URIBE VILLEGAS

EL A. B. C. DE LA CORRELACION Y SUS APLICACIONES SOCIALES

EL A.B.C. DE LA CORRELACION Y SUS APLICACIONES SOCIALES

INSTITUTO DE INVESTIGACIONES SOCIALES DE LA
UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

El A. B. C. de la Correlación y sus Aplicaciones Sociales

Por

OSCAR URIBE VILLEGAS

MEXICO

Primera edición 1962.

Impreso y Hecho en México.
Printed and Made in Mexico.

Al Prof. D. Luis Gutiérrez,

Mi Maestro de Matemáticas en el Instituto Científico y Literario Autónomo del Estado de México, hoy Universidad del Estado.

A la Memoria de D. Manuel María Contreras, Ilustre Matemático y Pedagogo Mexicano, paradigma para mí insuperable.

LIMINAR

La correlación es, probablemente, uno de los capítulos de la estadística que más importancia pueden tener para el sociólogo. Sus posibilidades de instrumento analítico no se encuentran agotadas aún. E incluso puede afirmarse que, sobre todo en los países latinoamericanos, apenas si —muy tímidamente— se han comenzado a explorar. Esa timidez en la utilización de instrumento tan fino y útil parece depender en buena parte de la forma en que brindan la enseñanza de la correlación la mayoría de los manuales elementales. Y no hablamos de los manuales superiores porque éstos dan por supuesto un conocimiento que deberían proporcionar los elementales y que no siempre brindan éstos en la forma deseable. Capítulos sobre la correlación los hay en ellos, pero generalmente los mismos tienen más aspecto de formulario o recetario que la apariencia de una teoría (así sea muy modesta) íntimamente trabada. Al través de su lectura y estudio, el estudioso puede llegar a obtener correlaciones incluso altamente elaboradas pero, en cuanto no hace sino reconocer los nódulos de una gran red sin lograr vislumbrar los hilos mismos que en tales nódulos se unen, no puede proceder sino con gran temor en un terreno cuyas acechanzas desconoce. Es por ello por lo que nos atrevemos a presentar en forma monográfica y un poco morosa este abecedario de la correlación. Su simbología; su forma de tratamiento, puede hacer suponer que en vez de simplificación se buscan complicaciones innecesarias. Nosotros creemos que, por el contrario, todas estas dificultades iniciales —reconocidas y no desconocidas desde el principio— salvadas en lo asequible (lo asequible a modestas preparaciones matemáticas que no rebasan el terreno del álgebra) permiten que el estudioso vea sin zozobra la posibilidad de aplicar los modestos conocimientos que se le brindan, así como la posibilidad de abrir sus ojos a horizontes más amplios de la teoría y de la práctica estadística en ciencias sociales. Sólo quien tras el estudio de este modesto esfuerzo proceda hacia terrenos más elevados y vea retrospectivamente el camino recorrido podrá decir, con justicia, si hubo en el caso acierto o desacierto por parte de

EL AUTOR.

I. MARCO DE CONJUNTO PARA EL ESTUDIO DE LA CORRELACION

C O R R E L A C I O N

Hay correlación entre dos o más fenómenos cuando puede predecirse el valor de uno de ellos (x_1) conocidos los valores del o de los restantes (x_2, x_3, x_4...) mediante el uso de una expresión matemática (ecuación de regresión o ecuación de estimación) que proporciona un valor estimado de dicho fenómeno ($\hat{x}_1$) sujeto a un error (error de estimación ε) menor o mayor, según sea más o menos intensa la correlación (medida por un "índice de correlación").

Existen varias formas de correlación, diferenciables sobre la base de los siguientes criterios:

I.—Imputabilidad de una relación de causalidad entre los fenómenos que se correlacionen.
II.—Carácter de los datos de la correlación.
III.—Número de fenómenos que se correlacionan.
IV.—Forma de la relación (que se manifiesta al través de la ecuación de estimación o de regresión).
V.—Sentido de la correlación (manifiesto en el signo del índice de correlación).

De acuerdo con la imputabilidad o falta de imputabilidad de una relación causal entre los fenómenos que se correlacionan, la correlación puede ser:

1. Aparente o puramente estadística, y
2. Real.

En rigor, esta distinción —que debe realizarse con ayuda de métodos extraestadísticos— lo único que hace es enfatizar el hecho de que al encontrar una correlación estadística, incluso alta entre dos o más fenómenos, no significa por sí mismo y sin ninguna prueba adicional el que pueda postularse entre ellos una relación de causalidad, pudiendo decirse en términos generales que:

a.—Basta que un índice de correlación sea distinto de cero (otras limitaciones o precisiones numéricas no nos interesan por el momento) para que exista, estadísticamente, correlación.

b.—Es necesario que, además de ser el índice de correlación distinto de cero, exista entre los fenómenos correlacionados una proximidad tal (definida en términos de la disciplina específica de que se trate; definida en términos específicos de causación social para el caso de la sociología) para que pueda postularse una de las siguientes posibilidades:

1º que uno de los fenómenos correlacionados es causa de los restantes,
2º que el restante o el conjunto de los restantes es causa del primero,
3º que el primero y los restantes menos uno son causa o efecto de éste,
4º que todos los fenómenos correlacionados son efectos de una causa común.

c.—Que cuando el índice de correlación sea distinto de cero, pero no exista entre los fenómenos la debida proximidad, no puede postularse una relación de causalidad, ya que ello sería dar justificación a los críticos de la noción de causa, para quienes, finalmente, el conjunto de todos los antecedentes (el universo mismo) es la causa de cualquier consecuente.

Según los caracteres estadísticos de los datos que se toman para la correlación, ésta se diferencia en:

1. Correlación de cantidades,
2. Correlación de cualidades.

Dentro de la correlación entre cantidades, y con vistas a la diferenciación de los procedimientos utilizables, puede hacerse una distinción entre:

1. Correlación entre series simples,
2. Correlación entre series de frecuencias,
3. Correlación entre series de clases y frecuencias,
4. Correlación entre series cronológicas.

Sobre la base de la distinción por el número de variables (y sin que dudemos en llevar la distinción hasta el límite, aunque tal límite parezca absurdo) puede distinguirse entre:

1. Correlación de una distribución univariada,
2. Correlación de una distribución bivariada,
3. Correlación de una distribución multivariada.

En el primer caso, se trata de la correlación que presenta una variable con respecto a sí misma. Naturalmente esta es una correlación perfecta, o sea, una correlación cuyo índice representa el máximo de intensidad de la correlación. Cuando, en el mundo sujeto a las variaciones aleatorias a que estamos acostumbrados (lo aleatorio, en estos contextos es "lo no explicable explícitamente"), una correlación es tan alta que llega a aproximarse excesivamente a la correlación perfecta, podemos y debemos preguntarnos si en realidad no estaremos encontrando la correlación de un fenómeno con respecto a sí mismo, por haber trabajado con los que, creyendo nosotros que eran dos fenómenos completamente distintos, son en realidad dos formas de expresión distintas del mismo fenómeno.

En el caso de una correlación de serie bivariada, se trata de encontrar la relación estadística existente entre una variable dependiente (desde el ángulo matemático) y otra independiente, o sea, en total, entre dos variables. A esta correlación también puede llamársele "correlación de primer orden", siendo obviamente la anterior una "correlación de orden cero".

En el caso de una correlación multivariada, se trata de la relación entre una variable considerada (matemáticamente) como dependiente con respecto a dos o más variables consideradas como independientes de la primera y que, además, son (o se consideran) independientes entre sí (o bien, para expresarlo en otra forma, con respecto a dos o más variables que no se interinfluyen o que se considera que no influyen entre sí).

Encabalgada en cierta forma entre la correlación bivariada, simple o de primer orden y la correlación multivariada, múltiple o de orden superior al primero, se encuentra la llamada correlación parcial. Si queremos conservar —mediante el uso de dos enunciados distintos— la convergencia-divergencia entre la correlación parcial y la simple y la correlación parcial y la múltiple, diremos que:

En la correlación parcial, se trata de encontrar la relación de una variable dependiente con respecto a otra variable independiente (con lo cual podría pensarse que se confundiría con la correlación bivariada, simple o de primer grado, *pero*) cuando otras variables independientes permanecen constantes.

O bien podríamos decir que:

La correlación parcial establece la relación entre una variable dependiente y dos o más independientes (con lo que se pensaría que se confunde con la correlación múltiple), pero las cuales se interinfluyen (por lo cual imponen la necesidad de eliminar sus mutuas influencias).

Más adelante, trataremos de precisar estas distinciones, con base en conocimientos que aún no hemos obtenido en este punto.

Con base en la forma de la relación entre las variables, se dice que la correlación es:

1. Rectilínea o del sistema rectilíneo, o
2. Curvilínea o del sistema de una curva particular.

El que la correlación sea rectilínea quiere decir, en el caso más sencillo (o sea, en el de la correlación rectilínea bivariada, simple o de primer orden) que a incrementos constantes de la variable independiente corresponden incrementos también constantes de la variable dependiente. En tal caso, la forma de la relación puede representarse por una recta. Si la correlación es del sistema rectilíneo y, además, es múltiple, la forma de la relación puede representarse mediante un plano (caso de tres variables) o por medio de un "hiperplano" (en el caso de más de tres variables).

El que la correlación sea curvilínea significa que a incrementos constantes de las variables independientes, corresponden incrementos variables (sujetos a una variabilidad regida a su vez por alguna expresión matemática) de la dependiente.

Para medir la intensidad de la correlación rectilínea, se emplea un índice que se representa con r minúscula y al que se sufijan como índices de cifras de las variables que se correlacionan.

De este modo:

r_{12} representa el índice de correlación rectilínea de la variable 1 en la variable 2.

r_{13} representa el índice de correlación rectilínea de la variable 1 en la variable 3.

r_{123} representa el índice de correlación del sistema rectilíneo (o, en el caso, "cuadrática") de la primer variable en la segunda y en la tercera consideradas simultáneamente (o sea, que se trata de una correlación múltiple).

En el caso de la correlación parcial, se emplea la misma literal, pero, en el sufijo o subíndice se separan las variables que se relacionan, de la o de las variables que se mantienen constantes o cuya influencia se reduce de este modo. Es así como:

$r_{12.3}$ representa el índice de correlación parcial de la primera variable en la segunda cuando permanece constante la tercera.

$r_{13.2}$ representa el índice de correlación parcial de la primera variable en la tercera cuando permanece constante la segunda.

La correlación curvilínea se aprecia mediante el cálculo de un índice conocido como "razón de correlación" que se acostumbra representa por η (eta) minúscula. Cuando η_{12} es igual a r_{12} la correlación no es curvilínea sino rectilínea.

Con base en el sentido del crecimiento, se diferencia entre:

1. Correlación directa,
2. Correlación indiferente,
3. Correlación inversa.

El sentido de la correlación se aprecia generalmente al través del signo del índice de correlación. Cuando el signo de correlación es positivo, la correlación es directa, o sea, que al aumentar la variable independiente aumenta la dependiente y al disminuir la primera disminuye la segunda. Cuando el signo del índice de correlación es negativo, la correlación es inversa, o sea, que al aumentar la variable independiente disminuye la dependiente y al disminuir la independiente aumenta la dependiente. Finalmente, cuando no hay signo por ser nulo el índice de correlación (en este sentido 0 es el signo de indiferencia) se dice que no existe correlación, puesto que no puede predecirse el valor de la variable dependiente con base en el conocimiento que se tenga del valor correspondiente de la variable independiente.

II. LA CORRELACION ESTUDIADA MEDIANTE EL METODO DE REGRESION

CORRELACION RECTILINEA DE PRIMER ORDEN

Series Sencillas.—Una serie sencilla bivariada es aquélla que se obtiene al representar a cada individuo de un conjunto, población o universo estadístico por un par de cifras. Una serie sencilla bivariada puede estar constituida por peso y estatura de cada uno de los miembros de una familia. Otra serie sencilla bivariada puede estar constituida por la lista de pares de valores que dan el monto de los ingresos y el monto de los egresos de una serie de familias. Una tercer serie sencilla bivariada estará constituida por la lista de pares de valores que den las edades de un conjunto de mujeres y el número de los hijos de cada una de ellas. Una cuarta serie puede estar representada por el conjunto pareado de valores que representen, por una parte, el por ciento del ingreso nacional manejado por los gobiernos y por otra, el por ciento de las funciones sociales realizadas por esos mismos gobiernos, en un conjunto de Estados. Tales series bivariadas seguirán siendo *sencillas,* en tanto no se hayan subsumido los posibles casos de repetición en uno solo. O sea, en tanto no se haya hecho acompañar cada pareja distinta de valores de una tercer cifra que indique cuántas son las veces en que se repite dicha pareja.

Consideremos, por ejemplo, dos fenómenos A y B. Supongamos que A puede adquirir los valores 6 y 12 y que B pueda adquirir los valores 3 y 5. Supongamos finalmente, un conjunto de individuos p, q, r, s, t, u, v, x. Supongamos asimismo que la observación nos ha mostrado que, para p, el fenómeno A vale 12 y el B vale 5; que para q, A vale 12 y B, 3; para r, A vale 6 y B, 3; para s, A vale 6 y B,5; para t, A=12 y B=3; para u, A=12 y B=3; para v, A=6 y B=5; para x, A=12 y B=3. Una serie sencilla, bivariada estará dada por:

	A	B
p	12	5
q	12	3
r	6	3
s	6	5
t	12	3
u	12	3
v	6	5
x	12	3

En esa serie *sencilla,* como puede verse, no se ha evitado la repetición de casos como el de q, t, u, x (a todos los cuales les corresponde la pareja 12,3) o como el de s, v (a los que corresponde la pareja 6,5). La serie *sencilla* bivariada podría convertirse en serie de frecuencias si se evitaran las repeticiones y haciendo perder su individualidad a cada observación, se subsumieran q, t, u, x bajo un 4 que les correspondería como la "frecuencia" de la pareja correspondiente 12,3) y se subsumiera a s, v bajo un 2 que sería la frecuencia de la pareja de valores 6, 3. En esta forma se obtendría una distribución bivariada de frecuencias como la siguiente:

A	B	frecuencia
12	5	1
12	3	4
6	5	2
6	3	1
		8 casos en total

En lo que sigue, trataremos de las series sencillas únicamente en relación con el tema de la correlación.

Se trata, por tanto, de la correlación rectilínea de primer orden (entre dos variables o sea en una serie bivariada) en serie sencilla.

$$\left.\begin{array}{l} \text{El que A } \dfrac{\text{crece}}{\text{decrece}} \text{ cuando B } \dfrac{\text{crece}}{\text{decrece}} \\[2ex] \text{o el que A } \dfrac{\text{decrece}}{\text{crece}} \text{ cuando B } \dfrac{\text{crece}}{\text{decrece}} \end{array}\right\} \begin{array}{c}\text{puede expresarse ma-}\\ \text{temáticamente:}\\[2ex] A = cB\end{array}$$

Expresión en la cual c representa una constante que multiplica a B y que puede ser positiva o negativa. Si c es positiva, esto quiere decir que cuando B crece, A también crece y cuando B decrece, A también decrece. Si c es negativa, esto significa que cuando B crece A decrece y cuando B decrece A crece.

Si los diferentes valores que puede asumir el fenómeno A (por ejemplo, los años en que puede expresarse la *Edad* de un conjunto de individuos) se representa por x_1 y si los diferentes valores que puede asumir el fenómeno B (por ejemplo, los años en que puede expresarse la *Edad* de los hijos de esos individuos) se representa por x_2 y se sabe, además que cada aumento unitario del valor del fenómeno B (en el ejemplo, por cada año de edad del hijo) existe un aumento de a_2 unidades en el valor del fenómeno B (a años más en la edad del padre), estas relaciones pueden expresarse:

$$x_1 = a_2 \, x_2$$

Aprovecharemos la ocasión que nuestro ejemplo nos proporciona para tratar de deshacer un malentendido que podría surgir (a no ser palpable el absurdo, en el caso) en la interpretación de los resultados de la correlación.

Si x_1 representa las edades de los padres del conjunto y x_2 las edad*es* de l*os* hijo*s* y a_2 en un conjunto determinado vale 2, por ejemplo:

¿La expresión $x_1 = 2x_2$ quiere decir que "por cada año que aumenta la edad de un hijo determinado se duplica la edad de su padre"? El absurdo es evidente. No sería posible que mientras el hijo envejecía un año, el padre envejeciera dos. Entonces, ¿qué es lo que esta expresión significa? Significa que "en el *conjunto* de los individuos estudiados *en un momento dado,* los padres tienen el doble de la edad de sus hijos".

El absurdo de la interpretación procede:

1º De que se habla de variaciones de edad no en sentido colectivo (no en el sentido de diversidad de edades de los individuos de un conjunto), sino individual (en el sentido de diversidad de edades de dos individuos vinculados entre sí, en el transcurso del tiempo) o

2º De que se habla de variaciones cronológicas cuando debiera de hablarse de variaciones no cronológicas; de variaciones diacrónicas cuando se debiera hablar de variaciones sincrónicas; de variaciones que se producen en series estáticas y no en series dinámicas.

La expresión dada anteriormente para esas variaciones simultáneas de dos fenómenos puede ampliarse mediante la igualdad:

$$x_1 = a_0 + a_2x_2$$

De acuerdo con esta expresión —más general que la previa, por la adición del sumando constante a_0— cuando x_2 vale 0, x_1 vale a_0 y por cada unidad de aumento de x_2 hay un aumento de a_2 unidades en x_1.

En nuestro ejemplo, podría darse el caso de que un individuo tuviera 18 años cuando naciera su hijo, que otro individuo del conjunto tuviera 20 años cuando su hijo tenía 1 año, que otro más cuyo hijo tuviera 2 años tuviera 22. En tal caso, la expresión específica de estos hechos sería:

$$x_1 = 18 + 2\,x_2$$

pues, en efecto, cuando x_2 valiera 0, x_1 valdría 18; cuando valiera 1, valdría x_1 20, y al valer x_2, 2, valdría x_1, 22.

Volveremos ahora a nuestra expresión genérica:

$$x_1 = a_0 + a_2\,x_2$$

Obtención del índice de correlación rectilínea a partir de la ecuación de regresión expresada en unidades originarias.—La expresión genérica anterior está expresada en unidades originarias. Si a esta expresión le aplicamos el operador sigma o sumador obtendremos:

$$\Sigma x_1 = Na_0 + a_2\,\Sigma x_2$$

Si la expresión que nos sirve de punto de partida se multiplica por x_2, se tiene:

$$x_1x_2 = a_0x_2 + a_2\,x_2^2$$

Al aplicar el sumador a esta expresión resultante, obtenemos:

$$\Sigma x_1x_2 = a_0\Sigma x_2 + a_2\Sigma x_2^2$$

Puesto que aplicar a una expresión el operador sigma equivale a aplicarlo separadamente a su primero y a su segundo miembro y, dentro de cada miembro, a cada término. Puede recordarse también que aplicar el sumador a un término constituido por un producto de constante y variable (como $a_0\,x_2$) equivale a aplicar el sumador a la variable (x_2) y multiplicar la sumatoria (Σx_2) por la constante (a_0 en el ejemplo).

Las dos expresiones sumatorias obtenidas, constituyen el siguiente sistema de dos ecuaciones con dos incógnitas (a_0, a_2):

$$\Sigma x_1 = Na_0 + a_2 \Sigma x_2$$
$$\Sigma x_1 x_2 = a_0 \Sigma x_2 + a_2 \Sigma x_2^2$$

Al resolver el sistema, se obtendrán valores determinados para a_0 y a_2 que pueden sustituirse en la ecuación de regresión a fin de estimar los valores de x_1 a partir de los valores de x_2. Si estos valores de x_1 estimados ($\hat{}$) a partir de los valores de x_2 los representamos por $\hat{x}_{12}$, tendremos:

$$\hat{x}_{12} = a_0 + a_2 x_2$$

Los valores de x_1 obtenidos a partir de x_2 o sean los valores de $\hat{x}_{12}$ diferirán más o menos de los valores observados de x_1, a menos que la correlación sea perfecta, en cuyo caso coincidirán. Una primera medida y muy burda del error de estimación estará dada precisamente por un promedio de las diferencias entre los valores estimados y los valores observados. Estas diferencias son:

$$x_1 - \hat{x}_{12}$$

y colocamos generalmente como minuendo el valor observado y como sustraendo el estimado en cuanto damos por supuesto que el valor observado es efecto no sólo de la causa representada por la variable dependiente o segunda variable, sino de otras múltiples causas (que, sin embargo, pueden ser negativas en su efecto).

Si estas desviaciones del valor observado y el estimado a partir de la ecuación de regresión las representamos por d_{12}, tendremos:

$$x_1 - \hat{x}_{12} = d_{12}$$

Hemos dicho que una primera medida del error de estimación puede obtenerse mediante el promedio de estas desviaciones. El paso siguiente, consistirá en preguntarse, ¿qué tipo de promedio?

Si se considera que, conforme a lo que acabamos de decir, esas desviaciones del valor observado respecto del estimado pueden ser positivas o negativas, esto quiere decir que el promedio que deberá usarse será aquél que considere los valores absolutos de las desviaciones o aquél otro al través del cual, gracias a la operación que con ellas se realice, haga que las desviaciones lleguen a tener el mismo signo, impidiendo las posibilidades que —en caso de conservarlo— tendrían en cuanto a anularse. Como se señala en las *Técnicas Estadísticas*[1] dicho promedio apropiado es la media cuadrática. En efecto, al emplear la media cuadrática debe principiarse por elevar al cuadrado, y todos los cuadrados —sean de números positivos como de números negativos— tienen signo más.

En estas condiciones, hablaremos de un primer indicador del error de estimación, calculado a partir de los datos mismos en unidades originales, que representaremos por e_{12} y que estará dado por la media cuadrática de las d_{12}, o sea:

$$e_{12} = \sqrt{\frac{\Sigma d_{12}^2}{N}}$$

En efecto, éste es un indicador apropiado, pues cuando los valores estimados coincidan con los observados, las desviaciones serán nulas y el error de estimación (e_{12}) también lo será.

Si se quiere elaborar un índice de correlación a partir del error de estimación, se tendrá que considerar que cuando el error de estimación sea nulo, el índice de correlación deberá de ser máxi-

[1] Obra del autor publicada por el Instituto de Investigaciones Sociales de la U.N.A.M., en 1958.

mo (para comodidad de lectura, más que para otra cosa); que conforme las desviaciones de los valores observados con respecto a los estimados crezcan en valor absoluto y, por lo mismo hagan crecer el error de estimación, disminuirá correspondientemente el índice de correlación, y que cuando el error de estimación sea máximo el índice de correlación deberá permitir una lectura de "nulo".

El error de estimación es una media cuadrática de desviaciones de los valores de una variable con respecto a las estimadas de esa misma variable a partir de otra. O sea que (de acuerdo con las equivalencias entre medias y momentos) el error de estimación elevado al cuadrado es igual al segundo momento de la distribución bivariada. A este segundo momento de la distribución bivariada de las variables 1 y 2 podremos representarlo por $N_2({}_{12})$. La piedra de toque o el elemento de comparación de este segundo momento tiene que serlo otro segundo momento. Será el segundo momento de la distribución univariada de la primer variable, o sea ν_{20}. ν_{20} es igual a:

$$\nu_{20} = \frac{\Sigma x_1^2}{N}$$

En estas condiciones este índice de correlación obtenido a partir de la ecuación de estimación expresada en unidades originarias de la variable estará dado por la fórmula siguiente:

$$= \nu_{20} - e_{12}^2$$

O, como un índice distinto, relacionado con él, pero dado en unidades lineales:

$$\sqrt{\nu_{20} - e_{12}^2}$$

Cálculo de la ecuación de estimación, del error de estimación y de la medida de la correlación de una distribución bivariada, a partir de los datos originales. Correlación no perfecta.

	x_1	x_2	$x_1\,x_2$	x_2^2	$6.87\,x_2$	$\hat{x}_{12}$	$\lvert d_{12}\rvert$	d_{12}^2
	20	3	60	9	20.61	19.69	0.31	0.0961
	40	6	240	36	41.22	40.30	0.30	0.0900
	46	7	322	49	48.09	47.17	1.17	1.3689
	68	10	680	100	68.70	67.78	0.22	0.0484
	75	11	825	121	75.55	74.65	0.35	0.1225
	34	5	170	25	34.35	33.43	0.57	0.3249
Sumas	283	42	2297	340				2.0508
entre 6								.3481

$$\varepsilon = \sqrt{.3481} = .59$$

$$283 = 6\,a_0 + 42\,a_2$$
$$2297 = 42\,a_0 + 340\,a_2$$
$$1981 = 42\,a_0 + 294\,a_2$$
$$316 = 46\,a_2$$
$$6.87 = a_2$$
$$283 = 6\,a_0 + 42 \times 6.87$$
$$283 = 6\,a_0 + 288.54$$
$$283 - 288.54 = 6\,a_0$$
$$-5.54 = 6\,a_0$$
$$-0.92 = a_0$$
$$x_1 = -0.92 + 6.87\,x_2$$

Ecuación de estimación en desviaciones respecto a la medida aritmética

Volveremos a nuestra expresión genérica:

$$x_1 = a_0 + a_2\,x_2$$

Como x_1 y x_2 constituyen una serie bivariada, compuesta, por lo mismo de dos series univariadas, es posible calcular la media de cada una de esas series univariadas. O sea, puede calcularse la media de todas las x_1 del conjunto, y separadamente, la media de todas las x_2 del conjunto. Así, se tendrá:

$\bar{x}_1$ = media aritmética del primer fenómeno

$\bar{x}_2$ = media aritmética del segundo fenómeno.

Calculadas las medias aritméticas de cada una de las distribuciones univariadas componentes de la bivariada, es posible expresar cada valor de cada serie como "desviación con respecto a la media aritmética de dicha serie". En esta forma:

$(x_1 - \bar{x}_1)$ desviación de cada caso del primer fenómeno con respecto a su media aritmética.

$(x_2 - \bar{x}_2)$ desviación de cada observación del segundo fenómeno con respecto a su media aritmética.

Si se substituyen los datos (x_1, x_2) por sus respectivas desviaciones con respecto a la media aritmética de cada uno de ellos, la expresión subsiste en su forma general, pero las constantes de esta nueva expresión serán distintas. Estas nuevas constantes (cuyo valor determinaremos en seguida) las representaremos por A_0 y A_2. La nueva expresión será:

$$(x_1 - \bar{x}_1) = A_0 + A_2 (x_2 - \bar{x}_2)$$

Esta expresión es válida para cada par de valores de x_1 y de x_2 en particular. O sea, que se pueden escribir tantas de estas expresiones como pares de valores haya en el conjunto. Si se suman esas diferentes expresiones, la expresión resultante será una igualdad. En esta igualdad, el primer miembro será igual a la suma de los primeros miembros de las expresiones del tipo señalado. El segundo miembro de la expresión resultante será igual a la suma de los segundos miembros de dichas expresiones. Estas sumas quedarán expresadas mediante el operador sigma mayúscula colocado antes de cada miembro de la expresión:

$$\Sigma(x_1 - \bar{x}_1) = \Sigma(A_0 + A_2[x_2 - \bar{x}_2])$$

La sumatoria del segundo miembro lo es de una suma (de A_0 por una parte y de $A_2(x_2 - \bar{x}_2)$ por otra y, por lo mismo es igual a la suma de las sumatorias de los sumandos:

$$= \Sigma A_0 + \Sigma A_2(x_2 - \bar{x}_2)$$

La sumatoria del primer término de este segundo miembro es la suma de una serie de constantes que es igual a la constante multiplicada por el número de veces que aparece, o sea NA_0. Según esto, el segundo miembro es:

$$= N\,A_0 + \Sigma A_2(x - \bar{x}_2)$$

En el primer miembro se tiene $\Sigma(x_1 - \bar{x}_1)$, o sea, la suma de las desviaciones de las x_1 con respecto a su media y esa suma es, como se sabe, igual con cero. En el segundo miembro figura, análogamente $\Sigma A_2(x - \bar{x}_2)$ o sea, $A_2\Sigma(x - \bar{x}_2)$ o sea, el producto de la constante A_2 por la suma de las desviaciones de las x_2 con respecto a su media; pero esa suma vale cero y, consiguientemente, se anula todo el término. Con lo cual se tiene como primer miembro cero y como segundo sólo NA_0

$$0 = N\,A_0$$

Si se despeja a A_0 se obtiene: $A_0 = 0/N = 0$

Si se sustituye este valor en la expresión general que liga las desviaciones de una variable con respecto a su media con las desviaciones de la otra respecto a la suya,

$$(x_1 - \bar{x}_1) = A_0 + A_2\,(x_2 - \bar{x}_2)$$

se convierte en:

$$(x_1 - \bar{x}_1) = A_2\,(x_2 - \bar{x}_2)$$

O bien, si se emplea como símbolo de desviación de la primera variable con respecto a su media d_1 y como símbolo de la desviación de la segunda variable con respecto a la suya d_2, se obtendrá

$$d_1 = A_2\, d_2$$

Obtención del índice de correlación a partir de la ecuación de regresión expresada mediante desviaciones respecto de la media aritmética.—Si la expresión anterior se somete al proceso de sumación, el primer miembro aparecerá afectado por el sumador y al afectar el segundo miembro con el sumador la constante A_2 aparecerá multiplicando a la variable d_2 previamente afectada por el sumador:

$$\Sigma d_1 = A_2\, \Sigma d_2$$

Sin embargo, debe recordarse que la suma de desviaciones de una variable con respecto a su media es nula, o sea, que la expresión anterior se convertirá en:

$$0 = A_2(0)$$

expresión de la que no es posible obtener ningún valor definido para A_2.

