuevamente que la covarianza máxima se encuentra con $\varphi = 0°$.

De igual forma, definida s_{xy} por el producto $s_x.s_y.\cos \varphi$ la covarianza será nula cuando los ejes sobre los que se sitúan las desviaciones típicas de las variables sean perpendiculares ($\varphi = 90°$) puesto que el coseno de 90° es igual a 0 y de esta forma encontraríamos que

$$s_{xy} = s_x.s_y.0 = 0$$

Geométricamente observamos (ver figura 4) que la proyección del punto P sobre el eje OO', que designamos por P', coincidirá con A y por tanto la distancia AP' será nula por ser ambos puntos coincidentes. De igual forma K coincide con D y L con C y siendo la covarianza el área determinada, indistintamente, por los rectángulos CDKL o ADIP', observamos que por ser cero una de las dimensiones, su área (covarianza entre las variables) sería también cero.

Nada nos impide, por tanto, definir la covarianza como EL AREA DEL RECTANGULO CUYAS DIMENSIONES VIENEN DADAS POR UNA DE LAS DESVIACIONES TIPICAS Y POR LA PROYECCION DE LA OTRA SOBRE ELLA MISMA (O SOBRE SU EJE DE DESPLAZAMIENTO).

De la misma forma que hemos procedido hasta ahora donde hemos situado uno de los ejes de las variables sobre el eje imaginario