De ahí que convenga, antes de aplicar el sumador, multiplicar ambos miembros de la ecuación originaria por d_2:

$$d_1 d_2 = A_2\, d_2^2$$

En seguida debe aplicarse el sumador a ambos miembros:

$$\Sigma d_1 d_2 = A_2 \Sigma d_2^2$$

Al despejar a A_2, se tiene:

$$A_2 = \frac{\Sigma\, d_1\, d_2}{\Sigma d_2^2}$$

Al sustituir este valor en la ecuación de estimación es posible obtener estimaciones de las desviaciones de la primer variable con respecto a su media ($\hat{d}_1$) a partir de las desviaciones de la segunda variable con respecto a la suya. Estas estimaciones las designaremos por $\hat{d}_{12}$. Según esto:

$$\hat{d}_{12} = A_2 \, d_2$$

Nuevamente, es posible calcular el error de estimación a partir de las desviaciones entre los valores estimados y los observados de las desviaciones de la primer variable. O sea, que partimos de las desviaciones dadas por la fórmula:

$$d_1 - \hat{d}_{12}$$

Para representar estas desviaciones de desviaciones podemos recurrir al símbolo d'_{12}:

$$d_1 - \hat{d}_{12} = d'_{12}$$

El error de estimación será, como en el caso anterior, la media cuadrática de estas desviaciones que representaremos por ε_{12}:

$$\varepsilon_{12} = \sqrt{\frac{\Sigma d'_{12}}{N}}$$

Por consideraciones parecidas a aquellas que se hicieron en el caso del índice de correlación obtenido a partir de los datos expresados en unidades originales, en este caso es necesario traducir la medida inversa de correlación que es el error de estimación a una medida directa de correlación. En este caso, como en el anterior, se tratará de un segundo momento, puesto que el cuadrado del error de estimación ε_{12} es el segundo momento de la distribución bivariada, pero tendrá que ser el segundo momento con respecto a la media aritmética de la primera distribución y no ya como en el caso anterior el segundo momento de esa primera distribución con respecto al origen. La medida de la correlación en este caso será:

$$\mu_{20} - \varepsilon_{12} = R_{12}^2$$

Cálculo de la ecuación de estimación, del error de estimación y de la medida de la correlación de una distribución bivariada, a partir de las desviaciones con respecto a las correspondientes medias aritmética. Correlación no perfecta

	x_1	x_2	d_1	d_2	d_1d_2	d_2^2	$6.87\, d_2$	d_{12}'	$d_{12}'^2$
	20	3	—27.17	—4	108,68	16	—27.48	0.31	0.0961
	40	5	− 7.17	−1	7.17	1	− 6.87	0.30	0.0900
	46	7	− 1.17	0	0.00	0	0.00	1.17	1.3689
	68	10	+20.83	+3	62.49	9	20.61	0.22	0.0484
	75	11	+27.83	+4	111.32	16	27.48	0.35	0.1225
	34	5	−13.17	−2	26.34	4	13.74	0.57	0.3249
Sumas	283	42			315.99	46			2.0508
Medias	47.17	7							

$$\varepsilon_{12} = \sqrt{\frac{34.81}{.3481}} = .59$$

$$d_1 = A_2\, d_2$$

$$\Sigma d_1 d_2 = A_2 \Sigma d_2$$

$$315.99 = 46\, A_2$$

$$6.87 = A_2$$

$$d_{12} = 6.87\, d_2$$

Relación entre el error de estimación de una serie bivariada calculado a partir de los datos originarios y el error de estimación de la misma serie calculado a partir de las desviaciones de cada serie univariada respecto de su media aritmética.—La desviación de los valores observados de la variable dependiente(x_1) con respecto a los valores de esa variable estimados a partir de la variable independiente (valores de x_1 estimados a partir de los valores de x_2, o sea $\hat{x}_{12}$) queda dada por la fórmula:

$$d_{12} = x_1 - \hat{x}_{12}$$

Puesto que toda desviación es una diferencia entre dos valores que se comparan.

En forma parecida, las desviaciones de los valores observados de la desviación de la variable dependiente con respecto a su media ($d_1 = x_1 - \bar{x}_1$) y los valores estimados de dicha desviación a partir de los valores de la desviación de la variable independiente te con respecto a su media ($d_2 = x_2 - \bar{x}_2$) queda dada por la formula:

$$d_{12}' = d_1 - \hat{d}_{12}$$

La estimada de la variable dependiente a partir de la independiente, en la primera fórmula (o sea $\hat{x}_{12}$) puede expresarse por:

$$\hat{x}_{12} = a_0 + a_2x_2$$

Valor que sustituido en la primera expresión, produce:

$$d_{12} = x_1 - a_0 - a_2x_2$$

En forma parecida, la estimada de la desviación de la variable dependiente a partir de la desviación de la independiente (con respecto a sus medias aritméticas correspondientes) puede expresarse como:

$$\hat{d}_{12} = a_2d_2$$

Que sustituido en la segunda expresión, produce:

$$d_{12}' = x_1 - a_2d_2$$

Para poder establecer una relación entre d_{12} y d_{12}', conviene sustituir d_1 y d_2 en esta última expresión por sus equivalentes $(x_1 - \bar{x}_1)$ y $(x_2 - \bar{x}_2)$. Con ello se tiene:

$$d_{12}' = x_1 - \bar{x}_1 - a_2(x_2 - \bar{x}_2) = x_1 - \bar{x}_1 - a_2x_2 - a_2\bar{x}_2$$

Esta expresión puede ordenarse como sigue:

$$d_{12}' = x_1 - a_2x_2 - \bar{x}_1 + a_2\bar{x}_2 = x_1 - a_2x_2 - (\bar{x}_1 - a_2\bar{x}_2)$$

Si en la expresión que da el valor de d_{12}

$$d_{12} = x_1 - a_0 - a_2x_2$$

despejamos a $x_1 - a_2x_2$ que son los dos primeros términos de la expresión a que habíamos llegado un momento antes, tendremos:

$$d_{12} + a_0 = x_1 - a_2x_2$$

Sustituyendo en la expresión alcanzada hacía un momento el valor de $x_1 - a_2x_2$ por su equivalente $d_{12} + a_0$, se tiene:

$$d_{12}' = d_{12} + a_0 - (\bar{x}_1 - a_2\bar{x}_2)$$

A partir de la ecuación de estimación:

$$x_1 = a_0 + a_2x_2$$

se puede obtener la sumatoria correspondiente:

$$\Sigma x_1 = N\, a_0 + a_2\, \Sigma x_2$$

si la sumatoria se divide entre el efectivo, se obtienen las medias correspondientes y en el caso del término Na_0, al dividírsele entre el efectivo N, se convierte en a_0:

$$\bar{x}_1 = a_0 + a_2\, \bar{x}_2$$

que, al despejar a a_0 produce:

$$a_0 = \bar{x}_1 - a_2\bar{x}_2$$

De acuerdo con eso, el paréntesis binomio que aparece como sustraendo de la expresión que dejamos pendiente de d_{12}' puede sustituirse por a_0, con lo cual se tiene:

$$d_{12}' = d_{12} + a_0 - a_0 = d_{12}$$

O sea, que las desviaciones obtenidas en uno y otro caso son iguales. Consiguientemente los errores de estimación obtenidos a partir de los datos originarios (e_{12} obtenida a partir de los datos x_1 y x_2) y a partir de las desviaciones con respecto a las correspondientes medias aritméticas (ε_{12} obtenido a partir de d_1 y d_2) son iguales: $e_{12} = \varepsilon_{12}$.

Obtención del índice de correlación rectilínea a partir de datos expresados en desviaciones con respecto a la media aritmética en unidades sigmáticas.—Partamos de la expresión:

$$d_1 = A_2 \, d_2$$

Si se dividen los dos miembros de la ecuación entre σ_1 (o sea, entre la desviación media cuadrática de la primera variable, que es la media cuadrática de las d_1), se obtiene:

$$\frac{d_1}{\sigma_1} = A_2 \frac{1}{\sigma_1} d_2$$

Si el segundo miembro se multiplica y divide por σ_2 (o sea, si se multiplica y divide por la desviación media cuadrática de la segunda variable, que es la media cuadrática de las d_2) se tiene:

$$\frac{d_1}{\sigma_1} = A_2 \frac{\sigma_2}{\sigma_1} \frac{d_2}{\sigma_2}$$

De acuerdo con una simbología ya aceptada (que hemos usado también en *Técnicas Estadísticas)* y puesto que d_1/σ_1 es la desviación con respecto a la media aritmética (d_1) dividida entre la desviación cuadrática media (σ_1), o sea la desviación sigmática de la primer variable, podremos representarla por δ_1 (delta minúscula sub-índice 1). En forma análoga, d_2/σ_2, desviación sigmática de la segunda variable, puede representarse por δ_2. Con ello, la expresión resulta ser:

$$\delta_1 = \frac{\sigma_2}{\sigma_1} A_2 \, \delta_2$$

Designaremos ésta como la "ecuación de estimación en desviaciones sigmáticas".

En esta ecuación, el conjunto constituido por $A_2\sigma_2/\sigma_1$ podremos representarlo por α_{12}. La expresión se convierte así en:

$$\delta_1 = \alpha_{12}\,\delta_2$$

De acuerdo con esto α_{12} indica cuál es el valor por el que es necesario multiplicar la desviación sigmática de la variable independiente, para obtener la desviación sigmática de la variable dependiente.

A partir de la ecuación de estimación expresada en desviaciones sigmáticas es posible obtener, como en casos anteriores, el error de estimación.

El error de estimación que designaremos como error sigmático de estimación y representaremos por $\varepsilon_{12\sigma}$ será, como en casos anteriores, una desviación media cuadrática, o una media cuadrática de las desviaciones entre los valores observados y los estimados de las desviaciones sigmáticas.

En estas condiciones, como en casos anteriores, comenzaremos por establecer que

$$\delta_1 - \hat{\delta}_{12}$$

o sea la diferencia entre la desviación sigmática de la primer variable tal y como se observa (o calcula a partir de los datos) y la desviación sigmática de esa primer variable estimada a partir de la desviación sigmática de la segunda variable, puede representarse por d_{12}''

$$d_{12}'' = \delta_1 - \hat{\delta}_{12}$$

Si estos valores se elevan al cuadrado, se suman los cuadrados, se divide la suma entre el efectivo y al cociente se le extrae la raíz cuadrada, se obtiene el error sigmático de estimación:

$$\sqrt{\frac{\Sigma d_{12}''^2}{N}} = \varepsilon_{12\sigma}$$

Relación entre los errores sigmáticos y no sigmáticos.—Aun cuando es fácil anticipar el tipo de relación existente entre los errores de estimación calculados a partir de los datos originarios o de las desviaciones aritméticas y el que se obtiene a partir de las desviaciones sigmáticas, conviene puntualizar esta relación.

Partiremos de la expresión que da el valor de d_{12}':

$$d_{12}' = d_1 - \hat{d}_{12} = d_1 - A_2\, d_2$$

y de la expresión para d_{12}''

$$d_{12}'' = \delta_1 - \hat{\delta}_{12}$$

en la cual se puede sustituir δ_1 por d_1/σ_1 y $\hat{\delta}_{12}$ por $\alpha_2\,\delta_2$ o $A_2\, d_2/\sigma_2$ obteniéndose:

$$d_{12}'' = \frac{d_1}{\sigma_1} - A_2\frac{d_2}{\sigma_1} = \frac{d_1 - A_2\, d_2}{\sigma_1} = \frac{d_{12}}{\sigma_1}$$

De este modo, al aplicar el sumador al primero y al último miembro de esta ecuación, se obtienen dos sumas que divididas entre el efectivo y sujetas a radicación producen:

$$\sqrt{\frac{\Sigma\, d_{12}''^2}{N}} = \sqrt{\frac{\Sigma\, d_{12}^2}{N\sigma_1^2}}$$

En el primer miembro de esta igualdad puede reconocerse de inmediato el error sigmático de estimación $\varepsilon_{12\sigma}$. El segundo miembro puede convertirse en:

$$\sqrt{\frac{\Sigma\, d_{12}^2}{N}} \quad \text{por} \quad \sqrt{\frac{1}{\sigma_1^2}}$$

O sea que ese segundo miembro está formado por dos factores, de los cuales el primero es el error no sigmático de estimación en tanto que el segundo (que puede simplificarse en $1/\sigma_1$) es el recíproco de la desviación media cuadrática de la primer variable. O sea, que volviendo a la igualdad de los radicales de donde partimos:

$$\varepsilon_{12\sigma} = \frac{\varepsilon_{12}}{\sigma_1}$$

Indice de correlación obtenido a partir de las desviaciones sigmáticas.—Conforme hemos venido insistiendo, el índice de correlación no es sino la forma de lectura directa de lo mismo que expresa el error de estimación que, en este sentido puede considerarse (complementariamente) como una medida de falta de correlación. Cuando el error de estimación crece, decrece la correlación entre las variables. Cuando el error de estimación decrece, crece la correlación entre las variables. En cambio, inversamente, el índice de correlación *se construye* a modo de que al crecer la correlación crezca el índice y al decrecer la correlación decrezca el índice.

En el caso al que nos enfrentamos, el error de estimación sigmático será cero si el error de estimación no sigmático es nulo, independientemente del valor que pueda tener la desviación cuadrática media de la primer variable. Cuando esto ocurra, el índice de correlación deberá ser máximo. En cambio, cuando el error sigmático de estimación sea máximo, el índice de correlación deberá ser mínimo. El error sigmático de estimación en la correlación rectilínea será máximo cuando el error no sigmático iguale a la desviación cuadrática media de la primer variable (o sea, cuando $\varepsilon_{12} = \sigma_1$). En tal caso, el error sigmático de estimación valdrá 1. Por tanto, este 1, valor máximo, puede servir de piedra de toque de la correlación o falta de correlación. Conforme más se acerque el error de estimación a 1, menos correlación habrá; conforme se aleje el error sigmático de estimación de 1, habrá más correlación. Cuando el error de estimación se aleje al máximo de 1 (o sea, cuando sea cero), la correlación será máxima.

De este modo, puede construirse un índice de correlación estableciendo una desviación entre la máxima del error sigmático y el valor realmente obtenido para la serie estudiada:

$$\varepsilon_{12\sigma_{\text{máx}}} - \varepsilon_{12\sigma}$$

El minuendo es, conforme ya se ha dicho, 1:

$$1 - \varepsilon_{12\sigma}$$

Sin embargo, como el error sigmático de estimación procede de una raíz cuadrada, cuyo signo es anfibológico, a fin de deshacer la anfibología del signo, se establece la relación por diferencia entre los cuadrados correspondientes del error de estimación y el error de estimación máximo y en seguida se extrae la raíz cuadrada para obtener unidades lineales y no cuadráticas:

$$\sqrt{1 - \varepsilon_{12\sigma}^2}$$

A este valor es al que se le conoce como índice de correlación rectilínea r_{12}

$$r_{12} = \sqrt{1 - \varepsilon_{12\sigma}^2}$$

De acuerdo con lo demostrado anteriormente, también es cierto que:

$$r_{12} = \sqrt{1 - \frac{\varepsilon_{12}^2}{\sigma_1^2}}$$

De este modo, el coeficiente de correlación resulta ser la raíz cuadrada del complemento aritmético de la relación entre las variancias (segundos momentos o cuadrados de desviaciones medias cuadráticas) de la distribución bivariada (ε_{12}^2 es la variancia de dicha distribución) y de la distribución univariada de la variable dependiente ($\sigma_1^2 = \mu_{20}$).

Cálculo del coeficiente de correlación a partir de las desviaciones sigmáticas. Caso de una correlación perfecta

	x_1	x_2	d_1	d_2	d_1^2	d_2^2	d_1	δ_2	$\delta_1\delta_2$	δ_2^2
	20	3	−27.17	−4	738.2089	16	−1.43	−1.44	2.0592	2.0736
	40	6	− 7.17	−1	51.4089	1	−0.37	−0.36	0.1368	0.1296
	46	7	− 1.17	0	1.3689	0	−0.06	0.00	0.0000	0.0000
	68	10	20.83	3	433.8889	9	1.09	1.08	1.1772	1.1664
	75	11	27.83	4	774.5089	16	1.46	1.44	2.1024	2.0736
	34	5	−13.17	−2	173.4489	4	−0.69	−0.72	0.4761	0.5184
Sumas	283	42			2172.8334	46			5.9517	5.9616
Medias	47.16	7			362.1389	7.66				
Raíces					19.03	2.77				

$$\delta_1 = \alpha_{12}\,\delta_2$$

$$\Sigma\delta_1\delta_2 = \alpha_{12}\Sigma\delta_2^2$$

$$5.9517 = \alpha_{12}(5.9616)$$

$$\therefore \alpha_{12} = \frac{5.9517}{5.9616} \doteq 1$$

Como es fácil comprender, si α_{12} es igual a 1, el error de estimación será cero, y el índice de correlación valdrá 1

$$r_{12} = 1$$

Relación entre el índice de correlación rectilínea de primer orden y el coeficiente de correlación en unidades sigmáticas.—Se llama "coeficiente de correlación en unidades sigmáticas" o "coeficiente de correlación sigmático" a la constante que multiplica a la desviación de la variable independiente (expresada en unidades sigmáticas) en la ecuación de estimación expresada en unidades sigmáticas. Es decir, se trata del valor que hemos designado por α_2 o por α_{12} (en forma más precisa).

Para encontrar la relación que buscamos, partiremos de la expresión

$$r_{12} = \sqrt{1 - \varepsilon_{12\sigma}^2}$$

En la cual,

$$\varepsilon_{12\sigma} = \Sigma d_{12}'{}^{12}/N$$

Como:

$$d_{12}'{}^{12} = (\delta_1 - \hat{\delta}_{12})^2 = (\delta_1 - \alpha_{12}\,\delta_2)^2 = \delta_1 - 2\,\alpha_{12}\delta_1\,\delta_2 + \alpha_{12}^2\delta_2{}^2$$

Al aplicar a este desarrollo el operador promediador $\Sigma\ldots/N$ también conocido como "esperanza matemática" (o sea, al aplicar el sumador y dividir las sumas entre el efectivo) se obtiene:

$$\varepsilon_{12\sigma}^2 = 1 - 2\alpha_{12}^2 + \alpha_{12}^2$$

Pues, en efecto, la esperanza matemática o promedio de las desviaciones del primer miembro $d_{12}'{}^{12}$ es el cuadrado del error sigmático de estimación; la esperanza matemática de los cuadrados de las desviaciones sigmáticas de la primer variable (δ_1^2) es el segundo momento sigmático de esa variable y como todo segundo momento sigmático, vale 1; la esperanza matemática del producto $\delta_1 \delta_2$ es el coeficiente de correlación sigmático α_{12} que multiplicado por α_{12} produce α_{12}^2 en el segundo término, y, en el último término, la esperanza matemática de los cuadrados de las desviaciones sigmáticas de la segunda variable es el segundo momento sigmático de dicha variable (que vale 1) que multiplicado por el cuadrado del coeficiente sigmático de regresión α_{12} produce la α_{12}^2 que aparece en último término.

La reducción de términos semejantes en ese desarrollo produce:

$$\varepsilon_{12\sigma}^2 = 1 - \alpha_{12}^2$$

Como el valor del índice de correlación elevado al cuadrado (r_{12}^2) se obtiene restando de la unidad el error sigmático, se tiene:

$$r_{12}^2 = 1 - (1 - \alpha_{12}^2) = \alpha_{12}^2$$

O sea que:

$$r_{12} = \alpha_{12}$$

Esto quiere decir que el índice de correlación es igual al coeficiente sigmático de correlación.

CORRELACION RECTILINEA DE ORDEN SUPERIOR AL PRIMERO

Series Sencillas.—Decir que entre una variable considerada como dependiente (x_1) y dos o más variables consideradas como independientes (x_2, x_3...) existe una correlación rectilínea, equivale a afirmar que si se multiplican las primeras potencias (de donde "rectilínea") de las variables independientes por ciertas constantes (a_{12}, a_{13}... designadas por un subíndice doble cuya primera parte indica cuál es la variable dependiente y cuya segunda parte indica cuál es la variable independiente a la que afectan como coeficientes) y se suman los resultados, se obtendrá una estimación de la variable dependiente ($\hat{x}_1$) sujeta a un cierto error de estimación ($\varepsilon_{1.23}$) que, como en el caso de la correlación de primer orden está en relación con un valor conocido como índice de correlación rectilínea de grado n (o, en forma más breve, como "índice de correlación múltiple") que mide la intensidad de la relación entre las variables.

De acuerdo con lo anterior, en la correlación múltiple rectilínea, la expresión que nos servirá como punto de partido será:

$$x_1 = a_0 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 + \ldots\ldots\ldots a_{1n}\,x_n$$

que, en el caso de una correlación rectilínea de segundo orden (o correlación rectilínea tri-variada) que es el que por comodidad trataremos, se reduce a:

$$x_1 = a_0 + a_{12}x_2 + a_{13}x_3$$

Como en el caso anterior, los datos pueden trasformarse en desviaciones con respecto a sus respectivas medias aritméticas, restando de cada dato la media de la serie univariada correspondiente ($d_1 = x_1 - \bar{x}_1$; $d_2 = x_2 - \bar{x}_2$; $d_3 = x_3 - \bar{x}_3$). Al realizar este cambio en las variables, la forma (rectilínea) de la ecuación que las liga no se altera; o sea, que en la nueva ecuación, las desviaciones aparecerán elevadas (como lo estaban los datos originarios) a la primera potencia. Sin embargo, sería fácil demostrar, como se hizo en el caso de la correlación rectilínea de primer orden, que las constantes que aparecen como coeficientes de las variables, cambiarán al tomar desviaciones en vez de datos originales (designaremos por mayúsculas los nuevos coeficientes, de modo que A_0 será el homólogo de a_0, A_{12} el de a_{12}, A_{13} el de a_{13}). Asimismo, como ya se demostró para la correlación rectilínea de primer orden, a_0 se anulará; o sea, que $A_0 = 0$. Es decir, no habrá en la nueva expresión término independiente:

$$d_1 = A_{12}\, d_2 + A_{13}\, d_3$$

Una última forma de expresar lo mismo puede obtenerse del modo siguiente:

Si se divide el primero y el segundo miembro de la ecuación entre la desviación media cuadrática de la variable dependiente (que designaremos por σ_1) tendremos:

$$\frac{d_1}{\sigma_1} = A_{12}\frac{1}{\sigma_1}\, d_2 + A_{13}\frac{1}{\sigma_1}\, d_3$$

Si en seguida, en cada uno de los términos del segundo miembro multiplicamos y dividimos por la correspondiente desviación cuadrática media de la variable independiente que figura en el término (o sea, si se multiplica y divide el primer término del segundo miembro por σ_2, y el segundo término por σ_3), ninguno de esos términos resulta afectado, pero la expresión se convierte en:

$$\frac{d_1}{\sigma_1} = A_{12}\frac{\sigma_2}{\sigma_1}\frac{d_2}{\sigma_2} + A_{13}\frac{\sigma_3}{\sigma_1}\frac{d_3}{\sigma_3}$$

En esta expresión puede observarse que: en el primer miembro figuran desviaciones respecto a la media aritmética entre la desviación cuadrática media de la variable dependiente, o sea las desviaciones sigmáticas de la variable dependiente (que hemos convenido en representar por δ_1); y que, en el segundo miembro figuran asimismo las desviaciones con respecto a la media aritmética correspondiente divididas entre la correspondiente desviación cuadrática media tanto de la segunda como de la tercera variable (ambas independientes), o sea, que en cada uno de esos dos términos figuran las desviaciones sigmáticas de las dos variables independientes (que convenimos en designar por δ_2 y δ_3). De este modo, la expresión puede escribirse:

$$\delta_1 = A_{12}\frac{\sigma_2}{\sigma_1}\delta_2 + A_{13}\frac{\sigma_3}{\sigma_1}\delta_3$$

Podemos, finalmente expresar los coeficientes de las desviaciones sigmáticas que aparecen en esta ecuación en forma más comprimida si hacemos:

$$A_{13}\frac{\sigma_3}{\sigma_1} = \alpha_{13}$$

$$A_{12}\frac{\sigma_2}{\sigma_1} = \alpha_{12}$$

Precaución.—Quien haya seguido hasta aquí los desarrollos correspondientes a las fórmulas de la correlación rectilínea de primer orden y las de la correlación rectilínea de orden superior al primero, puede sentirse tentado, en vista de la semejanza exterior de los primeros miembros de estas igualdades con una de las fórmulas que dejamos consignadas en el capítulo de la correlación de primer orden $\left(A_{12}\frac{\sigma_2}{\sigma_1} = r_{12}\right)$, a establecer unas igualdades como las siguientes:

$$\alpha_{12} = r_{12}$$

$$\alpha_{13} = r_{13}$$

Los errores de conceptuación a los que puede conducir esta identificación simbológica fundada en una apariencia más que en una realidad (como diremos en seguida) puede seguirse al través de las páginas siguientes que, *en cuanto incorrectas* están impresas en otro tipo distinto del general de estas páginas.

¿De dónde procede el error? Ya lo anticipábamos; de una abusiva identificación entre signos que no corresponde a una pertinente identificación entre los conceptos que los símbolos recubren. En una identificación a base de apariencias y no de realidades.

El error procede de que hemos usado A_{12} para representar el coeficiente de la variable 2 en la ecuación de estimación de una correlación de *primer orden* y también hemos usado A_{12} para representar el coeficiente de dicha variable en una ecuación de estimación de una correlación de *segundo orden* y que hemos hecho cosa parecida con A_{13} que muy bien podría representar el coeficiente de la variable 3, en una correlación de *primer orden*, entre la variable 1 y la variable 3, y que representa el coeficiente de la variable 3 en una correlación de *segundo orden* entre las variables 1, 2, 3.

A partir de la página siguiente, *en tipo distinto*, seguimos el desarrollo erróneo a que pudo conducir esa errónea identificación, con el fin de que el estudiante pueda evitar en un momento dado errores de este tipo, valorando anticipadamente la importancia que los mismos pueden tener. Una vez puesta de manifiesto la importancia de esta precaución de usar símbolos tan específicos y distintos como sea posible para realidades distintas, trataremos de mostrar cómo deshacer el equívoco mediante una simbología más precisa.

Ejemplo del error de conceptuación a que puede conducir una simbología imprecisa.—*Habíamos convenido en designar los pri-*

meros miembros de las igualdades siguientes por las literales que aparecen en sus segundos miembros:

$$A_{12}\frac{\sigma_2}{\sigma_1}=\alpha_{12}$$

$$A_{13}\frac{\sigma_3}{\sigma_1}=\alpha_{13}$$

Es fácil ver, en esta forma, la relación que existe entre estos coeficientes y el índice de correlación rectilínea de primer orden. El primer coeficiente, es el índice de correlación rectilínea de primer orden entre la variable dependiente y la primera de las independientes. El segundo coeficiente, es el índice de correlación de primer orden entre la variable dependiente y la segunda de las independientes. Consiguientemente, el primer término del segundo miembro de la ecuación de estimación en desviaciones sigmáticas de esta correlación trivariada, es la estimada de la desviación sigmática de la variable dependiente a partir de la desviación sigmática de la primera variable independiente, y el segundo término del segundo miembro es la estimada de esa misma desviación sigmática de la variable dependiente a partir de la desviación sigmática de la segunda variable independiente. En suma, que en esta correlación, la estimada de la desviación sigmática de la variable dependiente ($\hat{\delta}_1$) *es igual a la suma de las estimadas de la propia desviación sigmática a partir de las desviaciones sigmáticas de las variables independientes.*

Sustituyendo los coeficientes de las desviaciones sigmáticas por los equivalentes que acabamos de designar convencionalmente por α_{12} *y* α_{13} *la expresión presenta la siguiente apariencia:*

$$\delta_1=\alpha_{12}\,\delta_2+\alpha_{13}\,\delta_3$$

En esta expresión, las alfas indican los valores constantes por los que hay que multiplicar un valor dado de la desviación sigmá-

tica de la primer variable independiente y un valor dado de la desviación sigmática de la segunda variable independiente para obtener el correspondiente valor estimado de la variable dependiente. En vista de que el valor de la desviación sigmática de la variable dependiente es estimado *y no* observado *para mantener el rigor simbológico, la expresión anterior debe escribirse:*

$$\hat{\delta}_1 = \alpha_{12}\,\delta_2 + \alpha_{13}\,\delta_3$$

En esta expresión, el acento circunflejo de δ_1 significa precisamente que se trata de un valor estimado y no observado.

Pero, hay más. Puesto que la estimación de los valores de la desviación sigmática de la variable dependiente puede hacerse a partir de diferentes ecuaciones de estimación (de una ecuación propia de una correlación de primer orden o de una de segundo orden) y puede convenir establecer relaciones entre estas diferentes formas de estimación, emplearemos la siguiente simbología que, si bien puede parecer más compleja es, simultáneamente más clara o específica:

$\hat{\delta}_{1.2}$ representa la estimada de δ_1 con base en la correlación de primer orden entre las variables 1 y 2

$\hat{\delta}_{1.3}$ representa la estimada de δ_1 con base en la correlación de primer orden entre las variables 1 y 3

$\hat{\delta}_{1.23}$ representa la estimada de δ_1 con base en la correlación de segundo orden entre la variable 1 (dependiente) y las variables 2 y 3 (independientes)

De acuerdo con esta simbología, podemos escribir en símbolos lo que hemos expresado antes con palabras:

$$\hat{\delta}_{1.23} = \hat{\delta}_{1.2} + \hat{\delta}_{1.3}$$

Ecuación equivalente de la que, con otra simbología menos precisa dejamos asentada antes, y cuya forma definitiva por el momento sería:

$$\hat{\delta}_{1.23} = \alpha_{12}\,\delta_2 + \alpha_{13}\,\delta_3$$

O, incluso:

$$\hat{\delta}_{1.23} = r_{12}\,\delta_2 + r_{13}\,\delta_3$$

Que es la ecuación de estimación en unidades sigmáticas de una correlación rectilínea de segundo orden.

En general, puede escribirse como ecuación de estimación, en unidades sigmáticas, de una correlación rectilínea de orden superior al segundo:

$$\hat{\delta}^{1.23n} = r_{12}\,\delta_2 + r_{13}\,\delta_3 + \ldots\ldots\ldots r_{1n}\,\delta_n$$

O bien, en una forma más condensada:

$$\hat{\delta}_{1.23_n} = \sum_{n=1}^{n} r_{1n}\,\delta_n$$

Ajuste de la simbología de la correlación múltiple o de orden superior al primero.—En la correlación de primer orden empleamos los siguientes símbolos:

$$\begin{aligned} \hat{x}_1 &= a_0 + a_{12}x_2 \\ \hat{d}_1 &= \quad A_{12}d_2 \\ \hat{\delta}_1 &= \quad r_{12}\,\delta_2 \end{aligned}$$

Si la estimación de x_1, de d_1, o de δ_1 se hace no a partir de una ecuación de primer grado con una variable (x_2), sino a partir de una ecuación también de primer grado, pero con dos variables (o sea con una variable x_3 además de la variable x_2 ya considerada), el coeficiente de la variable considerada en la correlación de primer orden (x_2) ya no será el que figuraba en la ecuación de estimación de la correlación de primer orden (ya no será a_{12}) sino un coeficiente distinto (a menos que se trate de un caso particular que analizaremos ulteriormente). A ese nuevo coeficiente distinto podríamos designarlo arbitrariamente como a_{12}'. Sin embargo, la comodidad de la distinción no representa —simultáneamente— una claridad simbológica que tenga su correlato en una precisión conceptual creciente. En cambio si se utiliza un símbolo como $a_{12.3}$ esto podemos interpretarlo en el sentido de que se trata de *una constante* (a, una primera letra del alfabeto, según la convención que quiere que de preferencia se representen las constantes por las primeras letras del alfabeto) que sirve como *coeficiente a la segunda variable* (el 2 del subíndice) en una ecuación de *estimación de una primer variable* (el 1 del subíndice) en la cual figura también, como variable independiente, una tercer variable (el 3 del subíndice). En forma parecida, si la correlación es entre una variable dependiente y más de 2 independientes, el coeficiente podrá representarse por $a_{12.3...n.}$ Análogamente, para el coeficiente de la segunda variable independiente (o sea, la tercer variable que interviene en la correlación x_3), el coeficiente de una correlación de segundo orden podrá escribirse $a_{13.2}$ y, si se trata del coeficiente de esa misma variable en una correlación de orden superior, podrá registrarse como $a_{13.2...n.}$

Las designaciones de los coeficientes en las ecuaciones en que intervienen no datos originales, sino desviaciones con respecto a las medias aritméticas o desviaciones con respecto a las medias aritméticas en unidades sigmáticas, son completamente paralelas. De este modo, las expresiones de la correlación rectilínea de segundo orden se tendrá que escribir más precisamente:

$$\hat{x}_{1.23} = a_0 + a_{12.3}\, x_2 + a_{13.2}\, x_3$$

En donde el primer miembro representa: valor estimado (acento circunflejo) de la variable (x) primera (1 del subíndice) hecha a partir de sus relaciones con otras dos variables (2 y 3 del subíndice).

$$\hat{d}_{1.23} = A_{12.3}d_2 + A_{13.2}d_3$$

Y en desviaciones sigmáticas:

$$\hat{\delta}_{1.23} = \alpha_{12.3}\,\delta_2 + \alpha_{13.2}\,\delta_3$$

Con este ajuste de la simbología, el error contra el que precavimos y que mostramos en páginas anteriores, no puede producirse. Al obtener —como haremos en seguida— los valores de las alfas podrá verse el serio error en que se caería mediante tal identificación abusiva de los símbolos a la que no correspondía una correlativa identificación legítima de los conceptos simbolizados.

Obtención de los coeficientes de la ecuación de estimación en la correlación múltiple o de grado superior al primero.—Partiremos de la ecuación que acabamos de registrar, escrita de acuerdo con la simbología recién ajustada:

$$\hat{\delta}_{1.23} = \alpha_{12.3}\,\delta_2 + \alpha_{13.2}\,\delta_3$$

Si multiplicamos tanto el primer miembro como los dos términos del segundo miembro por δ_2 y en seguida aplicamos el sumador u operador sigma mayúscula, tendremos:

$$\Sigma\,\hat{\delta}_{1.23}\,\delta_2 = \alpha_{12.3}\Sigma\delta_2{}^2 + \alpha_{13.2}\,\Sigma\delta_2\delta_3{}^2$$

Si multplicamos en forma parecida por δ_3 y aplicamos el sumador, tendremos:

$$\Sigma\hat{\delta}_{1.23}\,\delta_3 = \alpha_{12.3}\Sigma\delta_2\delta_3 + \alpha_{13.2}\Sigma\delta_2{}^2$$

A partir de este sistema de dos ecuaciones con dos incógnitas ($\sigma_{12.3}$ y $\alpha_{13.2}$) podemos determinar los valores de dichas incógnitas.

Cálculo del índice de correlación rectilínea de segundo orden entre las variables x_1, x_2, x_3

	DATOS			ELABORACIONES								
	x_1	x_2	x_3	d_1	d_2	d_3	d_1^2	d_2^2	d_3^2	δ_1	δ_2	δ_3
	13	4	3	0	0	0	0	0	0	0	0	0
	8	2	2	−5	−2	−1	25	4	1	− .81	−1.55	− .46
	25	4	7	12	0	4	144	0	16	1.95	0	1.85
	15	3	4	2	−1	1	4	1	1	.33	− .77	.46
	6	6	0	−7	2	−3	49	4	9	−1.14	1.55	−1.39
	11	5	2	−2	1	−1	4	1	1	− .33	.77	− .46
Sumas	78	24	18				226	10	28			
Medias	13	4	3				37.7	1.67	4.67			
Raíces							6.14	1.29	2.16			
o Desviaciones Medias cuadráticas							σ_1	σ_2	σ_3			

Ecuación de estimación de la variable dependiente en función de las independientes:

$$\hat{\delta}_1 = \alpha_{12.3}\,\delta_2 + \alpha_{13.2}\,\delta_3$$

Ecuaciones de interpolación para encontrar los coeficientes de correlación (o sean las alfas):

$$\Sigma\delta_1\,\delta_2 = \alpha_{123}\Sigma\delta_2^2 + \alpha_{13.2}\Sigma\delta_2\,\delta_3$$

$$\Sigma\delta_1\,\delta_3 = \alpha_{12.3}\Sigma\,\delta_2\,\delta_3 + \alpha_{13.2}\Sigma\,\delta_3^2$$

Tabulación destinada a la obtención de los valores requeridos por las ecuaciones de interpretación.

	$\delta_1\,\delta_2$	$\delta_1\,\delta_3$	δ_2^2	$\delta_2\,\delta_3$	δ_3^2
	0.0000	0.0000	0.0000	0.0000	0.0000
	1.2555	0.3726	2.4025	−0.7130	0.2116
	0.0000	3.6075	0.0000	0.0000	3.4225
	−0.2541	0.1518	0.5929	−0.3542	0.2116
	−1.7670	1.5846	2.4025	−2.1545	1.9321
	−0.2541	0.1518	0.5929	−0.3542	0.2116
Sumas	−1.0197	5.8683	5.9908	−2.1499	5.9894

Sustituidos estos valores en las ecuaciones de interpolación, se tiene:

$$-1.0197 = 5.9908\,\alpha_{12.3} + (-2.1499)\,\alpha_{13.2}$$

$$5.8683 = (-2{,}1499)\,\alpha_{12.3} + 5.9894\,\alpha_{13.2}$$

El eliminante del sistema será:

$$\begin{vmatrix} 5.9908 & -2.1499 \\ -2{,}1499 & 5.9894 \end{vmatrix} = 5.9908 \times 5.9894 - (-2.1499)^2 =$$

$$= 35.8813 - 4.6221 = 31.2592$$

El determinante del numerador de $\alpha_{12.3}$ es:

$$\begin{vmatrix} -1.0197 & -2.1499 \\ 5.8683 & 5.9894 \end{vmatrix} = -6.1074 - (-12.6163) =$$
$$= 12.6163 - 6.1074 =$$
$$= 6.5089$$

El determinante del numerador de $\alpha_{13.2}$ es:

$$\begin{vmatrix} 5.9908 & -1.0197 \\ -2.1499 & 5.8683 \end{vmatrix} = 35.1558 - 2.1923 =$$
$$= 32.8635$$

Consiguientemente los valores de los coeficientes de correlación son:

$$\alpha_{12.3} = \frac{6.5089}{31.2592} = 0.2082$$

$$\alpha_{13.2} = \frac{32.9635}{31.2592} = 1.0545$$

Por tanto, puede escribirse:

$$\hat{\delta}_1 = 0.2082\,\delta_2 + 1.0545\,\delta_3$$

Tabulación para el cálculo de los valores estimados en δ_1
a partir de los valores de δ_2 y δ_3

δ_2	$0.2082\,\delta_2$	δ_3	$1.0545\,\delta_3$	$0.2082\,\delta_2 + 1.0545\,\delta_3 = = \delta_1$
0	0.000000	0	0.000000	0.000000
−1.55	−0.322710	−0.46	−0.485070	—0.807780
0	0.000000	1.85	+1.950825	1.950825
−0.77	−0.160314	0.46	+0.485070	0.324756
1.55	+0.322710	−1.39	−1.465755	—1.143045
0.77	+0.160314	−0.46	−0.485070	0.324756

Cálculo de las diferencias entre los valores estimados de $\hat{\delta}_1$ y sus valores calculados δ_1 a partir de los datos originarios.

$\hat{\delta}_1$	δ_1	$\delta_1 - \hat{\delta}_1$	$(\delta_1 - \hat{\delta}_1)^2$
0.000000	0	0	0.000000000000
−0.807780	−0.81	0.002220	0.000004928400
1.950825	1.95	−0.000825	0.000000680625
0.324756	0.33	0.005244	0.000027499536
−1.143045	−1.14	−0.003045	0.000009272025
0.324756	−0.33	0.005244	0.000027499536
		Suma de las diferencias	0.000069880122
		entre 5	0.00001397600244
		Raíz cuadrada:	0.003738 error de estimación

Como el error de estimación es, para propósitos prácticos, igual a cero, el índice de correlación r, será prácticamente igual a 1.

CORRELACION PARCIAL

Si de las desviaciones sigmáticas (δ) observadas de una variable considerada como dependiente (δ_1) se restan las desviaciones sigmáticas estimadas (de esa misma variable), obtenidas a partir de una ecuación de estimación propia del proceso de correlación con otra variable independiente ($\hat{\delta}_{1.2}$), se obtiene un nuevo tipo de desviación que representaremos por delta mayúscula (Δ) y que afectaremos de un subíndice igual al que afecta la estimada de la desviación sigmática de la variable dependiente (o sea, en el caso, 1.2 que afecta a $\hat{\delta}_{1.2}$). De este modo, $\Delta_{1.2}$ es un símbolo que debe interpretarse como "desviación de las desviaciones sigmáticas estimadas respecto a las calculadas" (Δ) "de la variable 1, a partir de la variable 2" (1.2). La fórmula corespondiente será, de acuerdo con lo dicho:

$$\Delta_{1.2} = \delta_1 - \hat{\delta}_2$$

En forma análoga, podemos establecer las dos fórmulas siguientes:

$$\Delta_{1.3} = \delta_1 - \hat{\delta}_{1.3}$$

$$\Delta_{2.3} = \delta_2 - \hat{\delta}_{2.3}$$

Si sustituimos las estimadas (o deltas minúsculas con circunflejo) por sus valores tomados de las correspondientes ecuaciones de estimación, tendremos:

$$\Delta_{1.3} = \delta_1 - r_{13}\, \delta_3$$

$$\Delta_{2.3} = \delta_2 - r_{23}\, \delta_3$$

Entre la desviación de la estimada sigmática de 2 en 3 ($\Delta_{2.3}$) y la desviación de la estimada sigmática de 1 en 3 ($\Delta_{1.3}$) puede establecerse una correlación al través de una ecuación de estimación en la que figurará un índice de correlación que provisionalmente designaremos por $r_{1.3:2.3}$

Como en todos los casos anteriores, comenzaremos por establecer una ecuación de estimación. La ecuación de estimación no sigmática a partir de los valores originales será:

$$\Delta_{1.3} = a_0 + a_{1.3:2.3}\,\Delta_{2.3}$$

Si hemos de trabajar con desviaciones de los valores originarios (deltas mayúsculas en este caso) con respecto a sus correspondientes medias aritméticas (medias aritméticas de dichas deltas mayúsculas) comenzaremos por determinar los valores de dichas medias como sigue:

$$\text{Si } \Delta_{1.3} = \delta_1 - \hat{\delta}_{1.3} \text{ y } \Delta_{2.3} = \delta_2 - \hat{\delta}_{3.2}$$

y las medias aritméticas correspondientes son (como siempre) sumas de variables (en este caso, sumas de deltas mayúsculas) entre el efectivo (en estos casos de series sencillas, número de pares de datos):

$$\bar{\Delta}_{1.3} = \frac{\Sigma\Delta_{1.3}}{N} \qquad \bar{\Delta}_{2.3} = \frac{\Sigma\Delta_{2.3}}{N}$$

Al sustituir en la primera fórmula los valores de las $\Delta_{1.3}$ se obtiene:

$$\bar{\Delta}_{1.3} = \frac{\Sigma(\delta_1 - \hat{\delta}_{1.3})}{N} = \frac{\Sigma(\delta_1 - r_{13}\,\delta_3)}{N} =$$

$$= \frac{\Sigma\delta_1 - r_{13}\,\Sigma\delta_3}{N}$$

Como $\Sigma\delta_1 = 0$ y $\Sigma\delta_3 = 0$ (porque la suma de desviaciones respecto a la media aritmética es siempre nula), podremos establecer que:

$$\bar{\Delta}_{1.3} = 0$$

Y, en forma análoga, que:

$$\bar{\Delta}_{2.3} = 0$$

Si las medias aritméticas de las deltas mayúsculas son nulas, ello quiere decir que las desviaciones de esas mismas deltas mayúsculas con respecto a sus propias medias serán las mismas deltas mayúsculas O sea que:

$$d_{1.3} = \Delta_{1.3} - \bar{\Delta}_{1.3} = \Delta_{1.3} - 0 = \Delta_{1.3} \quad \text{y} \quad d_{2.3} = \Delta_{2.3}$$

En tales condiciones, la ecuación de estimación, dada en desviaciones respecto a las medias aritméticas, en que se convertiría la expresión primitiva:

$$d_{1.3} = A_{1.3:2.3}\, d_{2.3}$$

acabará convirtiéndose en:

$$\Delta_{1.3} = A_{1.3:2.3}\, \Delta_{2.3}$$

A esta ecuación de estimación de la variable 1.3 a partir de la variable 2.3 en desviaciones respecto a là media aritmética, corresponde una ecuación inversa de estimación de la variable 2.3 a partir de la variable 1.3 en desviaciones respecto a la media aritmética:

$$\Delta_{2.3} = A_{1.3:2.3}\, \Delta_{1.3}$$

Finalmente, es posible pensar aquí también en una ecuación de estimación sigmática. Para obtenerla, será necesario que comencemos por calcular las desviaciones cuadráticas medias de las variables que se están correlacionando ($\Delta_{1.3}$ y $\Delta_{2.3}$).

Como la desviación cuadrática media es igual a la media cuadrática de las desviaciones:

$$\sigma_{1.3} = \sqrt{\frac{\Sigma d_{1.3}^2}{N}} = \sqrt{\frac{\Sigma \Delta_{1.3}^2}{N}}$$

Si dividimos cada delta mayúscula entre la correspondiente sigma así obtenida, obtendremos la correspondiente delta minúscula o desviación sigmática ($\delta_{1.3}$, $\delta_{2.3}$).

Entre estas desviaciones sigmáticas es posible establecer una ecuación sigmática de estimación, en la cual el coeficiente de la variable independiente será el índice de correlación de las variables (en forma análoga a como ocurrió en la correlación rectilínea de primer orden):

$$\delta_{1.3} = r_{1.3:2.3}\, \delta_{2.3}$$

Tanto para la correlación rectilínea de primer orden como para la correlación parcial puede demostrarse que la correlación directa de la primer variable a partir de la segunda como la inversa que parte de la primer variable y llega a la segunda, producen el mismo coeficiente de correlación sigmático o índice de correlación. Es decir, que puede escribirse también:

$$\delta_{2.3} = r_{1.3:2.3}\, \delta_{1.3}$$

Obtención del índice de correlación parcial a partir de la ecuación no sigmática de estimación de los residuos de las estimaciones.—Si tomamos como ecuación de estimación la que permite estimar los residuos de los valores observados de la variable dependiente respecto de sus estimaciones a partir de la variable eliminada ($\Delta_{1.3}$) con base en los residuos de los valores observados de la variable independiente respecto de sus estimaciones a partir de la eliminada ($\Delta_{2.3}$) tendremos:

$$\Delta_{1.3} = A_{1.3:2.3}\, \Delta_{2.3}$$

En esta ecuación, el único valor por determinar es el del coeficiente de correlación parcial $A_{1.3:2.3}$. Si aplicáramos un operador sigma a ambos miembros tropezaríamos con la dificultad con la que ya tropezamos en el cálculo de la correlación rectilínea: se producirían sumas de desviaciones que se anularían y cuyo cociente sería indeterminado. Por ello, como en aquella ocasión, comenzaremos por multiplicar ambos miembros por $\Delta_{2.3}$. Con ello obtendremos:

$$\Delta_{1.3}\,\Delta_{2.3} = A_{1.3:2.3}\,\Delta_{2.3}{}^2$$

Aplicando el sumador a ambos miembros de esta ecuación, se obtiene:

$$\Sigma\Delta_{1.3}\,\Delta_{2.3} = A_{1.3:2.3}\,\Sigma\Delta_{2.3}{}^2$$

Al despejar a $A_{1.3:2.3}$ se tiene:

$$A_{1.3:2.3} = \frac{\Sigma\Delta_{1.3}\,\Delta_{2.3}}{\Sigma\Delta_{2.3}{}^2}$$

Mediante el valor del coeficiente de correlación parcial y su sustitución en la ecuación de estimación, puede obtenerse una serie de estimadas de $\Delta_{1.3}$ a partir de $\Delta_{2.3}$ que diferirán más o menos de las calculadas, y que representaremos por $\hat{\Delta}_{1.3:2.3}$. Gracias a estas estimadas, es posible calcular una serie de residuos:

$$\Delta_{1.3} - \hat{\Delta}_{1.3:2.3}$$

cuya desviación media cuadrática no será sino el error no sigmático de estimación de la correlación parcial, a partir del cual, como en la correlación rectilínea de primer orden, puede obtenerse el correspondiente índice que en este caso será el índice correlación parcial $r_{13:23}$ que también representamos por $r_{1(3)2}$ o $r_{1.23}$.

Obtención del índice de correlación parcial a partir de la ecuación sigmática de estimación de los residuos de las estimaciones.—

Tomaremos la ecuación sigmática de estimación de los residuos que resultan de restar de los valores observados de la variable dependiente, los valores calculados a partir de la eliminada ($\delta_{1.3}$). En ella, dichos residuos se estiman a partir de otros residuos. Estos otros se obtienen al restar de los valores observados de la variable independiente los valores de ésta estimados a partir de la variable eliminada ($\delta_{2.3}$). Así, tendremos:

$$\delta_{1.3} = r_{1.3:2.3}\,\delta_{2.3}$$

Si multiplicamos ambos miembros por $\delta_{2.3}$, tendremos:

$$\delta_{1.3}\,\delta_{2.3} = r_{1.3:2.3}\,\delta_{2.3}^{\,2}$$

Al aplicar el sumador a los dos miembros de esta ecuación, se tiene:

$$\Sigma\delta_{1.3}\delta_{2.3} = r_{1.3:2.3}\,\Sigma\delta_{2.3}^{\,2}$$

De acuerdo con esto, tendremos, finalmente:

$$r_{1.3:2.3} = \frac{\Sigma\delta_{1.3}\,\delta_{2.3}}{\Sigma\delta_{2.3}^{\,2}}$$

Que es el valor del coeficiente-índice de correlación parcial de la primer variable en la segunda, con eliminación de la tercera.

$$r_{1.3:2.3} = \frac{\Sigma\delta_{1.3}\,\delta_{2.3}}{\Sigma\delta_{2.3}^{\,2}}$$

Por sustitución de las δ este índice también puede expresarse:

$$r_{1.3:2.3} = \frac{\Sigma\Delta_{1.3}\,\Delta_{2.3}}{\Sigma\Delta_{2.3}^{\,2}}\;\frac{\sigma_{1.3}}{\sigma_{2.3}}$$

Simplificación de la fórmula del índice de correlación parcial.—Si se considera que, como hemos mostrado anteriormente:

$$d_{1.3} = \Delta_{1.3} \qquad d_{2.3} = \Delta_{2.3}$$

La fórmula para el cálculo de la desviación media cuadrática:

$$\sigma_{1.3} = \sqrt{\frac{\Sigma d_{1.3}^2}{N}}$$

se transforma en:

$$\sigma_{1.3} = \sqrt{\frac{\Sigma \Delta_{1.3}^2}{N}}$$

Y, análogamente:

$$\sigma_{2.3} = \sqrt{\frac{\Sigma \Delta_{2.3}^2}{N}}$$

En estas condiciones, la relación de desviaciones sigmáticas $\sigma_{1.3}/\sigma_{2.3}$ que figura en la fórmula del índice de correlación parcial se convierte en:

$$\frac{\sigma_{1.3}}{\sigma_{2.3}} = \sqrt{\frac{\Sigma \Delta_{1.3}^2}{\Sigma \Delta_{2.3}^2}}$$

Al sustituir el equivalente de esta relación en la fórmula, se tiene:

$$r_{1(3)2} = \frac{\Sigma \Delta_{2.3}\, \Delta_{1.3}}{\Sigma \Delta_{1.3}^2} \sqrt{\frac{\Sigma \Delta_{1.3}^2}{\Sigma \Delta_{2.3}^2}}$$

Pero, el factor radicalizado puede descomponerse en un cociente de radicales, y el radical de numerador $\sqrt{\Sigma \Delta_{1.3}^2}$ se reduce a la unidad con uno de los dos radicales idénticos a él que constituyen el denominador de la otra fracción $(\Sigma \Delta_{1.3}^2)$ O sea, que el radical del numerador desaparece y $\Sigma \Delta_{1.3}^2$ del denominador queda radicalizado. El resultado final es:

$$r_{1(3)2} = \frac{\Sigma \Delta_{2.3}\, \Delta_{1.3}}{\sqrt{\Sigma \Delta_{2.3}^2}\ \sqrt{\Sigma \Delta_{1.3}^2}}$$

¿Qué representa $r_{1(3)2}$? $r_{1(3)2}$ de acuerdo con el proceso que hemos seguido, es el índice de correlación rectilínea entre dos desviaciones de valores sigmáticos observados y calculados.

En particular, $\Delta_{1.3}$ es la porción de la desviación sigmática observada, de la variable dependiente 1, que no llega a explicarse a partir de su correlación con la variable independiente 3.

En forma análoga, $\Delta_{2.3}$ es la porción sigmática observada de la variable dependiente 2 que no llega a explicarse a partir de su correlación con la variable independiente 3.

O sea, que las deltas mayúsculas representan las variaciones de la magnitud de las variables 1 y 2 que no se explican al través de las variaciones de magnitud de la variable 3.

En el momento en que, al través del cálculo de $r_{1(3)2}$ establecemos la relación estadística entre estos valores, estamos estimando la intensidad de la relación que liga a las variables 1 y 2 por encima de las variaciones de las mismas explicadas por las variaciones de magnitud de la variable 3.

Al índice $r_{1(3)2}$ que elimina la influencia de 3 y calcula la influencia que 2 tiene en 1 por encima de dicha influencia, se le conoce como índice de correlación parcial de 1 en 2 con eliminación de 3.

Naturalmente, en tratándose de las tres variables 1,2,3, es posible establecer no uno, sino los siguientes índices de correlación parcial (a los que corresponderán otras tantas ecuaciones de estimación de residuos de la estimada sigmática):

$$r_{1(3)2}$$
$$r_{1(2)3}$$
$$r_{2(1)3}$$
$$r_{3(2)1}$$
$$r_{3(1)2}$$
$$r_{2(3)1}$$

El carácter de cada investigación concreta que se haga, determinará cuáles, de entre ellos, habrá que calcular.

III. RELACIONES ENTRE LAS DIFERENTES MEDIDAS QUE INTERVIENEN EN LA CORRELACION

RELACIONES ENTRE LAS CONSTANTES ESTADISTICAS DE LA CORRELACION RECTILINEA DE PRIMER ORDEN

Fórmula del coeficiente de correlación.—A partir de la ecuación de estimación (que tomaremos en unidades sigmáticas para mayor facilidad), o sea, a partir de:

$$\delta_1 = \alpha_{12}\,\delta_2$$

puede obtenerse, mediante multiplicación de ambos miembros por δ_2 y aplicación ulterior del sumador a ambos miembros:

$$\Sigma\delta_1\,\delta_2 = \alpha_{12}\Sigma\,\delta_2^2$$

De donde, despejando a α_{12}, se obtiene:

$$\alpha_{12} = \frac{\Sigma\delta_1\,\delta_2}{\Sigma\delta_2^2}$$

Como en el denominador figura la suma de los cuadrados de desviaciones sigmáticas y dicha suma no es sino el segundo momento sigmático que ha sido multiplicado por el efectivo de la distribución, N, todo el denominador se reduce a N (ya que el segundo momento *sigmático* vale siempre 1 como puede verse en las *Técnicas Estadísticas)*. En tales condiciones, la fómula para el coeficiente de correlación se reduce a:

$$\alpha_{12} = \frac{\Sigma\delta_1\,\delta_2}{N}$$

Relación entre el error de estimación y el coeficiente de correlación.—En cuanto el error de estimación es la media cuadrática de las desviaciones entre los valores observados y los estimados de las desviaciones sigmáticas de la primer variable, puede escribirse, por definición:

$$\varepsilon_{12\sigma} = \sqrt{\frac{\Sigma(\delta_1 - \hat{\delta}_{12})^2}{N}}$$

O bien:

$$\varepsilon_{12\sigma}^{2} = \frac{\Sigma(\delta_1 - \hat{\delta}_{12})^2}{N}$$

Si se ejecutan las operaciones de exponenciación del binomio encerrado en el paréntesis y de aplicación del sumador a los resultados, se obtiene:

$$\frac{\Sigma\delta_1^{2} - 2\Sigma\delta_1\,\hat{\delta}_{12} + \Sigma\hat{\delta}_{12}^{2}}{N}$$

Al sustituir $\hat{\delta}_{12}$ por su equivalente tomado de la ecuación de estimación $\alpha_{12}\delta_2$ se obtiene:

$$\frac{\Sigma\delta_1^{2} - 2\alpha_{12}\Sigma\delta_1\,\delta_2 + \alpha_{12}^{2}\,\Sigma\delta_2^{2}}{N}$$

Pero, si como dijimos en el apartado anterior $\alpha_{12} = \Sigma\delta_1\,\delta_2/N$, $\delta_1\delta_2$ será igual a $N\alpha_{12}$, valor que sustituido, producirá:

$$\frac{\Sigma\delta_1^{2} - 2\alpha_{12}(N\alpha_{12}) + \alpha_{12}^{2}N}{N} = \frac{\Sigma\delta_1^{2} - 2\,N\alpha_{12}^{2} + N\alpha_{12}^{2}}{N}$$

$$= \frac{N - N\alpha_{12}^{2}}{N}$$

Porque, en efecto, tanto $\Sigma\delta_1^{2}$ como $\Sigma\delta_2^{2}$ pueden sustituirse por el efectivo N en cuanto son segundos momentos sigmáticos multiplicados por dicho efectivo, y los segundos momentos sigmáticos valen 1.

Consiguientemente, puede establecerse finalmente que:

$$\varepsilon_{12\sigma}^2 = 1 - \alpha_{12}^2$$

O sea, que el cuadrado del error sigmático de estimación de la correlación rectilínea de primer orden es el complemento aritmético del cuadrado del coeficiente de correlación.

Relación entre el error de estimación y el índice de correlación.—Esta relación la hemos establecido en forma de definición desde un principio. El error de estimación es el complemento aritmético del cuadrado del índice de correlación rectilíneo, pues si:

$$r_{12}^2 = 1 - \varepsilon_{12\sigma}^2$$

también es cierto que:

$$\varepsilon_{12\sigma}^2 = 1 - r_{12}^2$$

Relación entre el índice y el coeficiente de correlación en una correlación rectilínea de primer orden.—Puesto que:

$$\varepsilon_{12\sigma}^2 = 1 - r_{12}^2$$

Y también:

$$\varepsilon_{12\sigma}^2 = 1 - \alpha_{12}^2$$

Puede establecerse una igualdad entre los segundos miembros de estas igualdades, cuyos primeros miembros son iguales. Así, tendremos:

$$1 - r_{12}^2 = 1 - \alpha_{12}^2$$

Siendo como son iguales las restas (es lo que afirma la igualdad) y siendo como son iguales los minuendos (1), tendrán que ser iguales los sustraendos:

$$r_{12}^2 = \alpha_{12}^2$$

Y, consiguientemente, sus raíces cuadradas:

$$r_{12} = \alpha_{12}$$

Conviene recordar las condiciones de esta identidad:

Indice y coeficiente de correlación son iguales siempre que y sólo si:

1.—La correlación es rectilínea y de primer orden.
2.—El coeficiente de correlación procede de la ecuación de estimación dada en desviaciones sigmáticas (de ahí que aparezca alfa minúscula y no A mayúscula como cuando la ecuación de estimación está dada en desviaciones respecto a la media aritmética, o a minúscula como cuando la ecuación de estimación está dada en términos de los datos originarios).

Relación entre los coeficientes de correlación, los errores de estimación y los índices de correlación de la primer variable en la segunda y de la segunda en la primera.—Las ecuaciones de estimación de la primera variable en la segunda y de la segunda en la primera, dadas en desviaciones sigmáticas son, respectivamente:

$$\delta_1 = \alpha_{12}\,\delta_2 \quad \text{y} \quad \delta_2 = \alpha_{21}\,\delta_1$$

Si obtenemos los valores de los coeficientes de correlación (alfas) a partir de cada una de estas ecuaciones tomadas en forma independiente, obtendremos, respectivamente:

$$\alpha_{12} = \frac{\Sigma\delta_1\,\delta_2}{N} \quad \text{y} \quad \alpha_{21} = \frac{\Sigma\delta_2\,\delta_1}{N}$$

Por simple inspección (puesto que el orden de los factores de la expresión bajo el sumador no altera los productos correspondientes y, por consiguiente, no altera tampoco su suma), puede verse que los dos coeficientes de correlación son iguales:

$$\alpha_{12} = \alpha_{21}$$

Puesto que, en la correlación rectilínea de primer orden los coeficientes de correlación y los respectivos índices de correlación son iguales (o sea, que $\alpha_{12} = r_{12}$; $\alpha_{21} = r_{21}$ si los coeficientes han sido obtenidos de la ecuación sigmática de estimación), podremos establecer que:

$$r_{12} = r_{21}$$

O sea, que los índices de correlación obtenidos mediante la estimación de la primer variable a partir de la segunda y de la segunda a partir de la primera son idénticos.

Finalmente, en virtud de las relaciones que ligan al índice de correlación y al error de estimación sigmático, también puede establecerse que:

$$\varepsilon_{12\sigma} = \varepsilon_{21\sigma}$$

O sea, que el error que se obtiene de estimar la primera variable a partir de la segunda y el que se obtiene al estimar la segunda variable a partir de la primera son idénticos.

En general, se puede establecer que: en la correlación rectilínea de primer orden, los coeficientes *sigmáticos* de correlación, los errores de estimación *sigmática* y el índice de correlación (que siempre es una medida *sigmática)* son idénticos, sea que se obtengan de estimar la primer variable a partir de la segunda o la segunda a partir de la primera.

Relación entre los coeficientes de correlación no sigmáticos de la correlación rectilínea de primer orden.—Si partimos de las ecuaciones de estimación dadas en términos de desviaciones con respecto a las correspondientes medias aritméticas de cada serie univariada pero expresadas en unidades ordinarias y no en unidades sigmáticas, las ecuaciones correspondientes serán:

$$d_1 = A_{12}\, d_2 \quad \text{y} \quad d_2 = A_{21} d_1$$

A partir de estas ecuaciones, los valores de los coeficientes NO sigmáticos de correlación son:

$$A_{12} = \frac{\Sigma d_1\, d_2}{\Sigma d_2^{\,2}} = \frac{\Sigma d_1\, d_2}{N\sigma_2^{\,2}} \quad \text{y} \quad A_{21} = \frac{\Sigma d_1\, d_2}{N\sigma_1^{\,2}}$$

En este caso, puede observarse que, mientras los numeradores coinciden, los denominadores difieren.