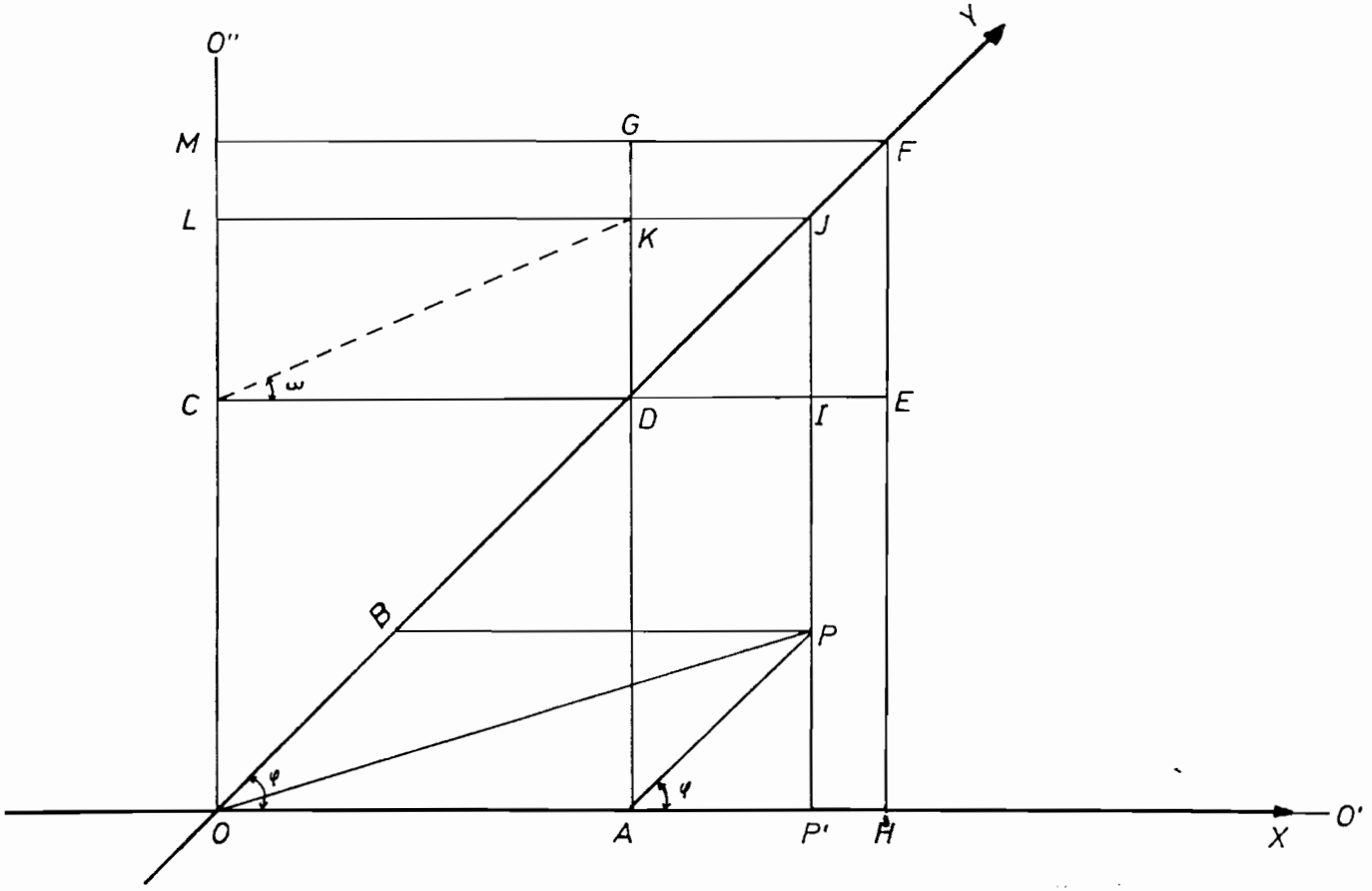


Fig 2

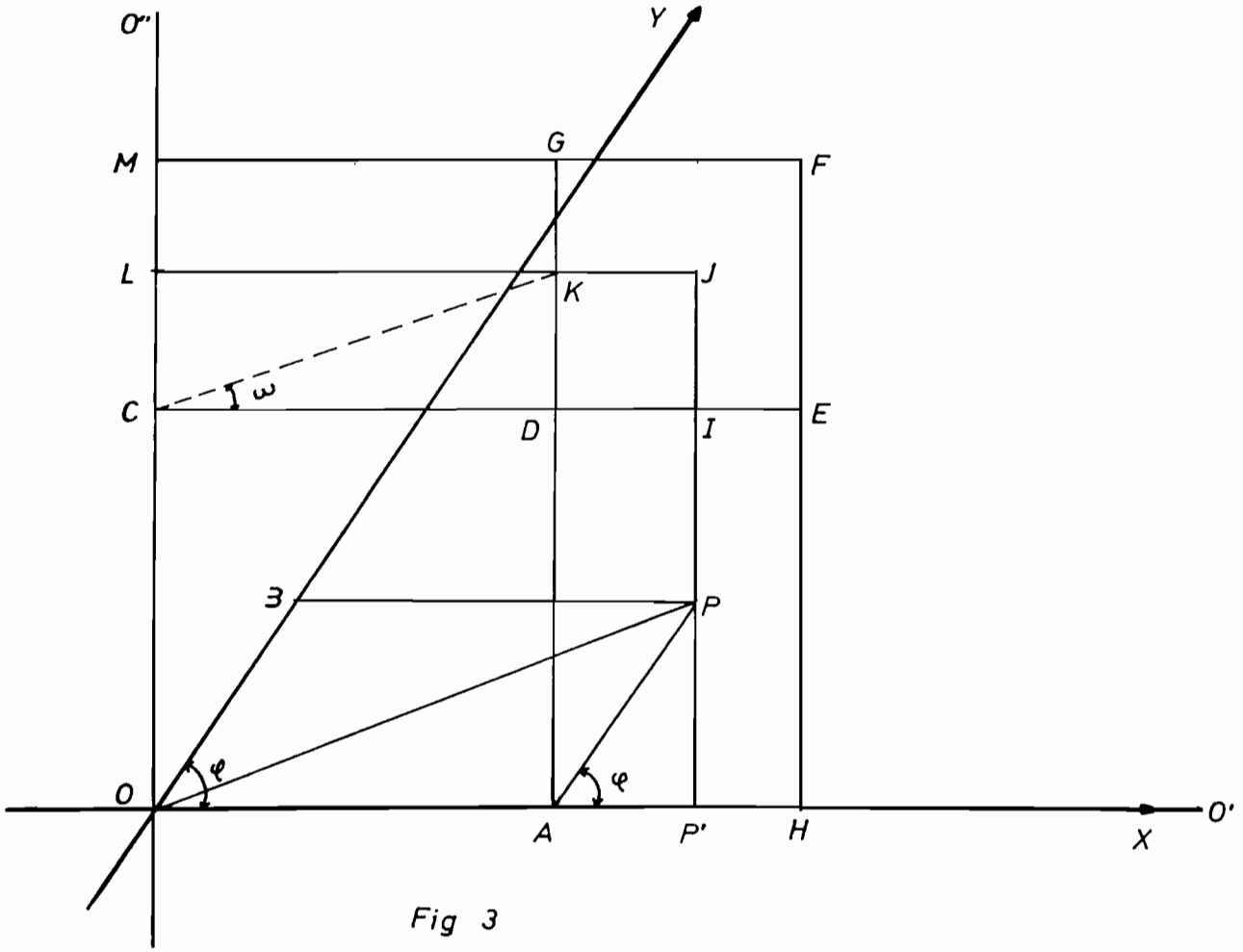


Fig 3

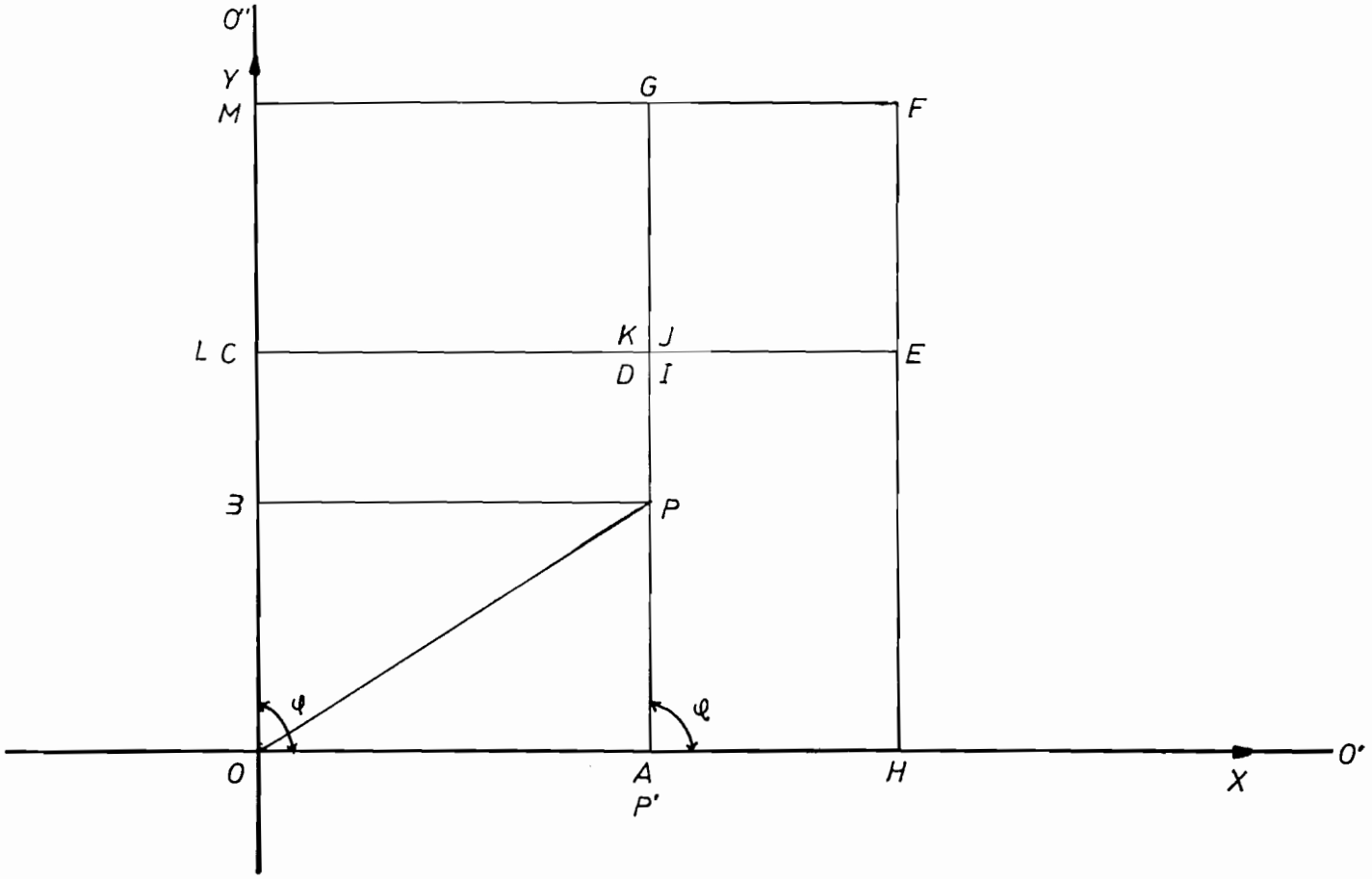


Fig 4

OO' y el eje de la otra en el primer cuadrante, podríamos haber razonado con el segundo eje dentro del segundo cuadrante ($90° < \varphi < 180°$) o dentro del tercer cuadrante ($180° < \varphi < 270°$) o dentro del cuarto ($270° < \varphi < 360°$), teniendo en cuenta que:

$$\cos (180° - \varphi) = - \cos \varphi$$
$$\cos (180° + \varphi) = - \cos \varphi$$
$$\cos (360° - \varphi) = \cos \varphi$$

Cuando el segundo eje se encuentra situado en el segundo cuadrante, es decir, cuando los ejes sobre los que se sitúan las desviaciones típicas de las variables X e Y, forman un ángulo superior a 90° e igual o inferior a 180°, para construir el modelo geométrico, hemos de proceder de forma similar a la que hasta ahora habíamos utilizado, pero trabajando sobre la base, no de OA y OB, sino de OB y OA', siendo OA' = — OA, por lo que las áreas obtenidas anteriormente a partir de los productos OA.OB, proyecciones o segmentos equivalentes a ambas dimensiones, vendrían ahora expresadas por OB.OA' y dando a OA' su valor, el producto anterior equivaldría a obtener áreas de valor OB.(—OA) y, por tanto, con carácter negativo.