Si multiplicamos entre sí los coeficientes NO sigmáticos de correlación, obtenemos:

$$A_{12}A_{21} = \frac{\Sigma d_1 d_2}{N\,\sigma_2^{\,2}}\ \frac{\Sigma d_1 d_2}{N\,\sigma_1^{\,2}} = \frac{(\Sigma d_1 d_2)^2}{N^2\,\sigma_1^{\,2}\sigma_2^{\,2}} = \left(\frac{\Sigma d_1 d_2}{N\sigma_1\sigma_2}\right)^2$$

En la última expresión se puede reconocer el cuadrado del índice de correlación r_{12}. O sea que:

$$A_{12}A_{21} = r_{12}^{\,2}$$

O bien que:

$$r_{12} = \sqrt{A_{12}A_{21}}$$

Esto es lo que permite afirmar que: el índice de correlación rectilínea de primer orden es la media geométrica de los coeficientes de correlación no sigmáticos.

Aplicación.—A partir de la ecuación de estimación en desviaciones sigmáticas, puede obtenerse el valor del índice de correlación rectilínea de primer orden (r_{12}) en la forma siguiente:

Si se aplica el operador sigma mayúscula a los dos miembros de la ecuación

$$\delta_1 = r_{12}\, \delta_2\ ,$$

se obtiene:

$$\Sigma\delta_1 = \Sigma r_{12}\,\delta_2$$

Pero, en el segundo miembro se tiene la sumatoria del producto (o más propiamente, "de los productos") de cada desviación sigmática (δ_2) por el índice de correlación (r_{12}), o sea, la sumatoria de una serie de variables por una constante. Y la sumatoria de una constante por una serie de variables es igual a la constante por la sumatoria de las variables (o, lo que es equivalente, la constante bajo el operador sigma o sumador puede salir de él sin que la expresión se afecte):

$$\Sigma\delta_1 = r_{12}\Sigma\delta_2$$

Si en esta expresión se despeja a r_{12}, se tendrá:

$$r_{12} = \frac{\Sigma\delta_1}{\Sigma\delta_2}$$

Pero como la suma de las desviaciones sigmáticas es nula, serán nulos tanto el numerador como el denominador y el valor de r_{12} quedará sin determinar (0/0 es signo de indeterminación).

Para deshacer la indeterminación, volveremos a tomar la expresión original

$$\delta_1 = r_{12}\,\delta_2$$

Si multiplicamos ambos miembros por δ_2:

$$\delta_1\,\delta_2 = r_{12}\,\delta_2^2$$

Aplicando el sumador a ambos miembros:

$$\Sigma\delta_1\,\delta_2 = r_{12}\,\Sigma\delta_2^2$$

$$\therefore\ r_{12} = \frac{\Sigma\delta_1\,\delta_2}{\Sigma\delta_2^2}$$

Pero como el denominador (la suma de los cuadrados de las desviaciones sigmáticas) es igual al segundo momento sigmático, o sea al efectivo de la distribución, tendremos:

$$r_{12} = \frac{\Sigma \delta_1 \, \delta_2}{N}$$

FALTA DE EQUIVALENCIA DE LOS INDICES DE CORRELACION MULTIPLE DE UNA DISTRIBUCION

Cuando la correlación es rectilínea, el índice de correlación no varía, sea que se calcule tomando la primera variable como dependiente y la segunda como independiente, o sea, que se calcule tomando la primera variable como independiente y la segunda como dependiente. O sea, que siempre es posible establecer que:

$$r_{12} = r_{21}$$

En cambio, no puede decirse lo mismo —aún en el caso de la correlación rectilínea— con respecto a los índices de correlación múltiple de una misma distribución. Es esto lo que intentaremos mostrar.

Podemos registrar como índices de correlación posibles, en el caso de una distribución trivariada: r_{123}, r_{132}, r_{213}, r_{231}, r_{312}, r_{321}, De éstos algunos son iguales entre sí, en tanto que los demás difieren de ellos.

Los índices que hemos consignado pueden parearse como sigue:

$$r_{123} \text{ y } r_{132}$$

$$r_{213} \text{ y } r_{231}$$

$$r_{312} \text{ y } r_{321}$$

Cada par está constituido por índices en los que la variable dependiente es la misma (1 en el primer par, 2 en el segundo, 3 en el tercero). En cambio, en cada par los dos índices constitutivos del par difieren entre sí por el orden en que aparecen las variables independientes (2 y 3, 3 y 2 en el primer par; 1 y 3, 3 y 1 en el segundo, 1 y 2, 2 y 1 en el tercero).

Es fácil anticipar que las variaciones dentro de cada par serán menos importantes que las variaciones de par a par (el cambio en el orden en que se tomen las variables independientes tiene que afectar menos al resultado que el cambio que se haga de una variable al trasformarla de independiente en dependiente).

Con respecto a las variaciones dentro de cada par, tomaremos como ilustración el primero de dichos pares.

Partiremos de la fórmula

$$r_{123}^{2} = 1 - \frac{R_{3x3}}{R_{11(3x3)}}$$

que se justifica en otro sitio.

En la cual:

$$R_{3x3} = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix}$$

siendo el denominador de la fracción el cofactor del primer elemento del determinante:

$$\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}$$

Cuando tratamos no con r_{123} sino con r_{132} se producen ciertas trasformaciones en los determinantes correspondientes. Es decir, que la fórmula para r_{132} será:

$$r_{132}^{2} = 1 - \frac{R_{3x3}'}{R_{11}'}$$

En esta fórmula:

$$R_{3x3}' = \begin{vmatrix} 1 & r_{13} & r_{12} \\ r_{13} & 1 & r_{23} \\ r_{12} & r_{23} & 1 \end{vmatrix}$$

Y, el cofactor del primer elemento de este nuevo determinante será:

$$\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}$$

Basta desarrollar tanto R_{3x3} como R_{3x3}' para ver que ambos son iguales. Es todavía más inmediato comprobar que los cofactores correspondientes, de los primeros elementos de dichos determinantes de tercer orden, son iguales. O sea, que $r_{123} = r_{132}$. En forma más general, puede establecerse que:

El índice de correlación múltiple no se altera cuando se altera el orden de las variables independientes. O sea, simbólicamente, que el valor de r no se altera si se cambia el orden de los índices que subsiguen al primero (que es el que representa la variable dependiente y que, por ello, suele separarse con un punto de los subíndices restantes).

$$r_{1.23} = r_{1.32}$$

$$r_{2.13} = r_{2.31}$$

$$r_{3.12} = r_{3.21}$$

Establecida la equivalencia entre estos pares de índices, trataremos de mostrar la falta de equivalencia de cada pareja con las restantes; o sea, que tataremos de demostrar que:

$$r_{1.23} \neq r_{2.13} \neq r_{3.12}$$

Si, para fines comparativos, aplicamos a $r_{2.13}$ la misma fórmula que hemos aplicado al principio a $r_{1.23}$, tendremos:

$$r_{2.13}^{2} = 1 - \frac{R_{3x3}''}{R_{11}''}$$

El determinante del numerador será:

$$R_{3x_3}'' = \begin{vmatrix} 1 & r_{21} & r_{23} \\ r_{21} & 1 & r_{31} \\ r_{23} & r_{31} & 1 \end{vmatrix}$$

Y el cofactor del primer elemento de este determinante, o determinante del denominador, será:

$$\begin{vmatrix} 1 & r_{31} \\ r_{31} & 1 \end{vmatrix}$$

Si desarrollamos la fórmula de $r_{1.23}^2$ (sustituyendo previamente los determinantes de tercero y de segundo orden en el numerador del substraendo del segundo miembro), obtendremos:

$$r_{123}^2 = \frac{r_{12}^2 - 2\, r_{12}r_{13}r_{23} + r_{13}^2}{1 - r_{23}^2}$$

Si, en forma parecida, sustituimos los determinantes de tercero y de segundo orden recién encontrados en el numerador y denominador del substraendo del segundo miembro, en la fórmula para r_{231}^2, tendremos:

$$r_{231}^2 = \frac{r_{23}^2 - 2\, r_{23}r_{21}r_{31} + r_{21}^2}{1 - r_{31}^2}$$

Para lograr la máxima semejanza, pueden hacerse las trasformaciones siguientes:

$$r_{231}^2 = \frac{r_{12}^2 - 2\, r_{12}r_{13}r_{23} + r_{23}^2}{1 - r_{13}^2}$$

A pesar de estos cambios de lugar de subíndices de coeficientes de correlación (índices de correlación simple), de factores y de sumandos, sigue estando de manifiesto que

$$r_{123}^2 \neq r_{231}^2$$

pues en los términos finales de numerador y denominados figura r_{13} en un caso y r_{23} en el otro.

Y, en forma parecida podría demostrarse que cada uno de ellos es distinto de

$$r_{312}^{2}$$

O sea que, como afirmamos antes:

$$r_{1.23} \neq {}_{2.31} \neq r_{3.12}$$

Es decir, que puede afirmarse que:

El índice de correlación múltiple *se altera* si se toma una de las variables que previamente se había tomado como independiente, como variable dependiente y la variable tomada previamente como dependiente se considera como independiente. En lenguaje simbólico, el valor de r se altera si uno de los subíndices cambia de lugar con otro que se encuentre de lado contrario al suyo con respecto al punto de separación de dichos subíndices.

RELACIONES ENTRE LAS CONSTANTES ESTADISTICAS DE LA CORRELACION DE SEGUNDO ORDEN Y ENTRE ESTAS Y LAS CONSTANTES DE LA CORRELACION DE PRIMER ORDEN

Fórmulas de los coeficientes de correlación.—A partir de la ecuación sigmática de estimación de la correlación de segundo orden:

$$\delta_1 = \alpha_{12.3}\, \delta_2 + \alpha_{13.2}\, \delta_3$$

Pueden obtenerse las siguientes ecuaciones de interpolación.

$$\Sigma\delta_1\, \delta_2 = \alpha_{12.3}\Sigma\, \delta_2^{\,2} + \alpha_{13.2}\Sigma\delta_2\, \delta_3$$

$$\Sigma\delta_1\, \delta_3 = \alpha_{12.3}\, \Sigma\delta_2\delta_3 + \alpha_{13.2}\, \Sigma\delta_3^{\,2}$$

Si cada una de estas ecuaciones se divide entre N, efectivo de la distribución, la primera suma de productos de desviaciones sigmáticas se convierte en el índice de correlación correspondiente, y lo mismo ocurre en todos los restantes casos análogos. En cuanto a las sumas de los cuadrados de las desviaciones sigmáticas, al dividirse entre N se convierten en los segundos momentos sigmático que, como se sabe, valen 1. Con esto, las ecuaciones de interpolación se convierten en:

$$r_{12} = \alpha_{12.3} + \alpha_{13.2}\, r_{23}$$

$$r_{13} = \alpha_{12.3}\, r_{23} + \alpha_{13.2}$$

A partir de estas ecuaciones, puede formarse, como eliminante del sistema:

$$\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix} = 1 - r_{23}^{\,2}$$

Para obtener el valor de $\alpha_{12.3}$, se necesitará formar una fracción, cuyo denominador sea el eliminante así encontrado, y cuyo numerador sea un determinante análogo en el que se hayan sustituido los elementos de la primera columna por los términos independientes:

$$\begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix} = r_{12} - r_{13}r_{23}$$

El numerador de la fracción que permitirá obtener $r_{13.2}$ (cuyo denominador es también el eliminante) se obtiene mediante la sustitución de los elementos de la segunda columna por los términos independientes:

$$\begin{vmatrix} 1 & r_{12} \\ r_{23} & r_{13} \end{vmatrix} = r_{13} - r_{12}\, r_{23}$$

En estas condiciones, los valores de los coeficientes de correlación de la ecuación sigmática resultan ser:

$$\alpha_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^{2}}$$

$$\alpha_{13.2} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^{2}}$$

El error sigmático de estimación de la correlación de orden segundo en términos de los coeficientes sigmáticos de correlación. —La fórmula que nos servirá de punto de partida es:

$$\varepsilon_{123\sigma}^{2} = \frac{\Sigma(\delta_1 - \hat{\delta}_{123})^2}{N}$$

Como:

$$\hat{\delta}_{123} = \alpha_{12.3}\,\delta_2 + \alpha_{13.2}\,\delta_3$$

$$\Sigma(\delta_1 - \hat{\delta}_{123})^2 = \Sigma(\delta_1 - \alpha_{12.3}\,\delta_2 - \alpha_{13.2}\,\delta_3)^2 =$$

Cuyo desarrollo produce: tres sumandos suma de cuadrados de desviaciones y tres sumandos duplos de productos de los términos del trinomio tomados de dos en dos (apareciendo en estos últimos, además del coeficiente 2, uno o los dos coeficientes de correlación y sumas de productos de desviación sigmáticas):

$$=\Sigma\delta_1^2 + \alpha_{12.3}^2\Sigma\delta_2^2 + \alpha_{13.2}^2\Sigma\delta_3^2 +$$
$$+ 2(-\alpha_{12.3}\Sigma\delta_1\,\delta_2 - \alpha_{13.2}\Sigma\,\delta_1\,\delta_3 + \alpha_{12.3}\,\alpha_{13.2}\,\Sigma\delta_2\,\delta_3)$$

Si se recuerda que la suma de los cuadrados de las desviaciones sigmáticas es igual al efectivo de la distribución, y que las sumas de productos de desviaciones sigmáticas es igual al correspondiente índice de correlación rectilínea multiplicado por el efectivo, y si, finalmente, puesto que en todos los términos aparece como factor dicho efectivo, se saca éste como factor común, se obtiene:

$$N[(1 + \alpha_{12.3}^2 + \alpha_{13.2}^2) + 2(\alpha_{12.3}\,\alpha_{13.2}r_{23} - \alpha_{12.3}r_{12} - \alpha_{13.2}r_{13})]$$

Para obtener el equivalente del cuadrado del error sigmático de estimación, basta con dividir entre el efectivo, o sea, con suprimir el factor N en el desarrollo anterior.

$$\varepsilon_{123\sigma}^2 = 1 + \alpha_{12.3}^2 + \alpha_{13.2}^2$$
$$+ 2(\alpha_{12.3}\,\alpha_{13.2}\,r_{23} - \alpha_{12.3}\,r_{12} - \alpha_{13.2}\,r_{13})$$

En forma parecida y por las relaciones ya establecidas entre el error sigmático de estimación y el índice de correlación, puede establecerse el equivalente del cuadrado de este último con sólo suprimir en el segundo miembro el primer término.

$$r_{123} = \alpha_{12.3}^2 + \alpha_{13.2}^2 + 2\,(\alpha_{12.3}\,\alpha_{13.2}\,r_{23} - \alpha_{12.3}\,r_{12} - \alpha_{13.2}\,r_{13}).$$

Equivalencia de los coeficientes de la ecuación de regresión en términos de un determinante de orden superior.—¿Puede construirse, a partir de los tres determinantes siguientes (mismos que hemos visto aparecer en el cálculo de los coeficientes de la ecuación de regresión), un determinante de tercer orden (una matriz de

3 x 3 elementos) que contenga a esos tres determinantes de 2º orden (matrices de 2 x 2 elementos)?

Los determinantes que nos sirven como punto de partida son:

$$\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix} \qquad \begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix} \qquad \begin{vmatrix} 1 & r_{12} \\ r_{23} & r_{13} \end{vmatrix}$$

De estos determinantes el primero puede dejarse intacto. En cuanto al segundo, si se toman las columnas por renglones y los renglones por columnas (1ª columna convertida en primer renglón, 2ª columna convertida en segundo renglón) el determinante no se altera. Esto también puede expresarse diciendo que ha habido un intercambio de elementos de la diagonal secundaria mientras permanecían intactos los de la principal. De este modo, el segundo determinante es equivalente a:

$$\begin{vmatrix} r_{12} & r_{13} \\ r_{23} & 1 \end{vmatrix}$$

En cuanto al tercer determinante y mediante un cambio semejante, puede trasformarse en:

$$\begin{vmatrix} 1 & r_{23} \\ r_{12} & r_{13} \end{vmatrix}$$

Si ahora cambiamos el primer renglón por el segundo y el segundo por el primero, se obtendrá un determinante igual al anterior en valor absoluto, pero de signo contrario, o sea, que el anterior será igual a:

$$\text{MENOS} \begin{vmatrix} r_{12} & r_{13} \\ 1 & r_{23} \end{vmatrix}$$

De acuerdo con esto, nuestros nuevos elementos de trabajo son los tres determinantes siguientes de segundo orden:

$$\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix} \qquad \begin{vmatrix} r_{12} & r_{13} \\ r_{23} & 1 \end{vmatrix} \qquad \begin{vmatrix} r_{12} & r_{13} \\ 1 & r_{23} \end{vmatrix}$$

Las transformaciones que hicimos en los dos últimos determinantes tenían como propósito poner de manifiesto los elementos comunes de estos tres determinantes.

El renglón 1 r_{23} es común al primero y al tercer determinante (en uno es primer renglón y en el otro segundo). Por ello puede presumirse que una conexión adecuada de los dos determinantes pudiera ser la que colocara en primer término al tercer determinante y lo superpusiera mediante la coincidencia de sus renglones comunes al primer determinante:

$$\begin{vmatrix} r_{12} & r_{13} \\ 1 & r_{23} \\ & \end{vmatrix} \qquad \begin{vmatrix} & \\ 1 & r_{23} \\ r_{23} & 1 \end{vmatrix} \text{ o sea } \begin{vmatrix} r_{12} & r_{13} \\ 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}$$

En esta sucesión de renglones, puede observarse que, si se omite el central, se reproduce el segundo determinante de segundo orden. Esto nos orienta en el sentido de que la serie de renglones que hemos obtenido puede constituir las dos columnas de un determinante de tercer orden, del que habría que escribir una tercera columna. Si volvemos a escribir esa serie de renglones en la forma siguiente:

$$\begin{vmatrix} r_{12} & r_{13} \\ r_{22} & r_{23} \\ r_{32} & r_{33} \end{vmatrix}$$

cosa que es posible porque $r_{23} = r_{32}$ y r_{22} y r_{33} son iguales a 1, observaremos un cierto patrón. En el primer renglón aparecen índices de la primera decena (*12, 13*); en el segundo renglón, índices de la segunda decena (*22, 23*); en el tercero, de la tercera (*32, 33*). Por otra parte, en el arreglo que tenemos a la vista, las unidades del subíndice de la primera columna son 2 (*12, 22, 32*) y las unidades del subíndice de la segundo columna son 3 (*13, 23, 33*). Puede constituirse, por lo mismo una columna cuyas unidades sean unos (ya que las variables con las que trabajamos fueron x_1, x_2, x_3) y cuyas decenas sean las del renglón correspondiente. Esta columna conviene que preceda a las anteriores y estará constituida, por tanto, por r_{11}, r_{21}, r_{31}. Las tres columnas integrantes del determinante de tercer orden serán:

$$\begin{vmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix}$$

O bien, en cuanto los índices de guarismo reduplicado han cumplido su misión, con vistas a la utilización práctica:

$$\begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$$

De acuerdo con la forma en que fue obtenido este determinante, no puede extrañar que si tomamos los cofactores (o determinantes de segundo orden) correspondientes a los elementos de su primera columna, obtengamos los determinantes que sirvieron para el cálculo de las alfas. Así:

$$C_{11} = (-1)^{1+1} \begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}$$

O sea, que sin cambio de signo, el que era eliminante del sistema no es en relación con este determinante de tercer orden sino el cofactor del elemento de la primera columna y el primer renglón.

Para el cofactor del segundo elemento de la primera columna (obtenible mediante la eliminación de la primera columna y el segundo renglón) se tiene:

$$C_{12} = (-1)^{1+2} \begin{vmatrix} r_{12} & r_{13} \\ r_{23} & 1 \end{vmatrix} = -(r_{12} - r_{13}r_{23})$$

O sea, que el determinante de segundo orden que sirvió de numerador para encontrar el valor de la primera de las alfas es igual al cofactor del elemento de la primera columna y del segundo renglón en el determinante de tercer orden, pero *con signo contrario.*

En consecuencia, el valor de $\alpha_{12.3}$ puede expresarse como:

$$\alpha_{12.3} = \frac{\Delta_1}{\Delta} = \frac{-C_{12}}{C_{11}}$$

Por otra parte, si tomamos el cofactor del elemento 13 (columna 1 renglón 3) tendremos:

$$C_{13} = (-1)^{1+3} \begin{vmatrix} r_{12} & r_{13} \\ 1 & r_{23} \end{vmatrix} = (r_{12}r_{23} - r_{13}) = -(r_{13} - r_{12}r_{23})$$

O sea, asimismo, en este caso, el determinante de segundo orden que sirvió de numerador para el valor de la segunda de las alfas es igual al cofactor del elemento de la primera columna y del tercer renglón en el determinante de tercer orden, también *con signo contrario.*

De ahí que el valor de $\alpha_{13.2}$ pueda escribirse:

$$\alpha_{13.2} = \frac{\Delta_2}{\Delta} = \frac{-C_{13}}{C_{11}}$$

Por extensión (estas extensiones aparentemente inútiles suelen ser útiles a veces para las deducciones y con propósito generalizantes) puede establecerse que $\alpha_{11.23} = C_{11}/C_{11}$ y, consiguientemente, igual a 1.

Según esto, la ecuación de estimación para la correlación rectilínea de segundo orden puede escribirse:

$$\hat{\delta}_{1.23} = -\frac{C_{12}}{C_{11}}\delta_2 - \frac{C_{13}}{C_{11}}\delta_3$$

A partir de esta ecuación de estimación, trataremos de encontrar las fórmulas correspondientes al índice de correlación múltiple y al error de estimación.

Error de estimación de la correlación múltiple o de orden superior al primero.—El error sigmático de estimación lo hemos definido en casos anteriores como la media cuadrática de las diferencias entre las desviaciones sigmáticas observadas y las desviaciones sigmáticas estimadas. O sea, que puede establecerse que:

$$(\varepsilon_{1.23\sigma})^2 = \frac{\Sigma(\delta_1 - \hat{\delta}_1)^2}{N}$$

En el numerador del segundo miembro pueden hacerse las siguientes sustituciones:

$$\Sigma(\delta_1 - \hat{\delta}_1)^2 = \Sigma(\delta_1 - \alpha_{12.3}\delta_2 - \alpha_{13.2}\delta_3)^2$$

La sustitución de los coeficientes de correlación (las alfas) produce:

$$\Sigma(\delta_1 + \frac{C_{12}}{C_{11}}\delta_2 + \frac{C_{13}}{C_{11}}\delta_3)^2$$

Expresión que también puede escribirse:

$$\frac{\Sigma(C_{11}\delta_1 + C_{12}\delta_2 + C_{13}\delta_3)^2}{C_{11}^2} \dots . (1)$$

Para simplificar, nos ocuparemos sólo del cuadrado del numerador (más propiamente, de la suma de cuadrados que figura en el numerador). Su desarrollo produce una suma de cuadrados y dobles productos de los términos del paréntesis y es el siguiente:

$$\Sigma(C_{11}^2\delta_1^2 + 2\,C_{11}C_{12}\delta_1\delta_2 + 2\,C_{11}C_{13}\,\delta_1\delta_3 + C_{12}^2\,\delta_2^2 + 2\,C_{12}C_{13}\delta_2\delta_3 + C_{13}^2\delta_2^2)$$

La sumatoria puede convertirse en una suma de sumatorias:

$$C_{11}^2\Sigma\delta_1^2 + 2\,C_{11}C_{12}\Sigma\delta_1\delta_2 + 2\,C_{11}C_{13}\Sigma\delta_1\delta_3 + C_{12}^2\Sigma\delta_2^2 + 2\,C_{12}C_{13}\Sigma\delta_2\,\delta_3 + C_{13}^2\Sigma\delta_3^2$$

Pero la suma de los cuadrados de las deltas o desviaciones sigmáticas dividida entre el efectivo de la distribución (N) que figura en el denominador de la expresión inicial del error sigmático de estimación es el segundo momento sigmático que vale 1 en cada caso, con lo que:

$$\frac{C_{11}^2 \Sigma\delta_1^2}{N} = C_{11}^2$$

Expresiones parecidas pueden consignarse para los equivalentes de los términos de este tipo de desarrollo anterior, que así se convierte en:

$$C_{11}^2 + C_{12}^2 + C_{13}^2 + \\ + \frac{2\,C_{11} + C_{12}\,\Sigma\delta_1\delta_2}{N} + \frac{2\,C_{12}\,C_{13}\,\Sigma\delta_1\delta_3}{N} + \\ \frac{2\,C_{12}\,C_{13}\,\Sigma\delta_2\delta_3}{N}$$

En este desarrollo, en los términos fraccionarios figuran sumas de productos de desviaciones sigmáticas que, divididas entre el efectivo de la distribución producen el momento producto de las correspondientes series bivariadas en unidades sigmáticas, o sea, los correspondientes índices de correlación de primer orden. Según esto, la expresión previa pasa a ser igual a:

$$C_{11}^2 + 2\,C_{11}C_{12}r_{12} + 2\,C_{11}C_{13}r_{13} + C_{12}^2 + C_{12}C_{13}r_{23} + C_{13}^2$$

Esta expresión puede escribirse:

$$C_{11}(C_{11} + C_{12}r_{12} + C_{13}r_{13}) + C_{12}(C_{12} + C_{11}r_{12} + C_{13}r_{23}) + \\ C_{13}\,(C_{13} + C_{11}r_{13} + C_{12}r_{23})$$

Toda esta expresión es el numerador de una previa (1) que tenía como denominador a C_{11}^2. Esto quiere decir que esa otra expresión se convertirá en:

$$\frac{(C_{11} + C_{12}r_{12} + C_{13}r_{13})}{C_{11}} + \frac{C_{12}}{C_{11}^{2}} (C_{12} + C_{11}r_{12} + C_{13}r_{23})$$

$$+ \frac{C_{13}}{C_{11}^{2}} (C_{13} + C_{11}r_{13} + C_{12}r_{23})$$

De estos términos, el primero tiene como numerador un determinante, cuyos elementos columnares son: 1, r_{12} y r_{13} o sea, los factores por los que aparecen multiplicados C_{11} C_{12} y C_{13} que son los cofactores correspondientes. O sea, que la primera columna de dicho determinante será:

$$\begin{matrix} 1 \\ r_{12} \\ r_{13} \end{matrix}$$

Con base en esto y en los valores de C_{11}, C_{12} y C_{13} que ya conocemos, el determinante del numerador del primer término es:

$$\begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix} = R$$

Según esto, todo el primer término del desarrollo que venimos siguiendo puede escribirse R/C_{11}.

En cuanto a los dos términos restantes, estos se anulan. En efecto, dentro de los paréntesis se tienen sumas de productos. Estos productos, a su vez, están formados por los elementos del determinante que acabamos de escribir (1, r_{12}, r_{13}) multiplicados por los cofactores de los elementos correspondientes a una columna que no es la suya (C_{12}, C_{11}, C_{13}). Y esto ocurre tanto en el caso del segundo término del desarrollo (que es el ejemplificado) como en el del tercer término (ya que 1, r_{13}, r_{23} aparecen multiplicados por C_{13}, C_{11}, C_{21}). Para comprobar esta anulación, pueden realizarse las operaciones indicadas dentro de cada paréntesis en estos dos últimos términos.

De este modo, todo el desarrollo conduce a la expresión:

$$(\varepsilon_{1.23\sigma})^2 = \frac{R}{C_{11}}$$

Pero, el error sigmático de estimación es igual a la raíz cuadrada del complemento aritmético del cuadrado del índice de correlación, según hemos visto en el caso anterior, o sea:

$$(\varepsilon_{1.23\sigma})^2 = 1 - r_{123}^2$$

Como dos cantidades iguales a una tercera son iguales entre sí:

$$1 - r_{123} = \frac{R}{C_{11}}$$

Despejando a r_{123}^2, se tiene:

$$r_{123}^2 = 1 - \frac{R}{C_{11}}$$

O sea, finalmente que:

$$r_{123} = (1 - \frac{R}{C_{11}})^{\frac{1}{2}}$$

Expresión en la que debe recordarse que:

$$R = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix}$$

C_{11}, por su parte es, en dicho determinante, el cofactor del elemento de la primera columna y el primer renglón.

Rasgo característico de la correlación múltiple (de orden superior al primero).—Como puede verse por todo lo anterior, a diferencia de lo que ocurre en la correlación de primer orden en la que los valores estadísticos por calcular eran un índice de correlación (r_{12}) y un error de estimación (ε_{12}) que afectaba las estimaciones ($\hat{\delta}$) hechas a partir de la ecuación de estimación ($\hat{\delta}_{1.2} = r_{12}\,\delta_2$), en el caso de una correlación de orden superior al primero aparecen además de los elementos correlativos (índice de correlación múltiple o de orden superior al primero, error de estimación y estimaciones obtenidas a partir de una ecuación de estimación) un cuarto elemento constituido por los coeficientes de correlación (que hemos designado con alfas minúsculas) y que son los factores constantes por los que aparecen multiplicadas las desviaciones sigmáticas de las variables independientes en la ecuación de estimación. En el caso de la correlación de primer orden, este elemento no se presentaba con la evidencia con que se presenta en correlaciones de orden superior, porque existía un solo coeficiente de correlación en la ecuación de estimación (la constante que multiplica el valor de δ_2 cuando se quiere obtener δ_{12}) y porque, además, dicho coeficiente de correlación se confundía con el índice de correlación (r_{12}). Esto explica el que, en tratándose de la correlación de primer orden (pero sólo de ella) r_{12} pueda designarse indistintamente "índice de correlación" o "coeficiente de correlación", puesto que es ambas cosas.

En el caso de la correlación de orden superior al primero, existe un "índice de correlación múltiple" ($r_{123\ldots}$) y tantos "coeficientes de correlación múltiple" como variables independientes intervengan en la ecuación de estimación ($\alpha_{12.3}, \alpha_{13.2}$ en el caso de la correlación entre una variable dependiente y dos independientes).

Presentación sintética de los cálculos necesarios para la obtención de la correlación de orden superior al primero.—De acuerdo con todo lo anterior, puede decirse que, básicamente, el estudio de la correlación múltiple implica:

1º La determinación de los valores de las desviaciones de cada serie en relación con sus respectivas medias aritméticas, expresadas en unidades sigmáticas (o sea, el cálculo de las "desviaciones sigmáticas" de cada una de las series univariadas componentes).

2º Determinación de los coeficientes de correlación de la ecuación de estimación (o sean, las "alfas").

3º Obtención de los valores estimados de las desviaciones sigmáticas de la variable dependiente a partir de los valores observados de las desviaciones sigmáticas de las independientes. Estimación es ésta que se hace a partir de la ecuación de estimación en la que figuran los coeficientes de correlación recién calculados. Se trata, en el caso, del cálculo de las "deltas minúsculas con circunflejo".

4º Cálculo del error sigmático de estimación (que también puede considerarse como un coeficiente de indeterminación y que hemos representado por epsilon subíndice sigma).

5º Cálculo del índice de correlación de orden n a partir del cálculo previo del error sigmático de estimación.

En seguida trataremos de aplicar este procedimiento a un ejemplo de tipo pedagógico en el que se han tomado valores extremadamente simples para facilitar las operaciones y poner únicamente de relieve, en lo concreto, cómo funciona el procedimiento. Partiremos del cálculo de los índices de correlación simple, en cuanto los pasos previos son conocidos.

EJEMPLO DE CORRELACION MULTIPLE

	$\delta_1 \delta_2$	$\delta_1 \delta_3$	δ_2^2	$\delta_2 \delta_3$	δ_3^2
	0.0000	0.0000	0.0000	0.0000	0.0000
	1.2555	0.3726	2.4025	0.7130	0.2116
	0.0000	3.6075	0.0000	0.0000	3.4225
	−0.2541	0.1518	0.5929	−0.3542	0.2116
	−1.7670	1.5846	2.4025	−2.1545	1.9321
	−0.2541	0.1518	0.5929	−0.3542	0.2116
Sumas	−1.0197	5.8683	5.9908	−2.1499	5.9894
entre N = 6	− .1699	0.9831	1.0000	−0.3583	1.0000
O índices de correlación de 1er. orden	r_{12}	r_{13}	r_{22}	r_{23}	r_{33}

A partir de este punto:

A.—Pueden sustituirse las sumas de estas columnas en las ecuaciones de interpolación, resolver el sistema y así obtener las alfas o

B.—Sustituir directamente los valores de los índices de correlación de primer orden en las fórmulas ya establecidas para las alfas.

Seguiremos el segundo procedimiento:

1º Calcularemos el eliminante del sistema (que figura en el denominador del valor de cada una de las alfas). El eliminante es:

$$\begin{vmatrix} r_{22} & r_{23} \\ r_{23} & r_{33} \end{vmatrix} = \begin{vmatrix} 1 & -0.3583 \\ -0.3583 & 1 \end{vmatrix} \begin{aligned} &= 1 - (-0.3583)^2 \\ &= 1 - 0.1284 \\ &= 0.8716. \end{aligned}$$

2º El determinante del numerador de $\alpha_{12.3}$ estará constituido por el eliminante en el que se habrá sustituido la primera columna por los términos independientes (o, en este caso, por r_{12} y r_{13})

$$\begin{vmatrix} r_{12} & r_{23} \\ r_{13} & r_{33} \end{vmatrix} = \begin{vmatrix} -0.1699 & -0.3583 \\ 0.9831 & 1 \end{vmatrix}$$

$$= -\,0.1699 - (0.9831x - 0.3583) =$$

$$= -\,0.1699 - (-0.3522) = 0.3522 - 0.1699 = 0.1823$$

3º El determinante del numerador de $\alpha_{13.2}$ estará constituido por el eliminante en el que se habrá sustituido la segunda columna por los términos independientes (que son, en el caso r_{12} y r_{13})

$$\begin{vmatrix} r_{22} & r_{12} \\ r_{23} & r_{13} \end{vmatrix} = \begin{vmatrix} 1 & -0.1699 \\ -0.3583 & 0.9831 \end{vmatrix}$$

$$= 0.9831 - (-0.1699x - 0.3583) =$$

$$= 0.9831 - (0.0609) = 0.9222$$

4º Consiguientemente, los valores de los coeficientes de correlación son:

$$\alpha_{12.3} = \frac{0.1823}{0.8716} = 0.2091 \qquad \alpha_{13.2} = \frac{0.9222}{0.8716} = 1.0580$$

5º Consiguientemente, la ecuación de estimación será:

$$\hat{\delta}_{12.3} = .2091\,\delta_2 + 1.0580\,\delta_3$$

CALCULO DE LAS VALORES ESTIMADOS DE LAS DESVIACIONES SIGMATICAS DE LA VARIABLE DEPENDIENTE

Datos obtenidos en elaboraciones previas		ELABORACION		
δ_2	δ_3	$.2091\,\delta_2$	$1.058\,\delta_3$	$\hat{\delta}_{1.23} = .2091\,\delta_2 + 1.058\,\delta_3$
0.00	0.00	0.000000	0.000000	0.00000
−1.55	−0.46	−0.324105	−0.486680	−0.81078
0.00	1.85	0.000000	1.957300	1.95730
−0.77	0.46	−0.161007	0.486680	0.32567
1.55	−1.39	0.324105	−1.470620	1.14651
0.77	−0.46	1.161002	−0.486680	−0.32567

CALCULO DEL ERROR SIGMATICO DE ESTIMACION DE LA CORRELACION MULTIPLE

DATOS originales y calculados mediante la elaboración previa				
Valores de la variable dependiente				
Observado	Estimado			
δ_1	$\hat{\delta}_{1.23}$	$\lvert\delta_1 - \hat{\delta}_{1.23}\rvert$	$\lvert\delta_1 - \hat{\delta}_{1.23}\rvert^2$	
0.00000	0.00000	0.00000	0.0000000000	
−0.81000	−0.81078	0.00078	0.0000006084	
1.95000	1.95730	0.00730	0.0000532900	
0.33000	0.32567	0.00433	0.0000187489	
−1.14000	1.14651	0.00651	0.0000423801	
−0.33000	−0.32567	0.00433	0.0000187489	
SUMA			0.0001337763	
entre N = 6			0.0000222960	
raíz cuadrada:			0.004721	error sigmático de estimación

Cálculo del índice de correlación de segundo orden.

Puesto que:

$$\varepsilon_{1.23\sigma} = 0.004721$$

y

$$r_{123} = (1 - \varepsilon_{1.23\sigma}{}^2)^{\frac{1}{2}}$$

El índice de correlación múltiple resulta ser:

$$\begin{aligned} r_{123} &= (1 - 0.004721^2)^{\frac{1}{2}} \\ &= (1 - 0.0000222960)^{\frac{1}{2}} = (0.99997704)^{\frac{1}{2}} \\ &= \sqrt{0.99997704} = .0000 \doteq 1.0000 \end{aligned}$$

Hemos seguido todo el procedimiento hasta llegar hasta este punto en que se revela que el índice de correlación múltiple es igual con la unidad (lo que indica que las tres series correlacionadas son equivalentes), porque nuestro interés radicaba en mostrar, en lo concreto, el modo de operar el procedimiento y no en obtener determinados resultados sustantivos. Sin embargo, desde temprano podía verse que, de haberse tomado una aproximación suficiente, las desviaciones estimadas de la variable dependiente hubieran resultado iguales a las desviaciones observadas; que, consiguientemente, el error de estimación sigmático hubiese resultado igual con cero y que r_{123} hubiera resultado igual a la unidad, mostrando con ello que la correlación entre las tres series es perfecta y que, por lo mismo, son equivalentes entre sí.

En seguida se encontrarán otro ejemplo pedagógico y diversas aplicaciones en los que no se da esta situación de correlación perfecta, pero que, procesalmente siguen los mismos lineamientos de cálculo seguidos en este ejemplo.

RELACIONES DE LOS INDICES DE CORRELACION PARCIAL CON LOS INDICES DE CORRELACION SIMPLE

Partiremos de la fórmula:

$$r_{i(e)d} = \frac{\Sigma \Delta_{d.e}\, \Delta_{e.i}}{[\Sigma(\Delta_{d.e})^2 \Sigma(\Delta_{i.e})^2]^{\frac{1}{2}}}$$

Fórmula en la que, para mayor generalidad, hemos sustituido en los subíndices por d la variable dependiente, por e la eliminada y por i la independiente.

De acuerdo con las fórmulas para las desviaciones de la estimada sigmática:

$$\Delta_{d.e} = \delta_d - r_{de}\, \delta_e$$

$$\Delta_{i.e} = \delta_i - r_{ie}\, \delta_e$$

El producto de estas deltas mayúsculas, será, consiguientemente, igual a:

$$(\delta_d - r_{de}\, \delta_e) \ (\delta_i - r_{ie}\, \delta_e) =$$

$$\delta_i\, \delta_d - r_{ie}\ \delta_d\, \delta_e - r_{de}\ \delta_i\, \delta_e + r_{ie}\, r_{de}\, \delta_e^{2}$$

Al aplicar el sumador a todos estos términos, se obtiene para el primero la suma de productos de desviaciones sigmáticas de las variables i y d, o sea su índice de correlación (r_{ie}). En el segundo mino en donde ya aparece r_{ie} la suma de los productos de las desviaciones sigmáticas de las variables d y e produce r_{de} y el término se convierte en $r_{ie} r_{de}$. En forma parecida, el tercer término pasa a ser $r_{de} r_{ie}$ y el último se reduce a $r_{ie} r_{de}$ puesto que la suma de los cuadrados de las desviaciones sigmáticas de la variable e es el segundo momento sigmático que, como sabemos, se reduce a la unidad. (En todo lo anterior hemos supuesto todos los términos divididos por N, lo cual puede justificarse fácilmente sin mayores desarrollos). En estas condiciones, el producto antes mencionado se convierte en:

$$r_{id} - r_{ie}r_{de} - r_{de}r_{ie} + r_{ie}r_{de} = r_{id} - r_{ie}r_{de}$$

Que es el equivalente del numerador en la fórmula para $r_{i(e)d}$. Para el denominador tendremos:

$$\Sigma(\Delta_{i.e})^2 = \Sigma(\delta_i - r_{ie}\,\delta_e)^2$$

Y una fórmula parecida para el otro factor (que no repetimos por economía).

El desarrollo del cuadrado del binomio encerrado en el paréntesis produce:

$$\Sigma(\delta_i^2 - 2\,r_{ie}\,\delta_i\,\delta_e + r_{ie}^2\,\delta_e^2)$$

Al aplicar el operador sigma a todos y cada uno de los términos de dentro del paréntesis (y dividir entre N, como se hizo en el numerador) se obtiene: como primer término la suma de los cuadrados de las desviaciones sigmáticas, o sea el segundo momento sigmático que se reduce a 1; como segundo miembro el doble producto de r_{ie} por la suma del producto de las desviaciones sigmáticas de las variables i y e, o sea, el doble producto de r_{ie} por r_{ie} (igual a $-2r_{ie}^2$), y como tercer término r_{ie}^2 puesto que la suma de los cuadrados de las desviaciones de la variable e o segundo momento sigmático de e se reduce a la unidad. Es decir, que la expresión anterior se convierte en:

$$1 - 2\,r_{ie}^2 + r_{ie}^2 = 1 - r_{ie}^2$$

Si tanto el equivalente previamente encontrado para el numerador, como este que recién acabamos de encontrar para el denominador se sustituyen en la expresión que nos sirvió de punto de partida, obtendremos:

$$r_{i(e)d} = \frac{r_{id} - r_{ie}r_{de}}{(1 - r_{de}^2)^{\frac{1}{2}}(1 - r_{ie}^2)^{\frac{1}{2}}}$$

Fórmula general que muestra la equivalencia entre los índices de correlación parcial de una correlación entre tres variables y los correspondientes índices de correlación simple.

Puede recordarse también que $1 - r_{ie}^2$ es el error sigmático $\varepsilon_{ie\sigma}{}^3$ (al cuadrado). O sea, que la fórmula anterior también puede escribirse:

$$r_{i(e)d} = \frac{r_{id} - r_{ie}r_{de}}{\varepsilon_{de\sigma}\varepsilon_{ie\sigma}}$$

Estas fórmulas permiten tener una idea más clara de lo que representa el índice de correlación parcial.

En la primera fórmula se establece dicho índice como la relación que existe entre la desviación (—) de la correlación simple entre variable independiente y variable dependiente y el producto de los índices de correlación entre cada una de dichas variables y la variable eliminada, y el producto de las raíces cuadradas (exponente $\frac{1}{2}$) de las diferencias entre los cuadrados de los índices de correlación simple y la unidad.

La segunda de dichas fórmulas muestra que se trata de la relación entre la diferencia del coeficiente de correlación de las variables independiente y dependiente y el producto de los de cada una de ellas con respecto a la variable eliminada, y el producto de los errores sigmáticos de estimación de cada una de esas dos variables (independiente y dependiente) cuando se calculan a partir de la variable eliminada.

RELACIONES ENTRE LOS COEFICIENTES SIGMATICOS DE CORRELACION MULTIPLE Y LOS INDICES DE CORRELACION PARCIAL

Conforme hemos podido mostrar anteriormente, los coeficientes sigmáticos de la correlación múltiple $\alpha_{12.3}$ y $\alpha_{13.2}$ pueden expresarse como sigue (en términos de los coeficientes-índices de correlación rectilínea de primer orden):

$$\alpha_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} = \frac{r_{12} - r_{13}r_{23}}{\varepsilon_{23\sigma}^2}$$

$$\alpha_{13.2} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} = \frac{r_{13} - r_{12}r_{23}}{\varepsilon_{23\sigma}^2}$$

En forma independiente, hemos podido llegar a establecer que:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\varepsilon_{13\sigma}\,\varepsilon_{23\sigma}}$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\varepsilon_{12\sigma}\,\varepsilon_{23\sigma}}$$

Si multiplicamos $\alpha_{12.3}$ por el cociente $\varepsilon_{23\sigma}/\varepsilon_{13\sigma}$ obtenemos:

$$\alpha_{12.3}\frac{\varepsilon_{23\sigma}}{\varepsilon_{13\sigma}} = \frac{r_{12} - r_{13}r_{23}}{\varepsilon_{23\sigma}^2}\,\frac{\varepsilon_{23\sigma}}{\varepsilon_{13\sigma}} = \frac{r_{12} - r_{13}r_{23}}{\varepsilon_{23\sigma}\,\varepsilon_{13\sigma}} = r_{12.3}$$

De acuerdo con esto:

$$\alpha_{12.3}\frac{\varepsilon_{23\sigma}}{\varepsilon_{13\sigma}} = r_{12.3}$$

Y, análogamente:

$$\alpha_{13.2}\frac{\varepsilon_{23\sigma}}{\varepsilon_{12\sigma}} = r_{13.2}$$

Recíprocamente:

$$r_{12.3} \frac{\varepsilon_{13\sigma}}{\varepsilon_{23\sigma}} = \alpha_{12.3}$$

El error de estimación parcial como cociente de los errores de estimación múltiple y simple.—De acuerdo con una de las fórmulas encontradas para el cuadrado del error sigmático de estimación de una correlación entre tres variables (una dependiente y dos independientes), se tiene:

$$\varepsilon_{123\sigma}^{2} = \frac{R_{3x3}}{C_{11(3x3)}}$$

O sea, que el error de estimación, elevado al cuadrado es igual al cociente que resulta de dividir una matriz cuadrada de tres elementos (los índices de correlación bivariada resultantes de tomar las tres variables de dos en dos) entre el cofactor del elemento de la primera columna y del primer renglón de dicho determinante.

El determinante de tercer orden al que se refiere la fórmula es:

$$R_{3x3} = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix}$$

El cofactor que figura en el numerador es, naturalmente:

$$C_{11(3x3)} = \begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}$$

Si esa misma fórmula la aplicamos a la determinación del cuadrado del error sigmático de estimación de una correlación entre dos variables (una dependiente y una independiente) que, por ulterior conveniencia, serán x_1 y x_3 o sus equivalentes δ_1, δ_3 tendremos:

$$\varepsilon_{12\sigma}^{2} = \frac{R_{2x2}}{C_{11(2x2)}}$$

En este caso, el determinante de segundo orden al que se refiere la fórmula es:

$$R_{2x2} = \begin{vmatrix} 1 & r_{13} \\ r_{13} & 1 \end{vmatrix}$$

Y el cofactor del elemento de la primera columna y del primer renglón de este determinante es:

$$C_{11(2x2)} = 1$$

Si desarrollamos R_{3x3} obtenemos:

$$R_{3x3} = 1 - r_{23}^{2} - r_{12}\,(r_{12} - r_{13}r_{23}) + r_{13}\,(r_{12}r_{23} - r_{13}) =$$
$$= 1 - r_{12}^{2} - r_{13}^{2} - r_{23}^{2} + 2\,r_{12}r_{13}r_{23}$$

El desarrollo de $C_{11(3x3)}$ produce:

$$C_{11(3x3)} = 1 - r_{23}^{2}$$

O sea, que el equivalente del cuadrado del error sigmático de estimación será:

$$\varepsilon_{123\sigma}^{2} = \frac{1 - r_{12}^{2} - r_{13}^{2} - r_{23}^{2} + 2\,r_{12}r_{13}r_{23}}{1 - r_{23}^{2}}$$

Por otra parte, el desarollo de R_{2x2} produce:

$$R_{2x2} = 1 - r_{13}^{2}$$

Y como el valor de $C_{11(2x2)} = 1$, el equivalente del cuadrado del error sigmático de estimación de esta otra correlación será:

$$\varepsilon_{13\sigma}^{2} = 1 - r_{13}^{2}$$

Si dividimos un cuadrado entre el otro, obtendremos:

$$\frac{\varepsilon_{123\sigma}^2}{\varepsilon_{13\sigma}^2} = \frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2\, r_{12}r_{13}r_{23}}{(1 - r_{13}^2)\ (1 - r_{23}^2)}$$

El valor de este cociente queremos compararlo con el valor de $\varepsilon_{1(2)3\sigma}^2$. Para ello partiremos de:

$$r_{1(3)2}^2 = \frac{(r_{12} - r_{13}r_{23})^2}{(1 - r_{13}^2)\ (1 - r_{23}^2)}$$

Como el error sigmático de estimación elevado al cuadrado es el complemento aritmético del cuadrado del índice de correlación:

$$\varepsilon_{1(3)2\sigma}^2 = 1 - r_{1(3)2}^2 = 1 - \frac{(r_{12} - r_{13}r_{23})^2}{(1 - r_{13}^2)\ (1 - r_{23}^2)}$$

Si se ejecutan las operaciones indicadas, se obtiene:

$$\varepsilon_{1(3)2\sigma}^2 = \frac{1 - r_{12}^2 - r_{13}^2 \div r_{23}^2 + 2\, r_{12}r_{13}r_{23}}{(1 - r_{13}^2)\ (1 - r_{23}^2)}$$

Si se compara el segundo miembro de esta ecuación con el segundo miembro de la ecuación que da el cociente de los cuadrados de los errores sigmáticos, se puede comprobar que dichos segundos miembros son idénticos, por lo que se pueden igualar los primeros miembros:

$$\varepsilon_{1(3)2}^2 = \frac{\varepsilon_{123\sigma}^2}{\varepsilon_{13\sigma}^2}$$

Si se extraen las raíces cuadradas de estos cuadrados (que sólo nos sirvieron para evitar la multiplicación tipográficamente embarazosa de radicales en el desarrollo) se obtiene:

$$\varepsilon_{1(3)2\sigma} = \frac{\varepsilon_{123\sigma}}{\varepsilon_{13\sigma}}$$

O sea que: el error de estimación de la correlación parcial entre variables de las que una queda eliminada es igual al cociente que resulta de dividir el error de estimación sigmático de la correlación múltiple entre las tres variables y el error de estimación resultante de correlacionar la variable dependiente con la eliminada.

Una notación alternativa podría poner mayormente de relieve la estructura de este error sigmático de estimación:

$$\varepsilon_{1(3)2\sigma} = \frac{\varepsilon_{123}}{\varepsilon_{1}\#_{3}}$$

En la que el signo # (empleado frecuentemente en lingüística para indicar "juntura" o cero contrastante) indica cuál es la variable que se toma en cuenta.

Relación entre los índices de correlación simple, múltiple y parcial.—Las relaciones entre los cuadrados de los errores de estimación, pueden expresarse también en la forma siguiente:

$$1 - r_{1(3)2}^{2} = \frac{1 - r_{123}^{2}}{1 - r_{13}^{2}}$$

Al pasar el denominador del segundo miembro al primer miembro como factor se obtiene:

$$(1 - r_{1(3)2}^{2})\ (1 - r_{13}^{2}) = 1 - r_{123}^{2}$$

$$1 - r_{1(3)2}^{2} - r_{13}^{2} + r_{13}^{2}\, r_{1(3)2}^{2} = 1 - r_{123}^{2}$$

Si se resta 1 a ambos miembros y se cambia signo a los resultados:

$$r_{13}^{2} + r_{1(3)2}^{2} - r_{13}^{2} r_{1(3)2}^{2} = r_{123}^{2}$$

O bien:

$$r_{123}^2 = (r_{13}^2 + r_{1(3)2}^2) - r_{13}^2 r_{1(3)2}^2$$

O sea, que el cuadrado del índice de correlación múltiple es igual a la suma de los cuadrados del índice de correlación simple y del correspondiente índice de correlación parcial menos el producto de esos mismos cuadrados.

IV. PROCEDIMIENTOS DIRECTOS DE CALCULO DE LOS INDICES DE CORRELACION

CALCULO DIRECTO DEL INDICE DE CORRELACION RECTILINEA DE PRIMER ORDEN

Al despejar a r_{12} de la ecuación $\Sigma\, \delta_1\, \delta_2 = r_{12}\, \Sigma\, \delta_2^2$, se obtiene:

$$r_{12} = \frac{\Sigma\, \delta_1\, \delta_2}{\Sigma \delta_2^2} = \frac{\Sigma\, \delta_1\, \delta_2}{N}$$

De acuerdo con esta fórmula puede diseñarse en sus términos más simples el procedimiento de cálculo. Para ello, hay que considerar que, con el fin de obtener las desviaciones con respecto a la media aritmética en unidades sigmáticas (o sea las deltas minúsculas) se necesitan las medias correspondientes y las sigmas (o desviaciones cuadráticas medias).

Procedimiento de cálculo.—Para calcular el coeficiente de correlación rectilínea de primer orden:

1º Calcúlense las dos medias aritméticas de las dos series univariadas componentes Para ello, habrá que sumar los datos de cada serie y dividir la suma entre el número de datos.	$\bar{x}_1, \bar{x}_2$
2º Abranse dos columnas y en ellas, frente a cada uno de los datos de cada una de las series univariadas, consígnese la diferencia entre el dato y su media o sea la desviación con respecto a la media aritmética.	d_1, d_2

3º En otras dos columnas, consígnense los cuadrados de dichas desviaciones.	d_1^2, d_2^2
4º Sumense al pie los valores de estas dos últimas columnas,	Σd_1^2 Σd_2^2
Divídanse las sumas entre el número de datos, y	$\Sigma d_1^2/N$; $\Sigma d_2^2/N$
Extráigase la raíz cuadrada del cociente	$\sqrt{\Sigma d_1^2/N}$
para obtener las desviaciones cuadráticas (o sigmas minúsculas)	σ_1, σ_2
5º Divídase cada una de las desviaciones con respecto a la media	d_1, d_2
de cada serie entre la correspondiente desviación cuadrática media	d_1/σ_1; d_2/σ_2
a fin de obtener las desviaciones con respecto a la media en unidades de la desviación cuadrática media (o desviaciones sigmáticas)	δ_1, δ_2
6º Multiplíquese cada desviación sigmática de la primera serie	δ_1
por la desviación sigmática de la segunda serie,	δ_2
consígnense los valores así obtenidos en una columna	$\delta_1\,\delta_2$
y súmense al pie de la columna	$\Sigma\,\delta_1\,\delta_2$
8º Divídase la suma de los productos de las desviaciones sigmáticas de la primera y de la segunda series	$\Sigma\,\delta_1\,\delta_2$
entre el efectivo de la distribución (o número de datos)	$\Sigma\,\delta_1\,\delta_2/N$

para obtener finalmente el índice de correlación rectilínea de primer orden r_{12}

Ejemplo pedagógico.—En el siguiente ejemplo se ha tomado una serie cortísima y valores tan sencillos como ha sido posible para cada uno de los fenómenos que tratan de correlacionarse, con el fin de que sea más fácil seguir la mecánica procesal. Esto no implica, en forma alguna, el que en la práctica sea aconsejable tomar un corto número de observaciones para buscar la correlación entre dos o más series, o el que, en esa misma práctica, las observaciones se concreten en cifras tan simples. La justificación del ejemplo estriba, repetimos, en su finalidad pedagógica ilustrativa.

Cálculo del índice de correlación rectilínea de primer orden entre x_1 y x_2 a partir de las desviaciones sigmáticas.

	DATOS		ELABORACIONES						
	x_1	x_2	d_1	d_2	d_1^2	d_2^2	δ_1	δ_2	$\delta_1\,\delta_2$
	21	3	−28	−4	784	16	−1.44	−1.44	2.0736
	42	6	− 7	−1	49	1	−0.36	−0.36	0.1296
	49	7	0	0	0	0	0.00	0.00	0.0000
	70	10	21	3	441	9	1.08	1.08	1.1664
	77	11	28	4	784	16	1.44	1.44	2.0736
	35	5	−14	−2	196	4	0.72	0.72	0.5184
SUMAS	294	42			2254	46			
MEDIAS	49	7			375.67	7.67			o, prácticamente
RAÍCES					19.39	2.77			6

$$r_{12} \quad \frac{\Sigma\,\delta_1\,\delta_2}{N} = \frac{5.9616}{6} = \text{práct. } 1$$

A fin de que el lector pueda reconstruir más fácilmente el procedimiento seguido, haremos en el cuadro las siguientes identificaciones de resultados parciales:

$$\bar{x}_1 = 49 \qquad \sigma_1 = 19.23$$

$$\bar{x}_2 = 7 \qquad \sigma_2 = 2.77$$

El resultado final para r_{12} es ligeramente superior a la unidad, lo cual teóricamente no es posible ni explicable, ya que este índice de correlación varía de -1 a $+1$ pasando por O, pero sin pasar nunca de $+1$ ni descender por debajo de -1 (o sea, sin que pueda tener valores como $+1.1$, $+1.2$, -1.3, -1.4, -3, etc.). En la práctica, el hecho de que el valor exacto obtenido para r_{12} en el ejemplo sea ligeramente superior a 1 (1.001) se debe a las aproximaciones que se tomaron al operar. Con la creciente aproximación, en este caso en que hay correlación perfecta (puesto que $r_{12} = 1$), el denominador tenderá al numerador de la expresión para r_{12}.

Obtenido el valor para r_{12} (que así como en este caso fue 1, en otros puede ser un valor diferente de 1), es posible sustituirlo en la ecuación de estimación en desviaciones sigmáticas:

$$\delta_1 = r_{12}\,\delta_2$$

con lo cual se obtiene:

$$\delta_1 = 1\,\delta_2$$

O sea, que en este caso, no hay necesidad de multiplicar por ningún número la desviación sigmática del segundo fenómeno para obtener la desviación sigmática del primero. En efecto, si en el cuadro se comparan las columnas δ_1 y δ_2, podrá observarse que son prácticamente iguales (que es precisamente lo que expresa la ecuación de regresión en este caso).

Si sustituimos las desviaciones sigmáticas simbolizadas por las deltas minúscula por el cociente del que proceden (desviaciones con respecto a la media aritmética divididas entre la desviación media cuadrática o sigma minúscula), la expresión anterior resulta:

$$\frac{d_1}{\sigma_1} = \frac{d_2}{\sigma_2}$$

De acuerdo con esta expresión: en el caso considerado, la relación matemática existente entre la desviación de los datos de la primera serie con respecto a su media y la desviación media cuadrática de la serie es igual a la que existe entre la desviación de los datos de la segunda serie y su desviación cuadrática media.

Si se despeja a las d_1, se tiene:

$$d_1 = d_2 \frac{\sigma_1}{\sigma_2}$$

O sea, de acuerdo con esta segunda expresión, que: para obtener las desviaciones de la primera serie con respecto a su media es necesario multiplicar las desviaciones de la segunda serie con respecto a la suya por la relación entre las desviaciones cuadráticas medias de la primera y de la segunda serie.

Si las desviaciones con respecto a la media aritmética de cada serie simbolizadas por des minúsculas se sustituyen por la diferencia entre cada dato y la media, la expresión que acabamos de registrar se trasforma en:

$$x_1 - \bar{x}_2 \frac{\sigma_1}{\sigma_2} (x_2 - \bar{x}_2)$$

Si se despeja a x_1, se obtiene:

$$x_1 = \bar{x}_1 + \frac{\sigma_1}{\sigma_2} (x_2 - \bar{x}_2)$$

De acuerdo con esta expresión, el valor de los datos de la primera serie es igual (en este caso de correlación perfecta) a la media aritmética de esa primera serie más el producto de las desviaciones de la segunda serie respecto a la media de esta segunda serie por la relación entre las desviaciones medias cuadráticas de la primera y de la segunda serie.

Como hemos dicho anteriormente, cuando la correlación es perfecta, puede pensarse que se trata de una correlación no de primer orden, sino de orden cero, o sea, de una correlación de una serie con respecto a sí misma. O sea, que puede pensarse que los datos de una serie no son sino los datos de la otra serie *transcritos en unidades diferentes.* En este sentido, las tres expresiones principales encontradas

$$d_1 = d_2$$

$$\frac{d_1}{\sigma_1} = \frac{d_2}{\sigma_2}$$

$$x_1 - \bar{x}_1 = \frac{\sigma_1}{\sigma_2}(x_2 - \bar{x}_2)$$

permiten trasformar una serie en otra igual pero cuya desviación cuadrática media sea diferente; o en otra igual pero cuya desviación cuadrática media y cuya media aritmética sean distintas.

Así si se tiene una serie cuya desviación cuadrática media sea 3 y se quiere trasformar en una serie equivalente, cuya desviación media cuadrática sea 5, habrá que obtener las desviaciones de dicha serie con respecto a su media aritmética y multiplicarlas por 5/3 obteniendo así las desviaciones de la serie buscada con respecto a su media aritmética. Si se desea que la serie continúe teniendo la misma media aritmética de la serie original, bastará agregar dicha media a las desviaciones obtenidas para obtener los datos de la nueva serie. Si se quiere cambiar también la media, en vez de agregar a las desviaciones la antigua media, habrá que agregarles la media deseada para obtener los datos de la nueva serie equivalente, con desviación cuadrática media y media aritmética distintas.

En caso de que r_{12} no sea igual a 1, como en el ejemplo, las expresiones generales con las que habrá que trabajar serán:

$$\delta_1 = r_{12}\,\delta_2$$

$$d_1 = r_{12}\frac{\sigma_1}{\sigma_2}d_2$$

$$x_1 = \bar{x}_1 + r_{12}\frac{\sigma_1}{\sigma_2}(x_2 - \bar{x}_2)$$

Esta última forma de expresión de la ecuación de estimación es la más generalmente empleada en cuanto más explícita. Sin embargo, es fácil de obtener a partir de la primera con sólo recordar la equivalencia de las desviaciones sigmáticas.

Obtención de la fórmula práctica de cálculo del índice de correlación rectilínea de primer orden o fórmula del momento producto. Conforme hemos visto:

$$r_{12} = \frac{\Sigma\,\delta_1\,\delta_2}{\Sigma\,\delta_2^{\,2}}$$

Esta fórmula puede permitir que la designemos con el nombre de fórmula del momento producto por lo siguiente:

Si dividimos numerador y denominador de la expresión por el número de datos N, tenemos:

$$r_{12} = \frac{\dfrac{\Sigma\delta_1\,\delta_2}{N}}{\dfrac{\Sigma\delta_2^{\,2}}{N}}$$

Como puede verse por esta expresión, en el denominador está claramente un momento, pues se trata de la suma de potencias de desviaciones entre el número de datos. En cuanto la potencia es la segunda, se trata del segundo momento. En cuanto las desviaciones son de la segunda serie univariada, se trata del segundo momento de la segunda variable. Y en cuanto las desviaciones son desviaciones sigmáticas, se trata del segundo momento de la segunda variable en unidades sigmáticas.