En este cuadrante, la covarianza obtenida mediante el área definida por el producto, por ejemplo, (DA').(A'P'), en donde A'D = OA' = — OA (por construcción) y A'P' es la proyección de A'P = OB tendrá carácter negativo, puesto que será negativo el producto (—OA).proy(OB), valores que corresponden a las dos dimensiones de cualquiera de los rectángulos A'P'ID o KDCL (figura 5). El área suma vendría determinada según la expresión (3) por:

$$(\vec{s_x} + \vec{s_y})^2 = s^2_x + 2.s_x.s_y \cos (180° - \varphi) + s^2_y$$

es decir,

$$(\vec{s_x} + \vec{s_y})^2 = s^2_x + 2.s_x.s_y.(-\cos \varphi) + s^2_y \qquad (6.a)$$

o bien

$$(\vec{s_x} + \vec{s_y})^2 = s^2_x - 2.s_x.s_y.\cos \varphi + s^2_y \qquad (6.b)$$

por tanto, el término central representa el área de los rectángulos

anteriormente mencionados y correspondientes a la figura 5, pero considerándola con carácter negativo: — $2.s_{xy}$.

Cuando el segundo eje se encuentra situado en el tercer cuadrante (figura 6), es decir, cuando los ejes sobre los que se sitúan las desviaciones típicas de las variables X e Y forman un ángulo superior a 180° e igual o inferior a 270°, la construcción del modelo geométrico es similar a la anterior, puesto que volvemos a trabajar con OB y OA' y donde OA' vuelve a ser igual a — OA. El área-suma vendría determinada por (6.a) o (6.b) puesto que el $\cos(180° + \varphi) = -\cos\varphi$ y por tanto:

$$(\vec{s_x} + \vec{s_y})^2 = s^2_x - 2.s_x.s_y.\cos\varphi + s^2_y \qquad (7)$$

donde el término central representa el área de los rectángulos P'A'DI y DCLK correspondientes a la figura 6, pero considerándola, igualmente, con carácter negativo.

Cuando el segundo eje se encuentra situado en el cuarto cuadrante (figura 7), es decir, cuando los ejes sobre los que se sitúan las desviaciones típicas de las variables X e Y forman un ángulo superior a 270° e igual o inferior a 360°, la construcción del modelo geométrico es similar a los de las figuras 1 a 4, puesto que volvemos a considerar las desviaciones típicas de las variables OA y OB en su sentido original y puesto que, además, $\cos(360° - \varphi) = \cos\varphi$. El área-suma vendría dada por

$$(s_x + s_y)^2 = s^2_x + 2.s_x.s_y\cos\varphi + s^2_y \qquad (8)$$

con el término central igual a $2.s_{xy}$ con carácter positivo.

Vemos por tanto que, considerando la varianza-unión como la varianza obtenida a partir de la suma de las desviaciones típicas, observamos que una parte de esta varianza-unión viene determinada, exclusivamente, por la desviación típica de X, siendo su valor s^2_x; otra parte viene determinada, exclusivamente, por la desviación típica de Y, cuyo valor es s^2_y y, finalmente, existe una tercera parte que perteneciendo a la unión (suma) viene determinada por una y otra desviación típica siendo su valor ($s_x.s_y.\cos\varphi$).

Desde esta perspectiva, podríamos considerar a los dos rectángulos cuyas áreas son, indistintamente, igual a la covarianza, como la

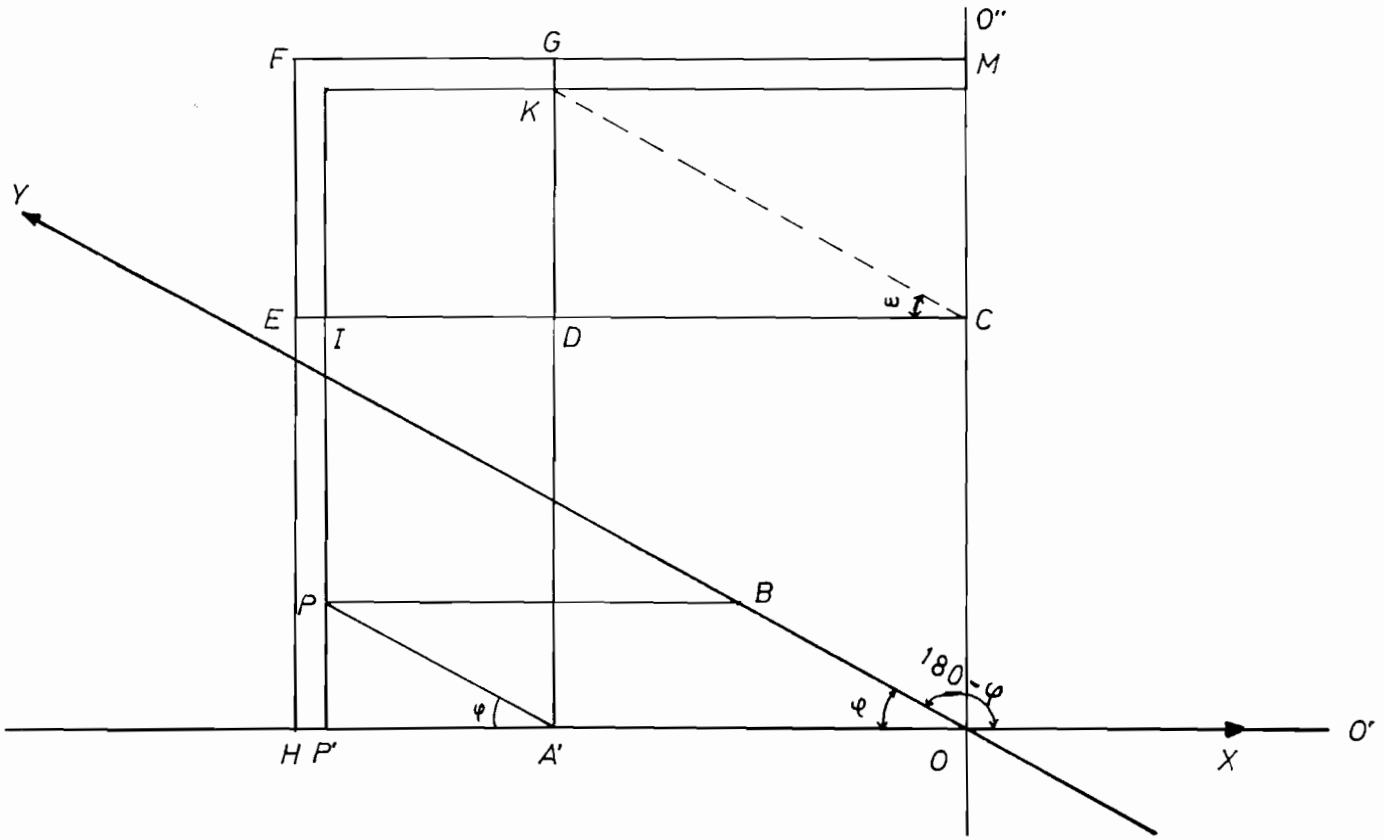


Fig 5

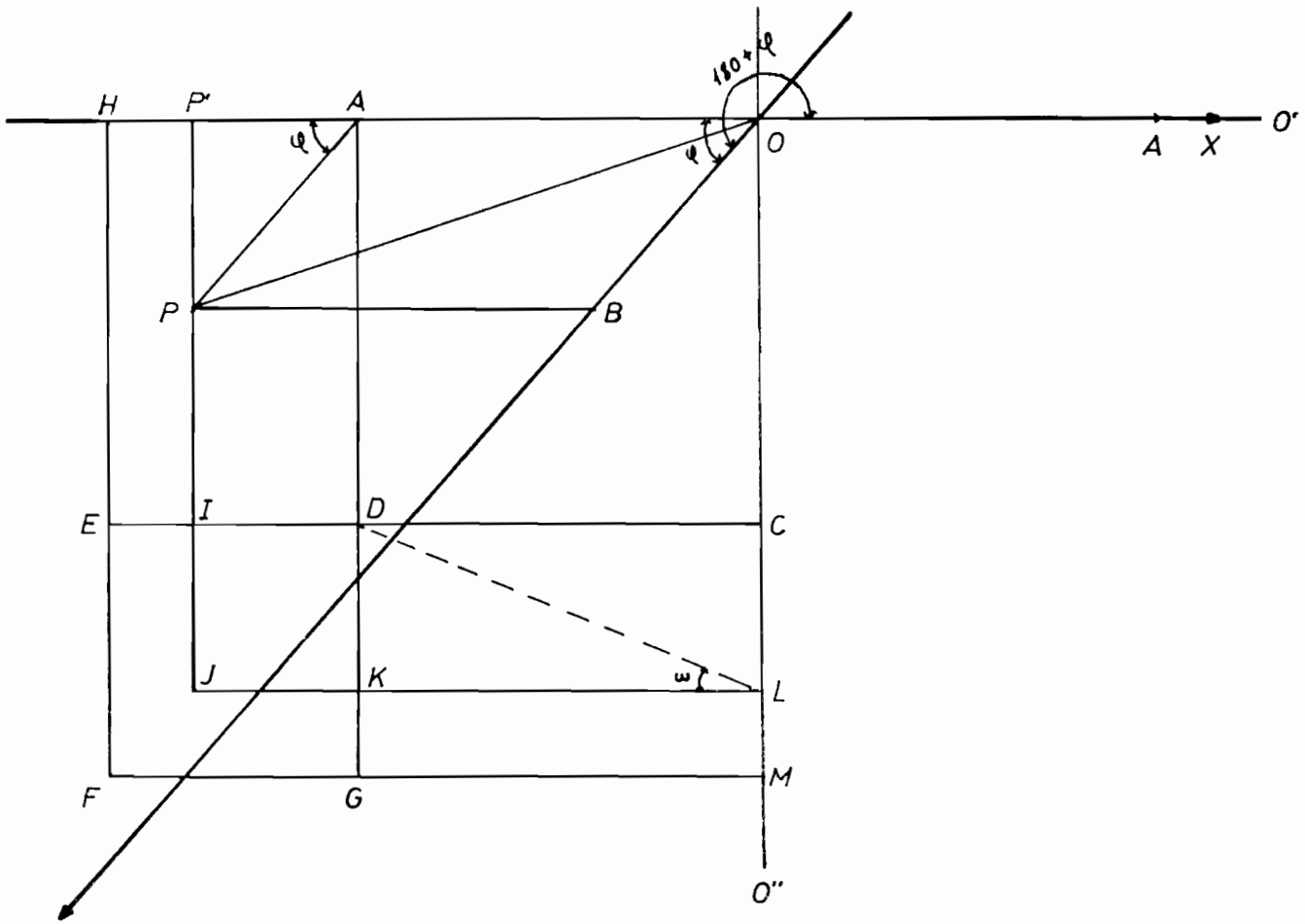


Fig 6

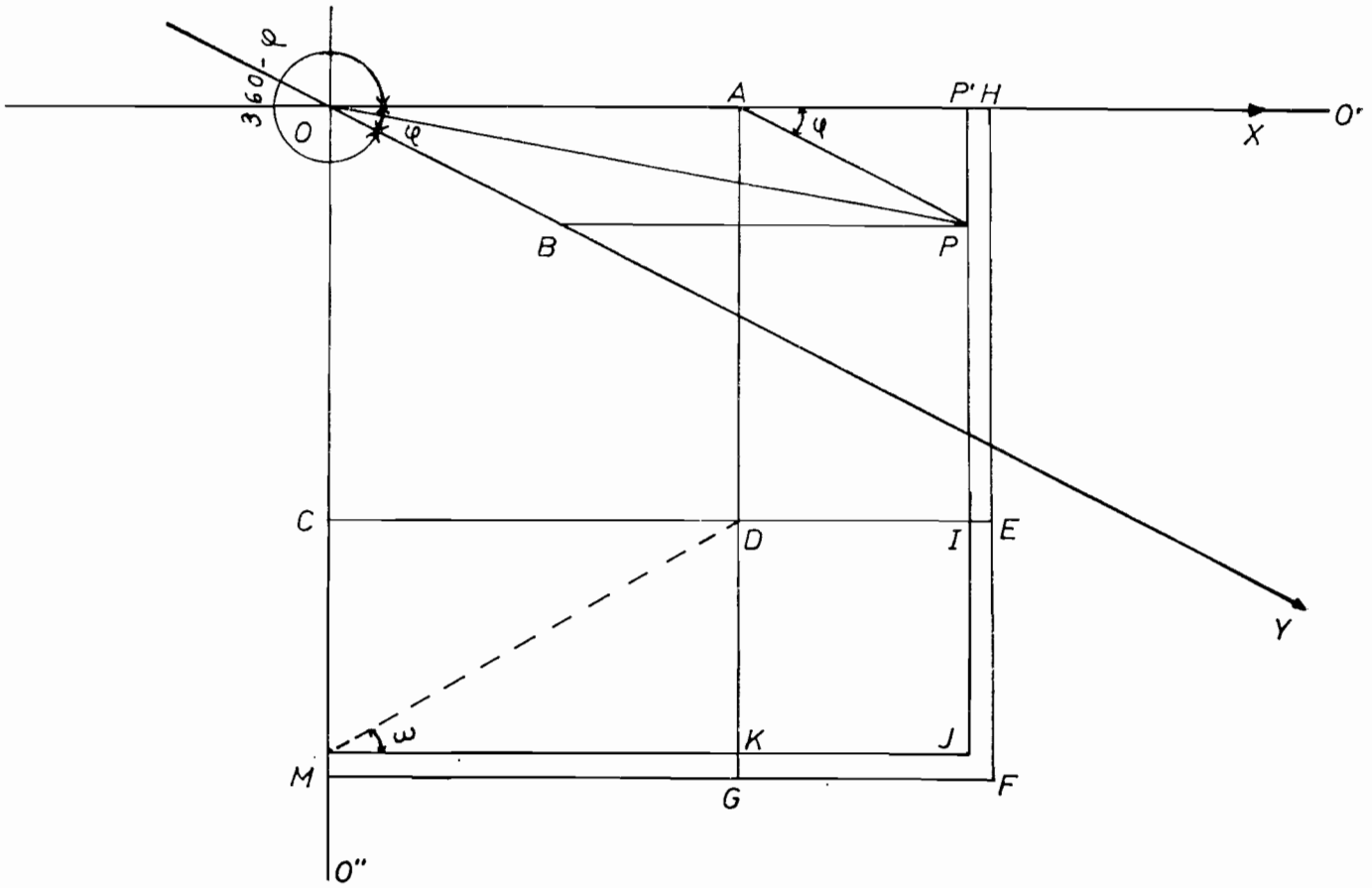


Fig 7

"varianza-intersección" de los conjuntos cuya "varianza-unión" viene determinada por $(s_x + s_y)^2$ y, por tanto la *covarianza* podría ser definida como LA MITAD DE LA VARIANZA INTERSECCION O VARIANZA COMUNITARIA DE DOS VARIABLES.

INTERPRETACION GEOMETRICA DE (r_{xy}).

En todas las figuras (figuras 1 a 7) hemos visto representadas dos covarianzas, una, la real, que depende de la inclinación de los ejes en donde se sitúan las desviaciones típicas de las variables X e Y, representada por el área de cualquiera de los rectángulos ADIP' (para las figuras 1 a 4), A'DIP' (para las figuras 5 a 7) o CDKL (para todas las figuras —1 a 7—), cuyo valor viene determinado por el producto $(s_x.s_y.\cos \varphi)$ con $\varphi \neq 0°$; la otra, la máxima, representada por el área de cualquiera de los rectángulos ADEH (para las figuras 1 a 4), A'DEH (para las figuras 5 a 7) o CDGM (para todas las figuras, de la 1 a la 7), cuyo valor viene determinado, asimismo, por el producto $(s_x.s_y.