Por lo que se refiere al numerador, la expresión que en él figura es parecida a la del denominador sólo que ahora no se trata de suma de potencias entre número de datos, sino suma de productos de desviaciones entre número de datos. De ahí que podamos hablar de un momento que, para diferenciarse de los momentos ordinarios, llamaremos "momento producto". En cuanto se trata de la primera potencia de las desviaciones de la primera serie y también de la primera potencia de las desviaciones de la segunda serie, hablaremos del momento de orden 11. En cuanto las desviaciones son las de las dos series, diremos que se trata del momento de orden 11 de la serie bivariada. Y en cuanto las desviaciones son sigmáticas, se trata del segundo momento (momento de orden 11) de la serie bivariada en unidades sigmáticas. De acuerdo con esto, la expresión puede escribirse:

$$r_{12} = \frac{\mu_{11\sigma}}{\mu_{2\sigma}} = \frac{M_{11}}{M_2}$$

Como el subíndice del momento del denominador no es unívoco, puesto que existen dos series univariadas y, consecuentemente dos segundos momentos, puede aceptarse la convención de designar como $\mu_{02}{}^{\sigma}$ o M_{02} al segundo momento sigmático de la segunda variable y designar como $\mu_{20}{}^{\sigma}$ o M_{20} el segundo momento sigmático de la primera variable.

De acuerdo con esto, la expresión resulta ser:

$$r_{12} = \frac{M_{11}}{M_{02}}$$

De acuerdo con esta expresión, el índice de correlación rectilínea de primer orden de una distribución bivariada está dado por el momento producto sigmático de la serie expresado en unidades del segundo momento sigmático de la variable independiente. Sin embargo:

$$M_{02} = N$$

Consecuentemente:

$$M_{11} = \frac{r_{12}}{N}$$

Es decir, que el índice de correlación rectilínea de primer orden es igual al momento producto de las dos variables.

Para usos prácticos el valor de r_{12} puede expresarse en forma sencilla mediante una fórmula obtenida como sigue:

Al sustituir en la expresión

$$r_{12} = \frac{\Sigma\,\delta_1\,\delta_2}{N}$$

las desviaciones sigmáticas por sus equivalentes d_1/σ_1 y d_2/σ_2 se tiene, para el numerador:

$$\delta_1\,\delta_2 = \Sigma\,\frac{d_1}{\sigma_1}\,\frac{d_2}{\sigma_2} = \frac{\Sigma d_1\,d_2}{\sigma_1\,\sigma_2}$$

Como el denominador es igual con N, al sustituir estos valores en la expresión que nos sirve de punto de partida, tendremos:

$$r_{12} = \frac{\dfrac{\Sigma d_1\,d_2}{\sigma_1\,\sigma_2}}{N}$$

O, finalmente, puesto que $\sigma_1 \sigma_2$ que figura en el denominador de la fracción numeradora puede pasar al denominador de la fracción principal como factor:

$$r_{12} = \frac{\Sigma\, d_1\, d_2}{N\, \sigma_1\, \sigma_2}$$

Esta es la fórmula clásicamente conocida como "fórmula del momento-producto" para el cálculo del índice de correlación.

Procedimiento.—De acuerdo con la fórmula anterior, el procedimiento de cálculo es el siguiente:

1º	Calcúlense las medias aritméticas de cada una de las series	x_1, x_2
2º	Réstese de cada dato (x_1, x_2) de cada serie la media aritmética correspondiente para obtener las desviaciones con respecto a las medias aritméticas correspondientes	d_1, d_2
3º	Multiplíquese cada par de desviaciones	$d_1 d_2$
4º	Súmense dichos productos	$\Sigma d_1 d_2$
5º	Elévense al cuadrado, separadamente las desviaciones de cada serie y regístrense en sendas columnas	$d_1^2\ d_2^2$
	súmense al pie esos cuadrados	$\Sigma d_1^2\ \Sigma d_2^2$
	divídanse esas sumas entre el número de datos	$\Sigma d_1\, 2/N,\ \Sigma d_2 2/N$
	y extráiganse las raíces cuadradas de los cocientes	$\sqrt{\Sigma d_1\, 2/N}\quad \sqrt{\Sigma d_2 2/N}$

6. Multiplíquense las desviaciones cuadráticas medias entre sí y por el número de pares de observaciones $N\sigma_1 \sigma_2$
entre el valor obtenido en el sexto paso $\sigma_1 \sigma_2$
a fin de obtener el índice de correlación r_{12}

Véase el ejemplo de aplicación de este procedimiento.

Observación.—Con vistas a ulteriores utilizaciones, a partir de la expresión general de r_{12}, si se pasa el denominador del segundo miembro como factor del primero, se tiene:

$$\Sigma d_1 d_2 = N r_{12} \sigma_1 \sigma_2$$

O sea, que la suma de los productos de las desviaciones es igual a tantas veces el producto de las desviaciones medias cuadráticas por el coeficiente de correlación como datos haya.

Una última simplificación en el cálculo directo del índice de correlación rectilínea de primer orden puede obtenerse si tomamos como punto de partida la más reciente de las fórmulas consignadas para dicho índice:

$$r_{12} = \frac{\Sigma d_1 d_2}{N \sigma_1 \sigma_2}$$

Los valores de las desviaciones cuadráticas medias de cada una de las dos variables son:

$$\sigma_1 = \sqrt{\frac{\Sigma d_1^2}{N}} \qquad \sigma_2 = \sqrt{\frac{\Sigma d_2^2}{N}}$$

Consiguientemente, el valor del denominador de r_{12} es:

$$N\sigma_1\sigma_2 = \sqrt{\frac{\Sigma d_1^2}{N}}\sqrt{\frac{\Sigma d_2^2}{N}}$$

$$= N\frac{\sqrt{\Sigma\delta_1^2\,\Sigma\delta_2^2}}{N}$$

Finalmente, en cuanto tanto el denominador de la expresión cocomo su coeficiente son N, se reducen a la unidad y toda la expresión se reduce al radical. Sustituido este valor en el denominador de r_{12}, se tiene:

$$r_{12} = \frac{\Sigma d_1\,d_2}{\sqrt{\Sigma d_1^2\,\Sigma d_2^2}}$$

A partir de esta fórmula, es posible diseñar el procedimiento —muy sencillo— de cálculo directo del índice de correlación rectilínea de primer orden. Dicho procedimiento consistirá en lo siguiente:

1º	Calcúlense las medias aritméticas de cada una de las dos series	$\bar{x}_1, \bar{x}_2$
2º	Réstese cada media de los valores de la serie correspondiente para obtener desviaciones con respecto a la media aritmética	d_1, d_2
3º	Multiplíquese cada desviación de la primera serie con respecto a su media, por la correspondiente desviación de la segunda serie con respecto a la suya	$d_1 d_2$
4º	Súmense esos productos	$\Sigma\, d_1\, d_2$
5º	Elévense al cuadrado las desviaciones de la primer variable	d_1^2
	y súmense los cuadrados obtenidos	$\Sigma\, d_1^2$

	Hágase lo propio con las desviaciones de la segunda variable	$\Sigma\, d_2^2$
6º	Multiplíquense esas sumas	$\Sigma\, d_1^2\, \Sigma\, d_2^2$
7º	Extráigase la raíz cuadrada del producto	$\sqrt{\Sigma\, d_1^2\, \Sigma\, d_2^2}$
8º	Divídase la suma obtenida en el cuarto paso entre la raíz obtenida en el 7º paso, para obtener el índice de correlación	r_{12}

La fórmula que nos ha permitido establecer este procedimiento muy simple, y que en seguida aplicaremos prácticamente, nos muestra que el índice de correlación rectilínea de primer grado es la relación entre el momento producto de las desviaciones con respecto a las medias aritméticas ($\Sigma\, d_1\, d_2$) y la media geométrica de los segundos momentos con respecto a las correspondientes medias aritméticas ($\Sigma\, d_1\, d_2$ y $\Sigma\, d_2^2$). ¿Por qué media geométrica en el denominador? Porque en la media geométrica es necesario *multiplicar* los valores promediados, y si en el numerador se trata de un momento *producto*, es necesario que, en forma correspondiente, en el denominador, se dé igualmente una multiplicación.

Error de estimación de la correlación rectilínea de primer orden.—Al tratar de definir la correlación, hemos hecho referencia a la ecuación de estimación, al índice de correlación y al error de estimación. El error de estimación, de acuerdo con nuestra definición inicial, mide el grado en que las estimaciones obtenidas de la ecuación de estimación se aproximan a los valores observados y el mismo error de estimación se encuentra íntimamente relacionado con el índice de correlación, de tal modo que, conforme es mayor la correlación es menor el error de estimación.

El error de estimación de la variable 1 a partir de la variable 2 se representa por $\varepsilon_{1.2}$.

Fundamentalmente, el error de estimación de una variable dependiente a partir de una independiente es una medida análoga a la desviación media cuadrática de una serie univariada. Pero, mientras en el cálculo de la desviación media cuadrática de una serie univariada intervienen las desviaciones *con respecto a la media aritmética,* de la serie, en el cálculo del error de estimación de una serie bivariada intervienen las desviaciones, pero *con respecto a los valores estimados* de la variable dependiente *a partir de la ecuación de estimación.*

De este modo:

$x_1 - \bar{x}_1 =$ desviacionesde los valores de la primer variable o variable dependiente con respecto a la media aritmética de dicha variable

$\sigma_1 = \sqrt{\dfrac{\Sigma\,(x_1 - \bar{x}_1)^2}{N}}$ desviación media cuadrática de la variable dependiente.

Pero si por $\hat{x}_1$ (x sub-1 estimada o con acento circunflejo) representamos la x_1 que podemos obtener a partir de la ecuación de estimación $\hat{x}_1 = \bar{x}_1 + r_{12} \dfrac{\sigma_1}{\sigma_2} (x_2 - \bar{x}_2)$

$x_1 - \hat{x}_1 =$ desviaciones de los valores observados de la primer variable con respecto a los valores *estimados* de esa misma variable a partir de la ecuación de estimación.

$\varepsilon_{.12} = \sqrt{\dfrac{\Sigma\,(x_1 - \hat{x}_1)^2}{N}}$ desviación media cuadrática de la variable dependiente con respecto a la independiente o ERROR DE ESTIMACION.

CALCULO DIRECTO DEL INDICE DE CORRELACION RECTILINEA DE ORDEN SUPERIOR AL PRIMERO

Hemos establecido, en páginas anteriores, que el índice de correlación rectilínea de orden superior al primero puede calcularse a partir de la fórmula:

$$r_{123}^{2} = 1 - \frac{R_{3x3}}{R_{11(3x3)}}$$

si se trata de una correlación múltiple entre tres variables.

El valor de R_{3x3} entre $R_{11(3x3)}$ o sea, el valor de un determinante de tres elementos por lado (las tres r minúsculas o índices de correlación rectilínea de primer orden) y el cofactor correspondiente al elemento de su primer renglón y su primera columna, lo hemos obtenido ya al hablar del error sigmático de estimación de la correlación múltiple. Ahí llegamos a la conclusión de que:

$$\frac{R_{3x3}}{R_{11(3x3)}} = \frac{1 - r_{12}^{2} - r_{13}^{2} - r_{23}^{2} + 2\, r_{12}r_{13}r_{23}}{1 - r_{23}^{2}} =$$

Al sustituir este equivalente en el valor de r_{123}^{2}, se obtiene:

$$r_{123}^{2} = 1 - \frac{1 - r_{12}^{2} - r_{13}^{2} - r_{23}^{2} + 2\, r_{12}r_{13}r_{23}}{1 - r_{23}^{2}} =$$

$$= \frac{1 - r_{23} - 1 + r_{12}^{2} + r_{13}^{2} + r_{23}^{2} - r_{12}r_{13}r_{23}}{1 - r_{23}^{2}} =$$

$$= \frac{r_{12}^{2} + r_{13}^{2} - 2\, r_{12}r_{13}r_{23}}{1 - r_{23}^{2}}$$

Consiguientemente, tendremos, como fórmula final, la siguiente:

$$r_{123} = \left(\frac{r_{12}^2 + r_{13}^2 - 2\, r_{12} r_{13} r_{23}}{1 - r_{23}^2} \right)^{\frac{1}{2}}$$

De acuerdo con esto, el procedimiento de cálculo no puede ser más simple. En efecto, se necesitará:

1º Calcular los índices de correlación rectilínea de primer orden entre la primera y la segunda, r_{12}
entre la primera y la tercera, r_{13}
entre la segunda y la tercera variables r_{23}

2º Multiplicar entre sí los coeficientes-índices de correlación así encontrados y multiplicar por 2 el producto $2r_{12}r_{13}r_{23}$

3º Elevar separadamente al cuadrado cada uno de los índices de correlación rectilínea de primer orden $r_{12}^2,\ r_{13}^3,\ r_{23}^2$

4º Sumar los dos primeros cuadrados y restar de la suma el producto encontrado en el 2º paso, para obtener el numerador.

5º Restar de uno el tercer cuadrado, para obtener el denominador.

6º Dividir el resultado del 4º paso entre el del 5º.

7º Extraer la raíz cuadrada del cociente obtenido en el 6º paso. r_{123}

PROCEDIMIENTO DE CALCULO DE LOS INDICES DE CORRELACION PARCIAL

1. Réstese de las desviaciones sigmáticas reales (calculadas mediante resta de la media y división entre la desviación media cuadrática) las desviaciones sigmáticas estimadas en el caso de cada una de las dos variables entre las que calcula el índice de correlación parcial (o sea, réstese de las desviaciones sigmáticas de la primer variable, obtenidas de restar la media de esa variable de los datos originales y dividir los resultados entre la desviación media cuadrática de dichos datos, la desviación sigmática de esa misma variable primera, pero calculadas mediante la ecuación de estimación a partir de la tercer variable). Así se obtienen $\delta_1 - \hat{\delta}_{1.3}$ o sea $\Delta_{1.3}$ y $\delta_2 - \hat{\delta}_{2.3}$ o sea $\Delta_{2.3}$

2. Obténgase el índice de correlación entre las desviaciones así calculadas $\Delta_{1.3}$ y $\Delta_{2.3}$.

Si se desciende al detalle del procedimiento, que puede seguirse en la ejemplificación subsecuente, se tendrá:

1º A partir de los datos originales de cada variable:

a.—Calcúlense las respectivas medias aritméticas $\bar{x}_1, \bar{x}_2\ \bar{x}_3,$

b.—réstense las medias de los datos para obtener desviaciones con respecto a la media aritmética d_1, d_2, d_3

c.—elévense al cuadrado, súmense los cuadrados, divídanse entre el número de datos y extráigase la raíz cuadrada del cociente para obtener las desviaciones medias cuadráticas $\sigma_1, \sigma_2, \sigma_3$

d.—divídanse las desviaciones con respecto a la media aritmética entre las desviaciones sigmáticas de las variables $\delta_1, \delta_2, \delta_3$

2º Multiplíquense por las desviaciones sigmáticas de la tercer variable (δ_3)

a.—las desviaciones sigmáticas de la primera variable $(\delta_1 \delta_3)$

b.—las desviaciones sigmáticas de la segunda variable $(\delta_2 \delta_3)$

c.—las desviaciones sigmáticas de la tercer variable $(\delta_3)^2$

d.—Súmense al pie los productos obtenidos en las tres columnas $\Sigma \delta_1 \delta_3, \Sigma \delta_2 \delta_3$ $\Sigma \delta_3^2$

e.—divídanse las dos primeras sumas entre la tercera para obtener los correspondientes índices de correlación simple r_{13}, r_{23}

3º Multiplíquense las desviaciones sigmáticas de la tercer variable (δ_3)

a.—por el índice de correlación de esta variable y de la primera para obtener la desviación sigmática de la primera *estimada* de la tercera $\hat{\delta}_{1.3}$

b.—por el índice de correlación de esta variable y de la segunda para obtener la desviación sigmática de la segunda *estimada* a partir de la tercera también $\hat{\delta}_{2.3}$

4º Réstese de las desviaciones sigmáticas *calculadas* (en el 1er. paso) de la primera y la segunda variable, las correspondientes desviaciones sigmáticas *estimadas* (en el tercer paso) a partir de la tercer variable, a fin de obtener desviaciones de desviaciones sigmáticas: $\Delta_{1.3}, \Delta_{2.3}$

5. A partir de las desviaciones de desviaciones sigmáticas,

a.—calcúlense las respectivas medias aritméticas $\bar{\Delta}_{1.3}, \bar{\Delta}_{2.3}$

b.—réstense de las respectivas deltas mayúsculas para obtener desviaciones de estas respecto de sus medias $d_{1.3}, d_{2.3}$

c.—elévense al cuadrado los resultados, súmense los cuadrados, divídanse entre el número de datos y extráigase la raíz cuadrada del cociente para obtener las desviaciones medias cuadráticas correspondientes $\sigma_{1.3}, \sigma_{2.3}$

d.—divídanse las desviaciones obtenidas (en b de este mismo paso) entre las desviaciones cuadráticas medias recién calculadas (en c) a fin de obtener desviaciones sigmáticas $\delta_{1.3}, \delta_{2.3}$

6º Multiplíquense por la desviación sigmática de 2 en 3 ($\delta_{2.3}$):

a.—la desviación sigmática de 1 en 3 para obtener $\delta_{1.3}\,\delta_{2.3}$

b.—la desviación sigmática de 2 en 3 para obtener $\delta_{2.3}{}^2$

7º Súmense separadamente los dos resultados del paso anterior para obtener $\Sigma\delta_{1.3}\,\delta_{2.3}$

y $\Sigma\delta_{2.3}{}^2$

8º Divídase la primera suma entre la segunda, a fin de obtener el índice de correlación parcial de la 1ª variable en la 2ª una vez eliminada la influencia de la 3ª $r_{1.3:2.3} = r_{1(3)2}$

Cálculo de uno de los índices de correlación parcial de una distribución trivariada.

Procedimiento Originario

Valores calculados a partir de los datos			Cálculo de los índices de correlación simple o de 1er. orden			Desviaciones sigmáticas estimadas		Residuos de las desviaciones sigmáticas	
δ_1	δ_2	δ_3	$\delta_1\delta_3$	$\delta_2\delta_3$	δ_3^2	$\hat{\delta}_{1.3}$	$\hat{\delta}_{2.3}$	$\Delta_{1.3}$	$\Delta_{2.3}$
−4.012	−2.674	−8	32.096	21.392	64	−4	−2.666	−0.012	−0.008
−3.009	−2.006	−6	18.054	12.036	36	−3	−2.000	−0.009	−0.006
−2.006	−1.337	−4	8.024	5.348	16	−2	−1.333	−0.006	−0.004
−1.003	−0.668	−2	2.006	1.336	4	−1	−0.666	−0.003	−0.002
0.000	0.000	0	0.000	0.000	0	0	0.000	0.000	0.000
1.003	0.668	2	2.006	1.336	4	1	0.666	0.003	0.002
2.006	1.337	4	8.024	5.348	16	2	1.333	0.006	0.004
3.009	2.006	6	18.054	12.036	36	3	2.000	0.009	0.006
4.012	2.674	8	32.094	21.392	64	4	2.666	0.012	0.008
SUMAS			120.360	80.224	240				
entre 240			0.501	0.334	1				
aproximadamente			0.5	0.33					
			Momentos producto sigmáticos.						
			Indices de correlación						
			r_{13}	r_{23}					

	Residuos de las desviaciones sigmáticas tomados del cuadro anterior		Desviaciones de los residuos (en milísimos)		Cálculo de un índice de correlación simple entre los residuos de las desviaciones sigmáticas, o cálculo del índice de correlación parcial entre las dos primeras variables					
	$\Delta_{1.3}$	$\Delta_{2.3}$	$d_{1.3}$	$d_{2.3}$	$d_{1.3}^2$	$d_{2.3}^2$	$\delta_{1.3}$	$\delta_{2.3}$	$\delta_{1.3}\delta_{2.3}$	$\delta_{2.3}^2$
	−0.012	−0.008	−12	−8	144	64	−1.55	−1.54	2.39	2.37
	−0.009	−0.006	− 9	−6	81	36	−1.16	−1.16	1.35	1.35
	−0.006	−0.004	− 6	−4	36	16	−0.78	−0.77	0.61	0.59
	−0.003	−0.002	− 3	−2	9	4	−0.39	−0.39	0.15	0.15
	0.000	0.000	0	0	0	0	0.00	0.00	0.00	0.00
	0.003	0.002	3	2	9	4	0.39	0.39	0.15	0.15
	0.006	0.004	6	4	36	16	0.78	0.77	0.61	0.59
	0.009	0.006	9	6	81	36	1.16	1.16	1.35	1.35
	0.012	0.008	12	8	144	64	1.55	1.54	2.39	2.37
SUMAS	0	0			540	240			9.00	$8.92 \doteq 9$
MEDIAS	0	0							Momentos producto	
Medias de desviaciones al cuadrado					60	27				
Raíces cuadrados de dichas medias o desviaciones medias cuadráticas					7.75	5.19				
Momento producto sigmático (9.00/9) = 1									$1 = r_{1(3)2}$	

CALCULO DIRECTO DE LOS INDICES DE CORRELACION PARCIAL

La fórmula que emplearemos para el cálculo de los índices de correlación parcial es:

$$r_{d(e)i} = \frac{r_{di} - r_{de}r_{ei}}{\left[(1 - r_{de}^{2})\ (1 - r_{ei}^{2})\right]^{\frac{1}{2}}}$$

Fórmula relativamente fácil de recordar, pues en el numerador existe una diferencia entre el índice de correlación simple de la variable dependiente y de la independiente y el producto de los índices de correlación simple de cada una de estas variables y la eliminada, y en el denominador se presenta la media geométrica de los cuadrados de los errores sigmáticos de estimación de cada una de las variables, dependiente e independiente, cuando dichas variables se estiman a partir de la variable eliminada.

El procedimiento para el cálculo de estos índices de correlación parcial entre tres variables, puede esquematizarse como sigue:

1º Calcular los índices de correlación rectilínea de primer grado
entre la variable dependiente y la independiente, r_{di}
entre la variable dependiente y la eliminada r_{de}
entre la variable independiente y la eliminada r_{ei}

4º Multiplíquense los dos últimos índices de correlación $r_{de}\ r_{ei}$

3º Réstese del primero de esos índices el producto obtenido en el segundo paso, para obtener el numerador de la expresión $r_{di} - r_{de}r_{ei}$

3º Réstese del primero de esos índices el producto obtenido en el segundo paso, para obtener el numerador de la expresión $r_{di} - r_{de}r_{ei}$

4º Elévense al cuadrado cada uno de los dos últimos índices de correlación r_{de}^2, r_{ei}^2

5º Réstese cada uno de los valores así obtenidos de 1 $1 - r_{de}^2, 1 - r_{ei}^2$

6º Multiplíquense los dos residuos obtenidos en el quinto paso

7º Extráigase la raíz cuadrada del producto, para obtener el denominador.

8º Divídase la diferencia obtenida en el 3er. paso entre la raíz cuadrada obtenida en el 7º paso, para obtener el índice de correlación parcial $r_{d(e)i}$

ciones con respecto a la media aritmética en unidades sigmáticas: $\delta_1, \delta_2, \delta_3, \delta_4$.

Hemos dicho, en páginas anteriores, que una correlación parcial es aquella en la que se encuentra la correlación que existe entre un cierto número de variables una vez que se elimina la influencia de otra variable sobre todas y cada una de ellas.

Esa eliminación de la influencia de una variable (en este caso elegiremos δ_4) sobre las otras variables ($\delta_1, \delta_2, \delta_3$) la podemos lograr, como en el caso de la correlación parcial entre tres variables, mediante la correlación entre los residuos de cada variable (dada en forma de desviación sigmática) con respecto a su estimación a partir de la variable cuya influencia se quiere eliminar. En el caso de tres variables ($\delta_1, \delta_2, \delta_3$) de las que se quería eliminar la influencia de la tercera (δ_3), calculamos los residuos del valor calculado de la primer variable con respecto al valor estimado de la misma a partir de la tercera ($\delta_1 - \hat{\delta}_{13}$); hicimos cosa parecida con la segunda variable ($\delta_2 - \hat{\delta}_{23}$) y correlacionamos las nuevas variables residuales Δ_{13} y Δ_{23} obteniendo así el índice de correlación simple entre esas variables residuales que, a su vez, es el índice de correlación parcial (con eliminación de la influencia de la tercer variable) entre las variables originales.

En el caso de cuatro variables, si se desea eliminar la tercera —por ejemplo— tendremos que estimar cada variable a partir de esa tercera (δ_1 estimada a partir de δ_3 o sea $\hat{\delta}_{13}$; δ_2 estimada a partir de δ_3, o sea $\hat{\delta}_{23}$ y δ_4 estimada de δ_3, o sea $\hat{\delta}_{43}$).

Los residuos que se obtengan de restar de cada valor observado el valor estimado ($\delta_1 - \hat{\delta}_{13}$, $\delta_2 - \hat{\delta}_{23}$, $\delta_4 - \hat{\delta}_{43}$) los designaremos por deltas mayúsculas $\Delta_{13}, \Delta_{23}, \Delta_{43}$.

Entre estas deltas mayúsculas o variables residuales, puede calcularse una correlación rectilínea de segundo orden (correlación múltiple de tres variables) por el procedimiento que ya hemos esquematizado en otras páginas. El resultado será un índice de correlación múltiple entre esas variables, que designaremos por $r_{13:23:43}$. Ese índice de correlación múltiple entre esas variables residuales es, a su vez, el índice de correlación parcial entre las variables originarias (δ_1, δ_2, δ_3, δ_4) que podemos designar por $r_{1(3)24}$ puesto que se trata de una corerlación en la que la variable eliminada es la tercera.

Si partimos del hecho de que tratamos con una correlación múltiple (o rectilínea de segundo orden) entre las variables Δ_{13}, Δ_{23}, Δ_{43}, podremos establecer una relación entre ese índice de correlación múltiple y los índices de correlación simple entre esas variables. O sea, que podremos establecer que:

$$r_{13:23:43}^2 = 1 - \frac{\begin{vmatrix} 1 & r_{13:23} & r_{13:43} \\ r_{13:23} & 1 & r_{23:43} \\ r_{13:43} & r_{23:43} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{23:43} \\ r_{23:43} & 1 \end{vmatrix}}$$

Esta fórmula expresa, simplemente, que la correlación múltiple ($r_{13:23:43}$) entre las variables residuales (Δ_{13}, Δ_{23}, Δ_{43}) es igual a la unidad menos una fracción, cuyo numerador es un determinante o matriz cuadrada que tiene por lado tantos elementos como variables residuales existen (tres en este caso), y cuyo denominador es el cofactor del elemento de la primer columna y del primer renglón. Por su parte, los elementos de estos dos determinante (determinante y cofactor) son los índices de correlación *simple entre cada par* de variables *residuales.*

Sin embargo, ya hemos visto que la correlación simple entre variables residuales no es sino la correlación parcial entre las variables originales de las que se obtuvieron los distintos residuos. Consiguientemente, puede afirmarse que:

La correlación parcial entre 4 raviables (δ_1, δ_2, δ_3, δ_4) originarias es igual a la unidad, menos una fracción cuyo determinante del numerador tiene por lado un elemento menos que el número de variables originales, y cuyo denominador es el cofactor del elemento de la primera columna y del primer renglón. Por su parte, los elementos de estos dos determinantes son los índices de correlación *parcial entre cada trío* de variables *originales,* del que se ha eliminado continuamente la influencia de una y de la misma variable (en todos los casos del ejemplo, se ha eliminado la influencia de la tercer variable, como en otro caso puede eliminarse siempre la influencia de la segunda, etc.).

Conforme a la notación de la correlación parcial, la misma fórmula anterior se convierte en:

$$r_{1(3)24}^{\ 2} = 1 - \frac{\begin{vmatrix} 1 & r_{1(3)2} & r_{1(3)2} \\ r_{1(3)2} & 1 & r_{1(3)2} \\ r_{1(3)2} & r_{1(3)2} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{2(3)4} \\ r_{2(3)4} & 1 \end{vmatrix}}$$

De ahí que, en forma general, pueda establecerse entre el índice de correlación parcial de orden superior al primero (de orden n = número de variable menos dos) y los índices de correlación parcial de primer orden (1 = 3 variables menos 2) una relación análoga a la establecida entre los índices de correlación total de orden superior al primero (de orden n = número de variables menos una) y los índices de correlación total de primer orden (1 = 2 variables menos 1). Esa relación puede expresarse, como en el caso anterior, por una fórmula parecida a la inicial de este apartado de nuestro estudio. En los determinantes, no aparecerá la columna en la que la variable eliminada figuraría como variable dependiente, y tampoco aparecerá el renglón en que la variable eliminada figuraría como variable independiente, y las restantes r minúsculas serán no los índices de correlación simple, sino los índices de correlación entre las mismas dos variables que designan los subíndices de la fórmula original, pero parcializados por la introducción entre ellos de paréntesis que indican cuál es la variable eliminada. De este modo $r_{123\ldots(e)\ldots n}{}^2$ es igual a:

$$1 - \frac{\begin{vmatrix} r_{1(e)1} & r_{1(e)2} & r_{1(e)3} & \ldots\ldots & (\, r_{1(e)e} \,) & \ldots & r_{1(e)n} \\ r_{2(e)1} & r_{1(3)2} & r_{2(e)3} & \ldots\ldots & (\, r_{2(e)e} \,) & \ldots & r_{2(e)n} \\ r_{3(e)1} & r_{3(e)2} & r_{3(e)3} & \ldots\ldots & (\, r_{3(e)e} \,) & \ldots & r_{3(e)n} \\ (\, r_{e(e)1} & r_{e(e)2} & r_{e(e)3} & \ldots\ldots & (\, r_{e(e)e} \,) & \ldots & r_{e(e)n} \,) \\ r_{n(e)1} & r_{n(e)2} & r_{n(e)3} & \ldots\ldots & (\, r_{n(e)e} \,) & \ldots & r_{n(e)n} \end{vmatrix}}{C_{11}}$$

En esta expresión, hemos colocado entre paréntesis los elementos de la columna y del renglón eliminados; hemos señalado con puntos suspensivos los puntos en que puede ampliarse más o menos el determinante, según el número de variables con que se cuente. Como es fácil suponer, sigue siendo válida la observación de que los elementos de la diagonal principal son unitarios (si se correlaciona el residuo que se obtiene de restar de una variable su estimada a partir de otra, con ese mismo residuo, el resultado tiene que ser uno). Por otra parte, la observación acerca de la simetría, sigue siendo válida. El cofactor que figura en el denominador es, naturalmente el cofactor del primer elemento del determinante del numerador en esta fórmula y no —naturalmente—, el cofactor del primer elemento del determinante del numerador en la fórmula previa.