\cos \varphi)$ pero con $\varphi = 0°$. Pues bien, definimos el coeficiente de correlación de Pearson (r_{xy}) como LA RELACION (RAZON) ENTRE LA COVARIANZA REAL Y LA COVARIANZA MAXIMA DE LAS VARIABLES, es decir:

$$\frac{s_x.s_y.\cos \varphi}{s_x.s_y.\cos 0°} = \cos \varphi = r_{xy} \text{ (puesto que } \cos 0° = 1)$$

INTERPRETACION GEOMETRICA DE (r^2_{xy}).

El cuadrado cuyo lado es OP' es decir, el cuadrado construido sobre $\text{Proy}(s_x + s_y)$ y que tiene por área

$$(OP')^2 = \left[\text{Proy}\,(\vec{s_x} + \vec{s_y})\right]^2$$

puede ser descompuesto, al igual que lo hacíamos con la covarianza máxima, en dos cuadrados y dos rectángulos (ver, por ejemplo, la figura 1), y puesto que

$$\text{Proy}(\vec{s_x} + \vec{s_y}) = \text{Proy}(\vec{s_x}) + \text{Proy}(\vec{s_y})$$

tendríamos que

$$(OP')^2 = \left[\text{Proy}(\vec{s}_x) + \text{Proy}\,(\vec{s}_y)\right]^2$$

y desarrollando el segundo miembro de la igualdad

$$(OP')^2 = \left[\text{Proy}(s_x)\right]^2 + 2.\left[\text{Proy}(s_x)\right].\left[\text{Proy}(s_y)\right] + \left[\text{Proy}(s_y)\right]^2$$

en donde el término central del trinomio no se encuentra multiplicado por el cos φ puesto que, por situarse las proyecciones sobre un mismo eje $\varphi = 0°$ y, por tanto, cos $\varphi = 1$. Además, Proy(s_x) $= s_x$ por encontrarse s_x sobre el eje de proyección y Proy(s_y) $= s_y.\cos\varphi$ como anteriormente vimos; por tanto, tendremos

$$(OP')^2 = s^2_x + 2.s_x.s_y.\cos\varphi + s^2_y.\cos^2\varphi$$

es decir

$$(OP')^2 = s^2_x + 2.s_{xy} + s^2_y.\cos^2\varphi$$

y como según (5)

$$(\vec{s}_x + \vec{s}_y)^2 = s^2_x + 2.s_{xy} + s^2_y$$

ambas áreas difieren en el tercer término del trinomio, es decir, en el área del cuadrado superior derecho (ver figuras 1 a 4), superior izquierdo (figura 5), inferior izquiedo (figura 6) o inferior derecho (figura 7).

Pues bien, la razón entre las áreas de los cuadrados DKJI y DEFG determinadas por los terceros términos de los trinomios anteriores recibe el nombre de coeficiente de determinación (r^2_{xy}).

Podemos definir, por tanto, el *coeficiente de determinación* como LA RELACION (RAZON) DE LA PARTE DE LA VARIANZA DE UNA DE LAS VARIABLES, ASOCIADA A LA COVARIANZA DE AMBAS, ENTRE LA VARIANZA TOTAL DE LA MISMA.