Como puede verse, a partir de esta relación muy sencilla entre índices de correlación del mismo tipo (totales todos o parciales todos), pero de diferente orden pueden calcularse tantos índices como sean necesarios, con sólo haber calculado previamente:

1º Los índices de correlación simple entre cada par de variables,

2º Los índices de correlación parcial entre cada trío de variables.

FORMULA SIMPLIFICADA PARA EL CALCULO DIRECTO DE LAS CORRELACIONES PARCIALES DE UNA DISTRIBUCION TETRAVARIADA A PARTIR DE LAS DE UNA DISTRIBUCION TRIVARIADA

Si en la fórmula siguiente:

$$r_{1(3)24}^2 = 1 - \frac{\begin{vmatrix} 1 & r_{1(3)2} & r_{1(3)4} \\ r_{1(3)2} & 1 & r_{2(3)4} \\ r_{1(3)4} & r_{2(3)4} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{2(3)4} \\ r_{2(3)4} & 1 \end{vmatrix}}$$

desarrollamos cada uno de los determinantes del numerador y del denominador, sustituimos sus equivalentes en dicha fórmula, hacemos las reducciones pertinentes, y extraemos la raíz cuadrada del primero y del segundo miembro, podremos obtener la expresión:

$$r_{1(3)24} = \left[\frac{r_{1(3)2}^2 - 2\, r_{1(3)2} r_{2(3)4} r_{1(3)4} + r_{1(3)4}^2}{1 - r_{2(3)4}^2} \right]^{\frac{1}{2}}$$

Esta fórmula, como podrá verse fácilmente, si se la compara con la que dejamos asentada páginas atrás, es en todo análoga a la fórmula para el cálculo del índice de correlación trivariada obtenido a partir de los índices de correlación bivariada. La diferencia entre una y otra fórmula estriba en que mientras en un caso se trata de un índice de correlación múltiple y de varios índices de correlación simple *total,* en este caso se trata de índices de correlación múltiple y de varios índices de correlación simple (si aceptamos esta denominación por extensión) *parcial.*

EJEMPLO DE CALCULO DE LAS CORRELACIONES SIMPLES, MULTIPLES Y PARCIALES DE UNA DISTRIBUCION TRIVARIADA

1.—Obtención de las desviaciones sigmáticas

	INDICES								
	Economía	Población	Trabajo	d_1	d_2	d_3	d_1^2	d_2^2	d_3^2
	21	9	49	0	0	+ 8	0	0	64
	30	3	48	+9	− 6	+ 7	81	36	49
	29	19	95	+8	+10	+54	64	100	2916
	18	6	20	−3	− 3	−21	9	9	441
	15	3	29	−6	− 6	−12	36	36	144
	15	4	47	−6	− 5	+ 6	36	25	36
	17	14	33	−4	+ 5	− 8	16	25	64
	28	14	36	+7	+ 5	− 5	49	25	25
	21	12	32	+0	+ 3	− 9	0	9	81
	16	6	21	−5	− 3	−20	25	9	400
SUMAS	210	90	410				316	274	4220
MEDIAS	21	9	41				31.6	27.4	422
RAÍCES							5.6	5.2	20.5

EJEMPLO DE CALCULO DE LAS CORRELACIONES SIMPLES, MULTIPLES Y PARCIALES DE UNA DISTRIBUCION TRIVARIADA

2.—Obtención de los índices de correlación bivariada

δ_1	δ_2	δ_3	$\delta_1\delta_2$	$\delta_1\delta_3$	$\delta_2\delta_3$
0.0	0.0	+0.4	0.00	0.00	0.00
+1.6	−1.1	+0.3	−1.76	+0.48	−0.33
+1.4	+1.9	+2.6	2.66	3.64	4.69
−0.5	−0.6	−1.0	+0.30	+0.50	+0.60
−1.1	−1.1	−0.6	+1.21	+0.66	+0.66
−1.1	−1.0	+0.3	+1.10	−0.33	−0.30
−0.7	+1.0	−0.4	−0.70	+0.28	−0.40
+1.3	+1.0	−0.2	+1.30	−0.26	−0.20
+0.0	+0.6	−0.4	+0.00	−0.00	−0.24
−0.9	−0.6	−1.0	+0.54	+0.90	+0.60
Sumas de términos positivos			+7.11	+6.46	−6.55
Sumas de términos negativos			−2.46	−0.59	−1.47
SUMAS			+4.65	+5.87	+5.08
SUMAS ENTRE EFECTIVO (10)			0.465	0.587	0.508
			r_{12}	r_{13}	r_{23}

$$r_{12} = 0.465$$
$$r_{13} = 0.487$$
$$r_{23} = 0.508$$

EJEMPLO DE CALCULO DE LAS CORRELACIONES SIMPLES, MULTIPLES Y PARCIALES DE UNA DISTRIBUCION TRIVARIADA

3.—Obtención directa del índice de correlación múltiple

$$r_{123} = \left(\frac{r_{12}^2 + r_{13}^2 - 2\, r_{12} r_{13} r_{23}}{1 - r_{23}^2} \right)^{\frac{1}{2}}$$

$r_{12} = 0.465$

$r_{13} = 0.587 \qquad \therefore\ 2\, r_{12} r_{13} r_{23} = 0.277322280$

$r_{23} = 0.508$

$r_{12}^2 = 0.465^2 = 0.216225$

$r_{13}^2 = 0.587^2 = \mathit{0.344569}$

$r_{12}^2 + r_{13}^2 = 0.560794$

$-2\, r_{12} r_{13} r_{23} = -\mathit{0.277322}$

0.283472 Numerador de la fracción.

$r_{23}^2 = 0.508^2 = 0.258064$

$1 - r_{23}^2 = 1 - 0.258064 = 0.741936$ Denominador de la fracción.

$$r_{123} = \left(\frac{0.283472}{0.741936} \right)^{\frac{1}{2}} = \sqrt{0.3822069}$$

$r_{123} = 0.618$

EJEMPLO DE CALCULO DE LAS CORRELACIONES SIMPLES, MULTIPLES Y PARCIALES DE UNA DISTRIBUCION TRIVARIADA

4.—Obtención directa de los índices de correlación múltiple

$$r_{2.13} = \left[\frac{r_{21}^2 + r_{23}^2 - 2\, r_{12} r_{13} r_{23}}{1 - r_{13}^2} \right]^{\frac{1}{2}}$$

$r_{12}^2 = 0.216225$

$r_{23}^2 = 0.258064$ $\qquad 2\, r_{12} r_{13} r_{23} = 0.77322280$

$r_{13}^2 = 0.344569$

$$\therefore r_{2.13} = \sqrt{0.300515} = 0.548$$

Cálculo de $r_{3.12} = \left[\frac{r_{31}^2 + r_{32}^2 - 2\, r_{12} r_{13} r_{23}}{1 - r_{12}^2} \right]^{\frac{1}{2}}$

$$\therefore r_{3.12} = \sqrt{0.415056} = 0.644$$

EJEMPLO DE CALCULO DE LAS CORRELACIONES SIMPLES, MULTIPLES Y PARCIALES

5.—Obtención directa de los índices de correlación parcial. Fórmulas:

$$_{1(3)2} = \frac{r_{12} - r_{13}\, r_{23}}{[\,(1 - r_{13}^{2})\ (1 - r_{23}^{2})\,]^{\frac{1}{2}}}$$

$$r_{1(2)3} = \frac{r_{13} - r_{12}\, r_{23}}{[\,(1 - r_{12}^{2})\ (1 - r_{23}^{2})\,]^{\frac{1}{2}}}$$

$$r_{2(1)3} = \frac{r_{23} - r_{21}\, r_{13}}{[\,(1 - r_{21}^{2})\ (1 - r_{13}^{2})\,]^{\frac{1}{2}}}$$

$$r_{1(3)2} = \frac{r_{12} - r_{13} r_{32}}{[(1 - r_{13}^{2})\ (1 - r_{23}^{2})\,]^{\frac{1}{2}}}$$

Substituciones: Cálculo de $r_{1(3)2}$.

$r_{12} = 0.465$

$r_{13} = 0.587$

$\therefore\ r_{13} r_{32} = 0.587 \times 0.508 = 0.298196$

$r_{23} = 0.508$

$r_{12} - r_{13} r_{32} = 0.465 - 0.298196 = 0.166804$ Numerador

$r_{13}^{2} = 0.587^{2} = 0.344569 \quad 1 - r_{13}^{2} = 1 - 0.344569 = 0.655431$

$r_{32}^{2} = 0.508^{2} = 0.258064 \quad 1 - r_{32}^{2} = 1 - 0.258064 = 0.741936$

$(1 - r_{13}^{2})\ (1 - r_{32}^{2}) = 0.655431 \times 0.741936 =$
$= 0.486287854416$

$\sqrt{0.486287854416} = 0.697342$ Denominador.

$$r_{1(3)2} = \frac{0.166804}{0.697342} = 0.24$$

Cálculo de $r_{1(2)3}$

$$r_{1(2)3} = \frac{r_{13} - r_{12}r_{23}}{[\,(1 - r_{12}^2)\ (1 - r_{23}^2)\,]^{\frac{1}{2}}}$$

$r_{12} = 0.465$
$r_{13} = 0.587$
$r_{23} = 0.508 \therefore r_{12}r_{23} = 0.465 \times 0.508 = 0.236220$

$r_{13} - r_{12}r_{23} = 0.587 - 0.236220 = 0.35078$ Numerador.

$r_{12}^2 = 0.465^2 = 0.216225 \quad 1 - r_{12}^2 = 1 - 0.216225 = 0.783775$

$r_{23}^2 = 0.508^2 = 0.258064 \quad 1 - r_{23}^2 = 1 - 0.258064 = 0.741936$

$(1 - r_{12}^2)\ (1 - r_{23}^2) = 0.783775 \times 0.741936 =$
$= 0.5815108884$

$\sqrt{0.5815108884} = 0.76256$ Denominador.

$$r_{1(2)3} = \frac{0.350780}{0.76256} = 0.46$$

Cálculo de $r_{1(2)3}$

$$r_{2(1)3} = \frac{r_{23} - r_{21}r_{13}}{[\,(1 - r_{21}^2)\ (1 - r_{13}^2)\,]^{\frac{1}{2}}}$$

$r_{23} = 0.508$
$r_{21} = 0.465$
$r_{13} = 0.587 \therefore r_{21}r_{13} = 0.465 \times 0.587 = 0.272955$

$r_{23} - r_{21}r_{13} = 0.508 - 0.272955 = 0.235045$ Numerador.

$r_{21}^2 = 0.465^2 = 0.216225 \quad 1 - r_{21}^2 = 1 - 0.216225 = 0.783775$

$r_{13}^2 = 0.587^2 = 0.344569 \quad 1 - r_{13}^2 = 1 - 0.344569 = 0.655431$

$(1 - r_{21}^2)\ (1 - r_{13}^2) = 0.783775 \times 0.655431 =$
$= 0.513710432025$

$\sqrt{0.513710432025} = 0.716735$ Denominador.

$$r_{2(1)3} = \frac{0.235045}{0.716735} = 0.3279$$

INTENTO DE INTERPRETACION DE LOS RESULTADOS DE LA CORRELACION

RESULTADOS:

$r_{12} = 0.465$	$r_{1.23} = 0.618$	$r_{1(3)2} = 0.239$
$r_{13} = 0.587$	$r_{2.13} = 0.548$	$r_{1(2)3} = 0.460$
$r_{23} = 0.508$	$r_{3.12} = 0.644$	$r_{2(1)3} = 0.328$

IDENTIFICACION DE LAS VARIABLES:

1.—Indice económico.
2.—Indice demográfico.
3.—Indice laboral.

COMPARACIONES: Entre los índices de correlación simple:

$$r_{12} = 0.465 < r_{13} = 0.587$$

Es menor la intensidad de la correlación entre economía y población que entre economía y trabajo.

O sea: conocido el valor del índice laboral se obtendrá un valor más aproximado del índice económico que si se conoce el valor del índice demográfico y se pretende estimar el del índice económico.

$$r_{21} = r_{12} = 0.465 < r_{23} = 0.508$$

Es menor la intensidad de la correlación entre población y economía que entre población y trabajo.

O sea: conocido el valor del índice laboral se obtendrá un valor más aproximado del índice demográfico que si se trata de estimar éste a partir del índice económico.

$r_{31} = r_{13} = 0.587 > r_{32} = r_{23} = 0.508$

Es mayor la intensidad de la correlación entre trabajo y economía que entre trabajo y población.

O sea: conocido el valor del índice económico es más aproximada la estimación que se obtenga del índice laboral que la que se obtendrá si se conociese el valor del índice demográfico.

$r_{13} > r_{23} > r_{12}$

O sea: que es la máxima intensidad en estas relaciones bipartitas (en las que aún no se elimina la influencia de la tercer variable) se da en el caso de las relaciones entre la economía y el trabajo; que la intensidad es media en el caso de las relaciones entre la población y el trabajo, y mínima en el caso de las relaciones entre la economía y la población.

COMPARACIONES: Entre los índices de correlación múltiple y los índices de correlación simple.

$r_{1.23} = 0.618 > r_{12} = 0.465$
$r_{1.23} = 0.618 > r_{13} = 0.587$

O sea: que la intensidad de la relación entre la economía por una parte, y la población y el trabajo por otra es mayor tanto si se la compara con la relación entre la economía y el trabajo solamente, como si se la comparara con la relación entre la economía y la población únicamente.

$$r_{2.13} = 0.548 > r_{21} = r_{12} = 0.465$$
$$r_{2.13} = 0.548 > r_{23} = 0.508$$

Como en el caso anterior la intensidad de la relación entre la población y la economía y el trabajo tomados conjuntamente es mayor que la intensidad de las relaciones de la población con la economía por una parte, y con el trabajo por otra.

$$r_{3.12} = 0.644 > r_{31} = r_{13} = 0.587$$
$$r_{3.12} = 0.644 > r_{32} = r_{23} = 0.508$$

Puede observarse nuevamente, que la intensidad de la relación aumenta cuando para la estimación del índice laboral se utilizan conjuntamente los índices económico y demográfico, en vez de utilizar cualquiera de ellos aisladamente.

Sin embargo, puede observarse también que en la estimación de "economía" el incremento fue más sustancial al agregar el factor 3 "trabajo" que al agregar el factor 2 "población"

$$r_{1.23} - r_{12} = 0.618 - 0.465 = 0.153$$

es mayor que

$$r_{1.23} - r_{13} = 0.618 - 0.587 = 0.031$$

En la estimación de la población, el incremento fue mayor al agregar el factor 1 "economía" que al agregar el factor 3 "trabajo" ($r_{2.13} - r_{21} = 0.548 - 0.465 = 0.083 > r_{2.13} - r_{23} = 0.548 - 0.508 = 0.040$)

En la estimación de "trabajo", los resultados muestran que
$r_{3.21} - r_{31} = 0.644 - 0.587 = 0.057 <$
$r_{3.21} - r_{32} = 0.644 - 0.058 = 0.136$

O sea, que hubo un incremento mayor cuando se agregó el factor 1 "economía", que cuando se agregó el factor 2 "población".

Todo lo anterior parece poner de manifiesto el gran peso que tiene el factor económico y, en íntima vinculación con él el que también tiene el factor trabajo, dentro del conglomerado social al que corresponden estos índices

COMPARACIONES: De los índices de correlación múltiple entre sí:

$r_{3.12} = .644 > r_{1.23} = .618 > r_{2.13} = .548$

O sea, que la correlación es más intensa cuando es la variable 3, o sea "trabajo" la que funciona como variable dependiente, y que es mínima su intensidad cuando es la variable 2, o sea "población" la que se considera como variable dependiente. De acuerdo con esto, puede considerarse que, contra lo que se postulaba implícitamente al tomar "economía" como primer variable en el cálculo de nuestras correlaciones, es el índice laboral el que depende de los índices económico y demográfico o sea, que, al menos en este caso, el trabajo depende de la economía y de la población mucho más de lo que éstas dependen de aquél. Lo cual no impide pensar que, en el momento siguiente (dentro de una concepción dinámica, diacrónica) el trabajo pueda pasar a ser variable independiente que explique, junto con la población, la economía de ese otro momento.

COMPARACIONES: De los índices de correlación simple y los índices de correlación parcial.

$$\frac{r_{1(3)2}}{r_{12}} = \frac{0.239}{0.465} = 51\%$$

$$\frac{r_{1(2)3}}{r_{13}} = \frac{0.460}{0.587} = 78.3\%$$

$$\frac{r_{2(1)3}}{r_{23}} = \frac{0.328}{0.508} = 64.5\%$$

Puede observarse que: el que menos se redujo fue $r_{1(2)3}$ en relación con el valor de la relación entre 1 y 3 antes de la eliminación de 2. O sea, que el factor 2 "población" es el que menos perturbaba esas relaciones. El que más se redujo fue el índice r_{12} al eliminar la influencia de 3, ya que el nuevo índice sólo representa la mitad (51%) del antiguo. O sea, que el factor 3 "trabajo" es el que mayor influencia perturbadora ejerce en las relaciones económico-demográficas.

EJEMPLO DE CALCULO DE LOS INDICES DE CORRELACION PARCIAL DE UNA DISTRIBUCION TETRAVARIADA

Las tres primeras variables son los índices económico, demográfico y laboral del ejemplo anterior. Son igualmente válidos para este ejemplo los cálculos de las desviaciones sigmáticas de las tres primeras variables del ejemplo previo, así como los índices de correlación bivariada calculados para el mismo. Según esto, el cálculo que sigue se reduce: al cálculo de la desviación sigmática de la cuarta variable y a la determinación de los índices de correlación bivariada en que interviene dicha variable cuarta (r_{14}, r_{24}, r_{34}) para obtener a partir de ellos los índices de correlación parcial que se buscan.

INDICE Educación	d_4	d_4^2	δ_4	$\delta_1 \delta_4$	$\delta_2 \delta_4$	$\delta_3 \delta_4$
70	+11	121	+0.7	0.00	0.00	+0.28
77	+18	504	+1.1	+1.76	−1.21	+0.33
82	+23	504	+1.1	+2.10	+2.85	+3.90
32	−27	729	−1.7	+0.85	+1.02	+1.70
40	−19	361	−1.2	+1.32	+1.32	+0.72
46	−13	169	−0.8	+0.88	+0.80	−0.24
60	+ 1	1	+0.1	−0.07	+0.10	−0.04
65	+ 6	36	+0.4	+0.52	+0.40	−0.08
58	− 1	1	−0.1	−0.00	−0.60	+0.04
60	+ 1	1	+0.1	−0.09	−0.06	−0.10
Positivos				+7.43	+6.49	+6.97
Negativos				−0.07	−1.87	−0.46
SUMAS				7.36	4.62	6.51
Entre el efecitvo (10)				0.736	0.462	0.651
				r_{14}	r_{24}	r_{34}

$r_{14} = 0.736$

$r_{24} = 0.462$

$r_{34} = 0.651$

$$r_{1(3)4} = \frac{r_{14} - r_{13}\, r_{34}}{\left[(1 - r_{13}^2)\ (1 - r_{34}^2) \right] \frac{1}{2}}$$

$r_{13}r_{34} = 0.587 \times 0.651 = 0.382137$

$\underline{r_{14} - r_{13}r_{34}} = 0.736 - 0.382137 = 0.353863$ Numerador.

$r_{13}^2 = 0.587^2 = 0.344569 \quad 1 - r_{13}^2 = 1 - 0.344569 = 0.655431$

$r_{34}^2 = 0.651^2 = 0.423801 \quad 1 - r_{34}^2 = 1 - 0.423801 = 0.576199$

$$(1 - r_{13}^2)\ (1 - r_{34}^2) = 0.655431 \times 0.576199 = \\ = 0.377658686769$$

$\sqrt{0.377658686769} = 0.614539$ Denominador

$$r_{1(3)4} = \frac{0.353863}{0.614539} = 0.576$$

Cálculo de $r_{2(3)4}$.

$$_{2(3)4} = \frac{r_{24} - r_{23}\, r_{34}}{\left[(1 - r_{23}^2)\ (1 - r_{34}^2) \right] \frac{1}{2}}$$

$r_{23}r_{34} = 0.508 \times 0.651 = 0.330708$

$\underline{r_{24} - r_{23}r_{34}} = 0.462 - 0.330708 = 0.131292$ Numerador.

$r_{23}^2 = 0.508^2 = 0.258064 \quad 1 - r_{23}^2 = 1 - 0.258064 = 0.741936$

$r_{34}^2 = 0.651^2 = 0.423801 \quad 1 - r_{34}^2 = 1 - 0.423801 = 0.576199$

$$(1 - r_{23}^2)\ (1 - r_{34}^2) = 0.741936 \times 0.576199 = \\ = 0.427502781264$$

$\sqrt{0.427502781264} = 0.653830$ Denominador.

$$r_{2(3)4} = \frac{0.131292}{0.653830} = 0.2008$$

EJEMPLIFICACION DEL CALCULO DE LOS COEFICIENTES DE CORRELACION PARCIAL DE LA DISTRIBUCION TETRAVARIADA

$$r_{1(3)24} = \left\{ \frac{r_{1(3)2}^2 - 2\, r_{1(3)2} r_{2(3)4} r_{1(3)4} + r_{1(3)4}^2}{1 - r_{2(3)4}^2} \right\}^{\frac{1}{2}}$$

$r_{1(3)2} = 0.239$

$r_{1(3)2}\, r_{2(3)4} = 0.048039$

$r_{2(3)4} = 0.201$

$r_{1(3)2} r_{2(3)4} r_{1(3)4} = 0.027670464$

$r_{1(3)4} = 0.576$

$$2\, r_{1(3)2} r_{2(3)4} r_{1(3)4} = 2 \times 0.027670464 = 0.055340928$$

$$r_{1(3)2}^2 = 0.239^2 = 0.057121$$

$$r_{1(3)4}^2 = 0.576^2 = 0.331776$$

$$r_{1(3)2}^2 + r_{1(3)4}^2 - 2\, r_{1(3)2} r_{2(3)4} r_{1(3)4}$$

$$= 0.057121 + 0.331776 - 0.055340928 =$$

$$= 0.388897 - 0.055340928 = 0.333556072$$

Numerador.

$$r_{2(3)4}^2 = 0.201^2 = 0.040401$$

$$1 - r_{2(3)4}^2 = 1 - 0.40401 = 0.959599$$

Denominador

$$r_{1(3)24} = \left[\frac{0.333556072}{0.959599} \right]^{\frac{1}{2}} = (0.347993)^{\frac{1}{2}}$$

$$r_{1(3)24} = \sqrt{0.347997} = 0.59$$

INDICES DE CORRELACION PARCIAL OBTENIDOS MEDIANTE LA ELIMINACION DE LA INFLUENCIA DE TODAS LAS VARIABLES MENOS DOS, EN SERIES DE MAS DE TRES VARIABLES

Hasta este momento, hemos hablado de la eliminación de la influencia de una variable independiente en la correlación entre dos (en el caso de una distribución trivariada) o entre más de dos (en el caso de distribuciones de más de tres variables) de las variables independientes. En seguida nos ocuparemos del caso de la eliminación de las influencias de todas las variables independientes menos una, lo cual permite estimar la correlación neta entre la variable dependiente y la variable indepediente, cuya influencia no se ha eliminado.

Caso fronterizo entre un tipo de eliminación (una de todas las variables independientes) y el otro tipo de eliminación (todas las variables independientes menos una) lo es el de la correlación parcial en una distribución trivariada. En efecto, la correlación parcial en una distribución de las variables x_1, x_2, x_3 puede considerarse bien como eliminación de todas las variables independientes (x_2, x_3) menos una (x_3, por ejemplo), o como la eliminación de una de las variables independientes (x_3).

En cambio, la distinción entre los dos tipos de eliminación de influencia de las variables independientes aparece desde el momento en que se trata con distribuciones tetravariadas. En efecto, si tenemos la distribución conjunta de las variables x_1, x_2, x_3, x_4 y consideramos x_1 como dependiente, y las restantes como independientes, hay dos tipos de eliminación posibles: 1º—Eliminación de la influencia de x_4 (una de las variables independientes) en la correlación entre la variable dependiente x_1 y las restantes variables independientes x_2 y x_3, o 2º—Eliminación de la influencia de x_3 y x_4 (todas las variables independientes menos una) en la correlación entre la variable dependiente x_1 y la variable independiente x_2.

Con el fin de encontrar la fórmula que nos permita calcular el índice de correlación neta entre dos variables cuando se han eliminado las influencias de todas las restantes variables, tomaremos el caso más simple de una distribución tetravariada, cuyos datos habremos reducido a desviaciones con respecto a las correspondientes medias aritméticas expresadas en unidades de la desviación cuadrática media (o sea, que trabajaremos con desviaciones sigmáticas que, como de costumbre, representaremos por deltas minúsculas).

$$\delta_1\ \delta_2\ \delta_3\ \delta_4$$

Mediante las correlaciones simples correspondientes que lleguen a establecerse entre estas cuatro variables pareadas, es posible estimar el valor de cada una de ellas a partir de cualquiera de las restantes. Para nuestros propósitos actuales, nos interesa particularmente la posibilidad que hay de estimar cada una de estas variables a partir de la tercera. Las estimadas correspondientes se representarán por las propias deltas minúsculas, afectadas de un acento circunflejo (que representa "estimación") y la adición de un nuevo subíndice (3 en el caso) que indica cuál es la variable a partir de la cual se estima la variable representada. En estas condiciones, dichas estimaciones quedarán representadas por:

$$\hat{\delta}_{13}\ \hat{\delta}_{23}\ \hat{\delta}_{33}\ \hat{\delta}_{43}$$

De estas estimadas, la tercera, o sea $\hat{\delta}_{33}$ es idéntica a la variable observada δ_3.

Si se quiere eliminar la influencia de la tercer variable en las restantes, será necesario restar de cada variable observada, la estimación de esa mismo variable hecha a partir de la tercer variable. O sea, que habrá que restar de cada uno de los elementos del primer renglón que hemos escrito el elemento correspondiente del segundo renglón que acabamos de escribir; esto produce:

$$\delta_1 - \hat{\delta}_{13}, \delta_2 - \hat{\delta}_{23}, \delta_3 - \hat{\delta}_{33}, \delta_4 - \hat{\delta}_{43}$$

De estas diferencias, la tercera es nula, puesto que la estimada de la tercer variable con respecto a la tercer variable es idéntica a dicha variable tercera. Consiguientemente, la serie queda reducida a *tres* nuevas variables:

$$\delta_1 - \hat{\delta}_{13}, \delta_2 - \hat{\delta}_{23}, \delta_4 - \hat{\delta}_{43}$$

Si a estas desviaciones las representamos por deltas mayúsculas, como ya hemos hecho anteriormente, las tres nuevas variables serán:

$$\Delta_{1.3}, \Delta_{2.3}, \Delta_{4.3}$$

Con esta nueva serie de variables, podemos repetir el procedimiento que seguimos hasta aquí; es decir, podemos intentar una nueva reducción entre ellas. Cada una de estas variables puede estimarse a partir de cada una de las restantes. Para nuestros intereses presentes, conviene que nos fijemos en la posibilidad de estimar las dos primeras a partir de la tercera (lo cual permitirá la posterior eliminación de ésta) Las estimaciones correspondientes las representaremos también por deltas mayúsculas, afectándolas de un circunflejo y añadiendo al subíndice las cifras que caracterizan a la variable a partir de la cual se hace la estimación. Estas estimadas serán:

$$\hat{\Delta}_{1.3:4.3}, \hat{\Delta}_{2.3:4.3}, \hat{\Delta}_{4.3:4.3}$$

Para eliminar la influencia de la tercer variable de esta nueva serie de las otras dos variables (lo que en el fondo equivale a eliminar la influencia de la cuarta variable de la serie original de un conjunto de variables en las que ya se eliminó la influencia de la tercera variable), restaremos de cada uno de los valores de las tres nuevas variables, los correspondientes valores estimados.

$$\Delta_{1.3} - \hat{\Delta}_{1.3:4.3} \quad \Delta_{2.3} - \hat{\Delta}_{23:4.3} \quad \Delta_{4.3} - \hat{\Delta}_{43:4.3}$$

Como toda variable estimada de sí misma es igual a sí misma, la última de estas desviaciones es nula. Por lo mismo, al eliminar la influencia de la tercer variable de la nueva serie (que equivale a eliminar la influencia de la cuarta variable en la serie originaria de la que ya se había eliminado la influencia de la tercer variable), se obtiene una distribución bivariada de las novísimas variables:

$$\Delta_{1.3} - \hat{\Delta}_{13:4.3} \text{ y } \Delta_{2.3} - \hat{\Delta}_{2.3:4.3}$$

Estas nuevas variables las representaríamos por la dental sonora correspondiente del alfabeto devanagari, a no ser por imposibilidad tipográfica que nos obliga a reemplazarla por la dental sorda del alfabeto griego (τ). En estas condiciones, la serie tetravariada original se habrá reducido a la serie bivariada:

$$\tau_{1.34} \qquad \tau_{2.34}$$

Entre estas dos novísimas variables, es posible calcular un índice de correlación rectilínea que, en el fondo será el índice de correlación entre las variables 1 y 2 (cuyas influencias no se han eliminado) cuando se eliminan las influencias de las variables 3 y 4. Consiguientemente, el índice de correlación que calculamos podrá representarse, según el uso ya establecido, por:

$$r_{1(34)2}$$

Obtención de la fórmula para el cálculo del índice de correlación neta entre dos variables con eliminación de la influencia de las restantes en una distribución tetravariada.—Para establecer la fórmula del índice de correlación neta entre las dos variables que se han obtenido tras la eliminación de las influencias de las otras variables partiremos en la ecuación de regresión correspondiente. De acuerdo con cuanto se ha dicho y repetido anteriormente, afirmar que existe una correlación rectilínea entre dos variables equivale a afirmar que si se multiplica la desviación sigmática de la variable independiente por el índice de correlación, se deberá obtener la desviación sigmática de la variable dependiente. Dicho en otra forma (más conveniente para nuestro propósito actual), que si la desviación de la segunda variable respecto de su media, dividida entre su desviación media cuadrática se multiplica por el índice de correlación rectilínea, se obtendrá el cociente que resulta de dividir la desviación de la primer variable respecto de su media entre la desviación media cuadrática de dicha variable.

En el caso presente:

$\tau_{1.34}$ — primer variable = desviación de la primer variable respecto a su media (pues es fácil probar que esta media de desviaciones es nula).

$\tau_{2.34}$ — segunda variable = desviación de la segunda variable respecto a su media.

$r_{1(34)2}$ — coeficiente de correlación entre las variables.