El mismo valor habríamos obtenido si las varianzas de X e Y se hubieran permutado (ver figura 8). En el primero de los casos tendríamos:

$$\frac{s^2_y.\cos^2\varphi}{s^2_y} = \cos^2\varphi = r^2_{xy}$$

y en el segundo

$$\frac{s^2_x.\cos^2\varphi}{s^2_x} = \cos^2\varphi = r^2_{xy}$$

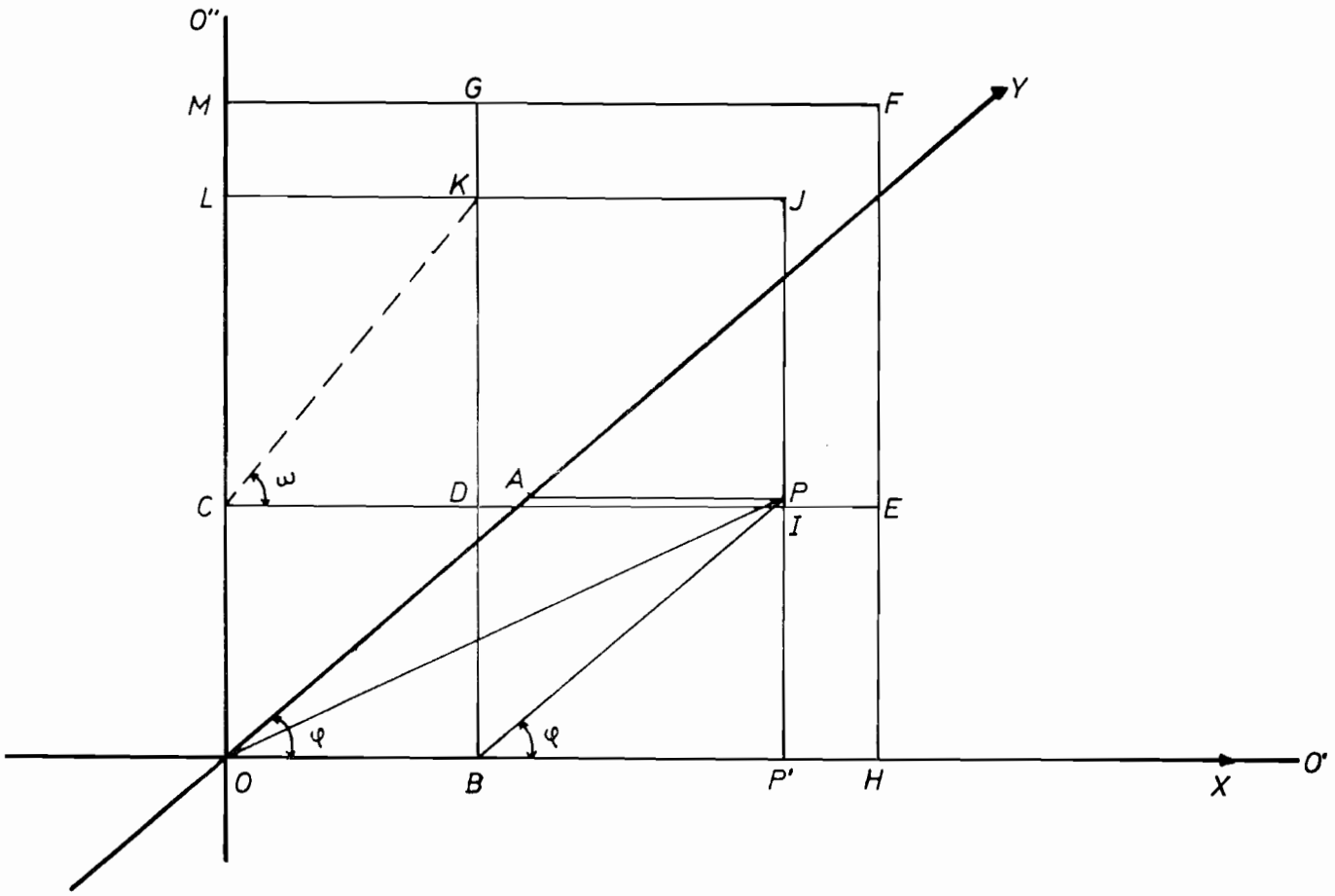


Fig 8

REFERENCIAS BIBLIOGRAFICAS

AMON, J.: *Estadística para psicólogos. Vol. 1: Estadistica descriptiva.* Ediciones Pirámide, S. A. Madrid, 1978.

CALOT, G.: *Cours de statistique descriptive.* Dunot. París, 1965. (Existe traducción española con el título de *Curso de estadística descriptiva*, en Paraninfo, Madrid, 1982).

HAYS, W. L.: *Statistics for the Social Sciences.* Holt, Rinehart and Winston. New York, 1980. 3rd. Edition.