En cuanto a las desviaciones medias cuadráticas de las variables, las representaríamos por las sibilantes dentales (correspondientes a las sigmas minúsculas del devanagari; pero por imposibilidad tipográfica las representamos por el nexo griego Ksi (ξ). En estas condiciones, la ecuación sigmática de estimación será:

$$\frac{\tau_{1.34}}{\xi_{1.34}} = r_{1(34)2} \frac{\tau_{2.34}}{\xi_{2.34}}$$

Para obtener el valor de r a partir de esta ecuación bastará con multiplicar ambos miembros de la ecuación por la fracción multiplicadora de $r_{1(34)2}$ pasar el cuadrado resultante en el segundo miembro al primero (tomando el numerador como denominador y viceversa), para obtener:

$$r_{1(34)2} = \frac{\Sigma\tau_{1.34}\,\tau_{2.34}}{\tau_{2.34}^{2}} \quad \frac{\xi_{2.34}}{\xi_{1.34}}$$

En el segundo miembro (como ya ha ocurrido en otro caso parecido) figura una relación entre desviaciones cuadráticas medias, que fácilmente puede reducirse a relación entre raíces cuadradas de sumas de cuadrados de las desviaciones correspondientes. O sea, que la expresión anterior equivale a la siguiente:

$$r_{1(34)2} = \frac{\Sigma\tau_{1.34}\,\tau_{2.34}}{\Sigma\tau_{2.34}^{2}} \quad \frac{\sqrt{\Sigma\tau_{234}^{2}}}{\sqrt{\Sigma\tau_{134}^{2}}}$$

Al reducir el radical del numerador con uno de los radicales que se encuentran implícitos en el denominador de la primer fracción, se obtiene:

$$r_{1(34)2} = \frac{\Sigma\tau_{1.34}\,\Sigma\tau_{2.34}}{\sqrt{\Sigma\tau_{1.34}^{2}\,\Sigma\tau_{2.34}^{2}}}$$

Expresión completamente análoga a la que ya obtuvimos antes para el coeficiente de correlación parcial-neta en el caso de series trivariadas.

RELACION ENTRE EL INDICE DE CORRELACION NETA DE UNA DISTRIBUCION TETRAVARIADA Y LOS INDICES DE CORRELACION PARCIAL-NETA DE LAS DISTRIBUCIONES TRIVARIADAS

Partiremos de la ecuación más sencilla del índice de correlación neta

$$r_{1(34)2} = \frac{\Sigma\tau_{1.34}\,\Sigma\tau_{2.34}}{\sqrt{\Sigma\tau_{2.34}^{2}\,\Sigma\tau_{2.34}^{2}}}$$

Para ello, estableceremos en primer término el equivalente del numerador, y en el segundo lugar el del denominador, para hacer en tercer término la sustitución final.

Para el numerador se tiene:

Si se sustituyen las das del alfabeto devanagari o sus sustitutos, las taus del alfabeto griego por sus equivalentes (diferencias entre deltas mayúsculas observadas y estimadas) el resultado es:

$$\Sigma(\Delta_{1.3} - \hat{\Delta}_{1.3:4.3})\ (\Delta_{2.3} - \hat{\Delta}_{2.3:4.3})$$

Al ejecutar las operaciones se obtiene:

$$\Sigma\Delta_{1.3}\,\Delta_{2.3} - \Sigma\Delta_{2.3}\,\hat{\Delta}_{1.3:4.3} - \Sigma\Delta_{1.3}\,\hat{\Delta}_{2.3:4.3} + \Sigma\hat{\Delta}_{1.3:4.3}\,\hat{\Delta}_{2.3:4.3}$$

En este desarollo, y de acuerdo con lo que ya asentamos en las primeras páginas consagradas al estudio de la correlación parcial:

1º El primer término que es suma de productos de desviaciones parcializadas es igual a un índice de correlación parcial entre las variables identificadas por los subíndices que preceden al punto, con eliminación de la influencia de la variable identificada por el subíndice (común) que subsigue al punto. O sea, que el primer término de dicho desarrollo es igual a:

$$r_{1(3)2}$$

2º El segundo término es una suma de productos. Los segundos factores son estimaciones que se pueden sustituir por un índice de correlación multiplicado por la variable correspondiente ($\hat{\Delta}_{1.3:4.3}$ puede sustituirse por $r_{1(3)4}\ \Delta_{4.3}$). Al hacer dicha sustitución se obtienen sumas de productos triples en los que los factores son un índice de correlación parcial ($r_{1(3)2}$) y dos factores delta mayúscula que pueden reemplazarse por otro índice de correlación parcial, de acuerdo con la secuela puesta de manifiesto en el primer término. De este modo, todo el segundo término se convierte en:

$$\text{MENOS } r_{1(3)4}\, r_{2(3)4}$$

3º En el tercer término se tiene nuevamente una delta mayúscula sin circunflejo y una delta mayúscula con circunflejo. La última puede reemplazarse por un índice de correlación parcial (en el caso $r_{2(3)4}$) multiplicado por una delta mayúscula sin circunflejo (en el caso $\Delta_{4.3}$). Mediante la última sustitución, quedan multiplicadas dos deltas mayúsculas sin circunflejo, que son sustituibles por un índice de correlación parcial (en el caso $r_{1(3)4}$). De este modo, el tercer término se reduce a:

$$\text{MENOS } r_{2(3)4}\, r_{1(3)4}$$

4º En el cuarto término se tiene el producto de dos deltas mayúsculas con circunflejo, cada una de las cuales puede sustituirse por un índice de correlación parcial y una delta mayúscula sin circunflejo (las sustituciones serían por $r_{1(3)4}\Delta_{4.3}$ y $r_{2(3)4}\Delta_{4.3}$). O sea, que se obtendría el producto de los dos índices de correlación parcial ya mencionados y la suma de los cuadrados de las $\Delta_{4.3}$ Pero como la suma de los cuadrados de las desviaciones es siempre unitaria, este último factor desaparece, y el cuarto término se reduce a

$$\text{MAS } r_{1(3)4}\, r_{2(3)4}$$

Al sustituir en la expresión que hemos logrado desarrollar anteriormente, el numerador de la expresión que nos sirve de punto de partida se convierte en:

$$r_{1(3)2} - 2\, r_{1(3)4} r_{2(3)4} + r_{1(3)4} r_{2(3)4} = r_{1(3)2} - r_{1(3)4} r_{2(3)4}$$

Obtenido este equivalente del numerador, pasaremos a determinar, en forma parecida el equivalente del denominador.

Para el denominador se tiene que, la cantidad subradical equivale al producto de dos expresiones análogas (sumas de cuadrados de das o taus). De ahí que sólo investigaremos el equivalente de uno de los factores del subradical.

$$\Sigma\tau_{2.34}^{2} = \Sigma(\Delta_{2.3} - \hat{\Delta}_{2.3:4.3})^{2} =$$

$$= \Sigma\Delta_{2.3}^{2} - \Sigma\Delta_{2.3}\,\hat{\Delta}_{2.3:4.3} + \Sigma\hat{\Delta}_{2.3:4.3}^{2}$$

Los equivalentes de los términos de este desarrollo son los siguientes:

1º En el primer término se tiene una suma de cuadrados de desviaciones, que es igual a la unidad; o sea, que el primer término vale

$$1$$

2º En el segundo miembro figura una suma de productos de una delta mayúscula sin circunflejo por una con circunflejo. Esta última puede sustituirse por un índice de correlación parcial multiplicado por una delta sin circunflejo (en el caso, por $r_{2(3)4}\,\Delta_{4.3}$). Al hacer la sustitución se obtiene un triple producto de la r correspondiente y de dos deltas mayúsculas sin circunflejo; estas dos deltas mayúsculas pueden sustituirse por un índice de correlación parcial (en el caso $r_{2(3)4}$). Con ello se obtiene el doble producto de dos índices de correlación parcial que, por ser iguales producen a su vez un cuadrado. Con lo cual, todo el segundo término se reduce a:

$$\text{MENOS } 2\,r_{2(3)4}^{\;2}$$

3º En el tercer miembro figura una suma de cuadrados de deltas mayúsculas con circunflejo. Como la delta mayúscula con circunflejo puede sustituirse por un índice de correlación parcial y una delta mayúscula sin circunflejo, el término será el cuadrado de dicho producto (el cuadrado de $r_{2(3)4}\,\Delta_{4.3}{}^{2}$) bajo el sumador. Pero como la suma de los cuadrados de las deltas mayúsculas sin circunflejo es igual a la unidad, todo el término se reduce a:

$$\text{MAS } r_{2(3)4}^{\;2}$$

De este modo, uno de los factores del subradical del denominador en la expresión que sirve de punto de partida se reduce a:

$$1 - r_{2(3)4}^{\;2}$$

Análogamente, puede mostrarse que el otro factor de dicho subradical se reduce a:

$$1 - r_{1(3)4}^{\;2}$$

Consiguientemente, todo el denominador equivale a

$$\left[(1 - r_{2(3)4}^2) \; (1 - r_{1(3)4})^2 \right]^{\frac{1}{2}}$$

La sustitución del equivalente del numerador y del equivalente del denominador en la expresión que nos sirve de punto de partida, permite establecer como fórmula de las relaciones entre una de las correlaciones netas de la distribución tetravariada y las correspondientes la siguiente:

$$r_{1(34)2} = \frac{r_{1(3)2} - r_{1(3)4}\, r_{2(3)4}}{(1 - r_{2(3)4}^2) \; (1 - r_{1(3)4}^2) \quad \frac{1}{2}}$$

En forma más simple, recordando las equivalencias con los errores de estimación se puede hacer figurar en el denominador el producto $\varepsilon_{1(3)4} \times \varepsilon_{2(3)4}$

FORMULA GENERAL DE LA CORRELACION NETA

El índice de correlación neta entre dos variables ("dependiente" e "independiente") con eliminación de las n restantes ("variables por eliminar") es igual a un cociente entre:

1.—La diferencia que se obtiene al restar de:

A.—El índice de correlación neta entre la variable dependiente con eliminación de las n-1 anteriores a la última de las variables por eliminar, y

B.—El producto formado por:

a.—El índice de correlación neta entre la variable dependiente y la última de las variables por eliminar, con eliminación de la influencia de las n-1 restantes por eliminar, y b.—El índice de correlación neta entre la variable independiente y la enésima de las variables por eliminar, con eliminación de la influencia de todas las variables restantes por eliminar, y

2.—El producto formado por:

A.—El error de estimación correspondiente al primero de los índices de correlación neta que figuran en el substraendo del numerador y

B.—El error de estimación que corresponde al segundo de dichos índices.

En la fórmula siguiente identificaremos con los siguiente índices las diferentes variables que intervienen en una correlación neta:

Subíndice d.—variable dependiente,
Subíndice i.—variable independiente,
Subíndices a, b, c, d, e... n-1, n variables por eliminar.
En estas condiciones:

$$r_{d(abcd...n)i} = \frac{r_{d(abcd...n-1)i} - r_{d(abcd...n-1)n}\, r_{i(abcd...n-1)n}}{\varepsilon_{d(abcd...n-1)n_{\sigma}}\ \varepsilon_{i(abcd...n-1)n_{\sigma}}}$$

EJEMPLO DE CALCULO DE UN INDICE DE CORRELACION NETA ENTRE DOS VARIABLES EN UNA DISTRIBUCION TETRAVARIADA

Tomaremos como ejemplo el de la distribución tetravariada de la que hemos calculado ya los índices-coeficientes de correlación simple, los índices de correlación múltiple, los índices de correlación parcial-neta de las distribuciones trivariadas componentes y uno de los índices de correlación parcial por eliminación de una sola de las variables de la distribución conjunta.

La fórmula para el cálculo de la correlación neta es:

$$r_{1(34)2} = \frac{r_{1(3)2} - r_{1(3)4}\, r_{2(3)4}}{\sqrt{(1 - r_{1(3)4}{}^2)\ (1 - r_{2(3)4}{}^2)}}$$

De acuerdo con los resultados obtenidos previamente:

$r_{1(3)2} = 0.239$

$r_{1(3)4} = 0.576$

$r_{2(3)4} = 0.201$

Consiguientemente:

$r_{1(3)4}\, r_{2(3)4} = 0.576 \times 0.201 = 0.115776$

$r_{1(3)2} - r_{1(3)4} r_{2(3)4} = 0.239 - 0.115776 = 0.123224$ Numerador.

$1 - r_{1(3)4}{}^2 = 1 - 0.576^2 = 1 - 0.331776 = 0.668224$

$1 - r_{2(3)4}{}^2 = 1 - 0.201^2 = 1 - 0.040401 = 0.959599$

Producto = 0.641227082176

Raíz cuadrada = 0.800766 Denominador.

$$r_{1(34)2} = \frac{0.123224}{0.800766} = 0.1539 = 15.39\%$$

Intento de Interpretación.—De acuerdo con la identificación que se hizo de las variables:

1.—Indice económico.
2.—Indice demográfico.
3.—Indice laboral.
4.—Indice educativo.

Estos términos no se emplean en sentido estadístico estricto. En este contexto empleamos "índice" tan sólo en el sentido de "indicador". La razón de esta salvedad puede verse en relación con la existencia de "correlaciones espurias" cuando se trabaja con "índices" tomados en su estricto sentido estadístico.

$r_{1(34)2}$ indica la correlación neta entre la economía y la población cuando se han eliminado las influencias que sobre una de estas variables ejercen tanto el trabajo como la educación. En un cierto sentido, el índice es el mismo tipo que $r_{1(3)2}$ y que r_{12} así como del mismo que otro semejante (que no hemos calculado por el momento $r_{1(4)2}$). Si se registran sucesivamente los valores de estos diversos índices se tiene:

$$r_{12} = 0.465$$
$$r_{1(3)2} = 0.239$$
$$r_{1(3422} = 0.154$$

Al través de este simple registro y de una sencilla comparación se muestra cómo disminuye la dependencia de la economía con respecto a la población en cuanto se eliminan las influencias que sobre ambas ejerce, en primer término, el trabajo y, en segundo término y en forma conjunta, tanto el trabajo como la educación.

Si comparamos, por cociente, el índice de correlación neta entre la economía y la población con eliminación de las influencias de trabajo y educación con el índice-coeficiente de correlación simple entre esas dos variables se verá que

$$\frac{r_{1(34)2}}{r_{12}} = \frac{0.154}{0.465} = .33$$

O sea que la relación neta entre economía y población es apenas el 33 por ciento de la relación bruta entre esas mismas variables.

V.—CORRELACION EN SERIES DE FRECUENCIAS

CORRELACION EN SERIES DE FRECUENCIAS

Series de Frecuencias.—Una serie bivariada de frecuencias es aquella que se obtiene al representar a cada individuo de un conjunto, población o universo estadístico por un trío de cifras. Este trío de cifras lo constituyen:

1.—El valor de la primer variable: x_1
2.—el valor de la segunda variable: x_2
3.—la frecuencia conjunta del par de valores constituido por las dos cifras anteriores: f_{12}

Al hablar de las series sencillas, tomábamos como ejemplo dos fenómenos A y B cuyo registro de valores producía la siguiente distribución o serie sencilla bivariada:

	A	B
p	12	5
q	12	3
r	6	3
s	6	5
t	12	3
u	12	3
v	6	5
w	12	3

En este caso, puesto que se trata de una serie sencilla bivariada, sólo tenemos un par de valores: el valor de la variable x_1 que está dado por las cifras de la columna A, y el valor de la variable x_2 que está dado por las cifras de la columna B.

En la serie sencilla bivariada anterior no se han evitado las repeticiones de unos mismos pares de valores (hay tres pares 12, 3, hay dos pares 6, 5, etc.) y por ello se dice que la serie es sencilla.

Más adelante, evitamos las repeticiones consignando frente a cada par distinto de valores el número de veces que aparecía dicho par (no cada valor por separado) y así obtuvimos la distribución:

A	B	frecuencia
12	5	1
12	3	4
6	5	2
6	3	1

Esta pequeña tabulación indica que dentro del conjunto hubo un solo caso en el que mientras el valor de la primer variable era 12 el segundo era 5; que hubo 4 casos en los cuales mientras el valor de la primer variable era 12 el de la segunda era 3; que hubo 2 en los que en tanto el valor de la primera era 6 el de la segunda era 5, y que, finalmente hubo uno en el que mientras el de la primera era 6 el de la segunda era 3. A los valores 1, 4, 2, 1 les damos el nombre de "frecuencias conjuntas" de la distribución bivariada de frecuencias, las representamos por r_{12}, y las designamos así para distinguirlas de las frecuencias simples de cada una de las distribuciones univariadas componentes. En efecto, si tomamos separadamente el fenómeno A, nuestra lista nos da:

	A
p	12
q	12
r	6
s	6
t	12
u	12
v	6
w	12

como serie sencilla univariada. Si esta serie univariada sencilla queremos convertirla en una serie univariada de frecuencias, evitaremos las repeticiones asociando una cifra que indique cuántas veces aparece cada valor. Así se obtiene:

A	frecuencia no conjunta
12	5
6	3

O sea que en el conjunto e independientemente de los valores que asuma B, existen 5 casos en que A vale 12 y 33 casos en los que A vale 3. En forma parecida se pueden determinar en el mismo ejemplo frecuencias no conjuntas para B, que serían:

B	frecuencia no conjunta
5	3
3	5

puesto que, en efecto, independientemente de los valores que adquiere A, existen 3 casos en los cuales B vale 5 y cinco casos en los que B vale 3.

De acuerdo con esto, para no confundir los diferentes tipos de frecuencia que pueden intervenir en una distribución bivariada, representaremos las frecuencias conjuntas de la distribución por la literal f afectada de un subíndice doble f_{12}, y las frecuencias no conjuntas de cada una de las distribuciones univariadas componentes por la misma lateral f afectada de subíndices sencillos: f_1 para las frecuencias del primer fenómeno: f_2 para las frecuencias del segundo fenómeno.

El Cuadro de Doble Entrada.—Una distribución bivariada de frecuencias se tabula y manipula más fácilmente mediante la constitución de un cuadro de doble entrada.

Un cuadro de doble entrada usa:

1.—Los encabezados o casillas capitulares de las columnas para la serie univariada de los valores de la primer variable: x_1

2.—Los rubros o casillas capitulares de las filas para la serie univariada de los valores de la segunda variable: x_2

3.—Las casillas del cuerpo de la tabla para las frecuencias conjuntas de los pares de valores constituidos por la cabeza y el rubro de la casilla correspondiente.

De este modo, el ejemplo muy sencillo que hemos tomado como punto de partido quedaría consignado como sigue en un cuadro de doble entrada:

x_2 \ x_1	12	6
5	1	2
3	4	1

La lectura de este cuadro nos indica que hay un caso en el que siendo el valor de la primer variable 12 (encabezado de la primer columna) el valor de la segunda es 5 (rubro del primer renglón); que hay 2 casos (frecuencia conjunta del extremo del renglón superior) en que siendo 6 el valor de la primer variable (segundo encabezado) es 5 el de la segunda (rubro del primer renglón); que hay 4 casos en que siendo 12 el valor de la primera es 3 el de la segunda, etc.

Pueden obtenerse, a partir del cuadro de doble entrada las frecuencias no conjuntas de las distribuciones univariadas componentes. Para ello:

1.—Si se suman las frecuencias conjuntas de cada renglón se obtienen las frecuencias no conjuntas de la segunda variable

$x_2 \backslash x_1$	12	6	f_2
5	1	2	$1+2=3$
3	4	1	$4+1=5$

2.—Si se suman las frecuencias conjuntas de cada columna, al pie, se obtienen las frecuencias no conjuntas de la primer variable

$x_2 \backslash x_1$	12	6	f_2
5	1	2	3
3	4	1	5
f_1	$1+4=5$	$2+1=3$	

3.—El efectivo de la distribución queda dado por el total de las frecuencias conjuntas Σf_{12}, y también puede obtenerse

a.—sumando al pie de la columna correspondiente todas las frecuencias no conjuntas de la segunda variable Σf_2

b.—sumando al extremo del renglón correspondiente todas las frecuencias no conjuntas de la segunda variable Σf_1

$$\Sigma f_{12} = \Sigma f_1 = \Sigma f_2$$

De ahí que, el sitio más apropiado, dentro del cuadro de doble entrada, para consignar el efectivo de la distribución sea el cruce de la columna destinada a las frecuencias no conjuntas de la segunda variable y de la hilera destinada a las frecuencias no conjuntas de la primer variable.

De ahí también que el cuadro quede totalmente constituido y listo para manipularse, en la forma siguiente:

$x_2 \backslash x_1$	12		6		f_2
5	1		2		3 +
3	4		1		5
f_1	5	+	3	=	$8 = \Sigma f_{12} = \Sigma_2 = \Sigma f_1$

Cálculo de las estadísticas de las distribuciones univariadas en un cuadro de doble entrada.—Antes de hacer el cálculo de la correlación en una distribución bivariada de frecuencias, conviene mostrar cuál es la forma en que se calculan las medidas características de las distribuciones univariadas correspondientes que se requerirán para dicho cálculo. El cálculo de la media aritmética de la primer variable ($\bar{x}_1$) y de la media aritmética de la segunda variable ($\bar{x}_2$) son suficientemente demostrativas.

Según ocurre en general cuando se calcula la media aritmética de una serie de frecuencias (en este caso la serie x_1, f_1 por una parte y la serie x_2, f_2 por otra) en la fórmula de la media aritmética deben intervenir las frecuencias correspondientes como factores de ponderación. De este modo, las fórmulas que habrá que emplear son:

$$\bar{x}_1 = \frac{\Sigma x_1 f_1}{\Sigma f_1}$$

$$\bar{x}_2 = \frac{\Sigma x_2 f_2}{\Sigma f_2}$$

En el primer caso, comenzaremos por obtener los productos de cada x_1 por su f_1 correspondiente. En la tabla de doble entrada, habrá que multiplicar cada encabezado por cada uno de los valores contenidos en la correspondiente casilla de la hilera cuyo rubro es f_1, anotando el producto correspondiente en la casilla de la hilera inmediatamente siguiente que tendrá como rubro $x_1 f_1$. Al final de esta fila, mediante la suma de todas sus casillos, se obtendrá $\Sigma x_1 f_1$. El cuadro de doble entrada presentará el aspecto siguiente:

$x_2 \backslash x_1$	12	6	f_2
5	1	2	3
3	4	1	5
f_1	5	3	8
$x_1 f_1$	$12 \times 5 = 60$	$6 \times 3 = 18$	$60 + 18 = 78 = \Sigma x_1 f_1$

En el segundo caso, para obtener los productos de las x_2 por sus frecuencias correspondientes f_2, multiplicaremos cada rubro por el valor que aparece en la misma hilera en la columna de las f_2, sumaremos los productos al pie de la columna que habremos encabezado $x_2 f_2$ y así obtendremos:

$x_2 \backslash x_1$	12	6	f_2	$x_2 f_2$
5	1	2	3	$5 \times 3 = 15$
3	4	1	5	$3 \times 5 = 15$
f_1	5	3	8	$15 + 15 = 30$
$x_1 f_1$	60	18	78	

En forma parecida, para obtener medidas análogas, se procederá como si la columna inicial y la columna final (x_2 y f_2) constituyeran una serie sencilla univariada y no existiera el cuerpo del cuadro que contiene las frecuencias conjuntas, y, en forma semejante, como si la hilera inicial y la hilera final (x_1 y f_1) constituyeran por su parte la otra serie univariada, con prescindencia total del resto del cuadro. En estas condiciones, siempre que en la teoría general de las distribuciones univariadas (ver *Técnicas Estadísticas)* se dice: "ábrase una columna y consígnese en ella... (los productos, las potencias, las desviaciones, etc.)", la expresión debe entenderse al pie de la letra para la distribución de x_1, f_1, y debe transformarse en "ábrase una hilera y consígnense en ella..." en tratándose de la distribución x_2, f_2.

Como puede comprenderse, asimismo, el procedimiento de la correlación rectilínea de primer orden no sufre modificación considerable por el hecho de tratarse de una distribución bivariada de frecuencias y no de una distribución sencilla bivariada, a no ser que se trate:

1.—Del efectivo de la distribución que, ahora, no estará dado por N, sino por la suma de las frecuencias conjuntas Σf_{12} o, en ciertos casos, por las sumas de las frecuencias no conjuntas de las distribuciones univariadas componentes (que como sabemos son iguales entre sí y con la suma de las frecuencias conjuntas).

2.—De productos en los que intervengan las x_1 y las x_2 o valores derivados de ellas como las d_1 y las d_2 (o desviaciones de cada una de esas variables respecto a sus medidas aritméticas) o como las δ_1 y las δ_2 (o desviaciones sigmáticas de ambas variables), pues en estos casos se requiere de una ponderación particular que pasamos a exponer.

Ponderación de los productos de las dos variables de una distribución bivariada o de valores derivados de dichas variables en un cuadro de doble entrada.—Ponderar, en estadística significa, básicamente, multiplicar por ciertos valores. En el caso de las series de frecuencias, la ponderación más frecuente empleada es la que utiliza como factor de ponderación a las frecuencias. Pero, en el caso de una distribución bivariada hay tres clases de frecuencias: frecuencias de la primer variable, frecuencias de la segunda variable y frecuencias conjuntas de la primera y de la segunda variable. En tales condiciones, ¿cuál de esas tres frecuencias será el factor ponderal? Es fácil pensar que el factor ponderal deberá estar constituido por las frecuencias conjuntas en tratándose de productos de las dos variables y no por cada frecuencia aislada como factor de cada variable.

Supongamos que tenemos que obtener la suma ponderada de los productos de las dos variables x_1 y x_2 consignadas en el cuadro de doble entrada con que venimos trabajando. El proceso que necesitaríamos seguir sería el siguiente:

Primer valor de x_1 (12) por primer valor de x_2 (5) $= 60$ por número de veces que aparece el par (12, 5) $= 60 \times 1 = 60$
Segundo valor de x_1 (6) por primer valor de x_2 (5) por frecuencia del par (2) $= 60$
Primer valor de x_1 (12) por segundo valor de x_2 (3) por frecuencia de par (4) $= 144$
Segundo valor de x_1 (6) por segundo valor de x_2 (3) por frecuencia del par (3) $= 54$

Suma ponderada de los productos: $60 + 60 + 144 + 54 = 318$

Este proceso puede mecanizarse como se indica en seguida:

Mecanizaciones del proceso de ponderación de productos en los que intervienen las dos variables en cuadros de doble entrada.— Vamos a tomar el cuadro del doble entrada en su forma más simple (es decir, prescindiendo de las columnas de frecuencias no conjuntas que hemos consignado después de formado el cuadro). A partir del cuadro:

	x_1	12	6
x_2			
5		1	2
3		4	1

1.—Duplicaremos el cuerpo del cuadro mediante la apertura de dos columnas adicionales (o, en general, mediante la apertura de tantas columnas adicionales como haya en el cuadro originario), cuyos encabezados serán los mismos de las columnas originales (12 y 6 en este caso) y que se extenderán por tantas hileras como se extiende el cuadro originario (en este caso, por dos hileras, cuyos rubros siguen siendo 5 y 3:

x_2 \ x_1	12	6	/	12	6
5	1	2	/		
3	4	1	/		

En el cuadro aparece una separación entre el cuadro y su duplicación porque entre ellos es posible intercalar todas las columnas que requiera un cálculo concreto (columna de las frecuencias univariadas, productos de la segunda variable por sus frecuencias, de la segunda variable respecto a su media aritmética, etc.).

Las casillas de la reduplicación del cuerpo del cuadro se llenan como sigue:

2.—Se toma cada frecuencia conjunta del cuerpo del cuadro (1, por ejemplo) y se multiplica por su encabezado (12 en el ejemplo) y por su rubro (5 en el ejemplo). El producto de estos tres factores se consigna en la casilla correspondiente del duplicado del cuerpo del cuadro. En la mismo forma se procede con todas las frecuencias conjuntas restantes. De este modo el cuadro presenta la siguiente apariencia:

x_2 \ x_1	12	6	/	12	6
5	1	2	/	60	60
3	4	3	/	144	54

3.—La suma ponderada se obtiene sumando al extremo de cada renglón los valores contenidos en el duplicado del cuerpo del cuadro y sumando al pie de la columna correspondiente, cuyo encabezado será $x_1x_2f_{12}$ o bien, sumando al pie de cada columna de la reduplicación del cuerpo del cuadro y sumando al extremo todas estas sumas, siendo en tal caso, el rubro de la hilera el mismo que se señaló anteriormente, a sea $x_1x_2f_{12}$. La parte reduplicada del cuadro se presentará como sigue:

12		6		$x_1x_2f_{12}$
60		60		120
				+
144		54		198
204	+	114	=	318 $= x_1x_2f_{12}$

Como es fácil comprender, también la última hilera puede ser separada del resto del cuadro por las correspondientes hileras interpoladas que den las frecuencias de la primera distribución univariada, sus productos por los de la variable misma, las desviaciones de ésta con respecto a su media aritmética, etc.

Obtención de la ecuación de estimación de una distribución bivariada de frecuencias.—En la forma más sencilla de cálculo de la ecuación de regresión en una correlación rectilínea de primer orden, partimos de la expresión:

$$x_1 = a_0 + a_2 x_2$$

De esta ecuación formamos dos ecuaciones de interpolación aplicando el sumador a todos los términos de la misma, y multiplicando cada término por x_2 y aplicando a los productos el sumador:

$$\Sigma x_1 = Na_0 + a_2\Sigma x_2$$

$$\Sigma x_1x_2 = a_0\Sigma x_2 = a_2\Sigma x_2^2$$

Estas ecuaciones son válidas en tratándose de series sencillas. Cuando se trata de series de frecuencias es necesario ponderar cada una de las sumas que aparecen en las ecuaciones de interpolación. Dicha ponderación se hará como sigue:

1º—Cuando se trate de sumas de variables aisladas, dichas sumas se trasformarán en sumas de productos de la variable correspondiente por su frecuencia.
2º—La ponderación, en el caso de las sumas de cuadrados, tiene que hacerse mediante la frecuencia de la distribución univariada correspondiente.
3º—La ponderación, en el caso de los productos de las variables, deberá hacerse mediante las frecuencias conjuntas.

De este modo, las ecuaciones de interpolación se convierten en:

$$\Sigma x_1 f_1 = \Sigma f_{12} a_0 + a_2 \Sigma x_2 f_2$$
$$\Sigma x_2 x_1 f_{12} = a_0 \Sigma x_2 f_2 + a_2 \Sigma x_2^2 f_2$$

Mediante el procedimiento ordinario de despeje de incógnitas pueden obtenerse los valores de a_0 y a_2 que pueden substituirse en la ecuación general. De este modo, los valores estimados de la primer variable a partir de la segunda variable pueden obtenerse con la ecuación:

$$\hat{x}_{12} = a_0 + a_2 x_2$$

Obtención del error de estimación.—El proceso para obtener el error de estimación, en el caso de una serie bivariada de frecuencias consignada en un tabla o cuadro de doble entrada procede fácilmente mediante la apertura de:

1.—Una columna destinada a las desviaciones de los valores observados de la primer variable x_1 con respecto a los valores estimados a partir de los valores de la segunda variable ($\hat{x}_{12}$) que se encabezará d_{12}.
2.—Una columna para los cuadrados de dichas desviaciones.

3.—La suma al pie de los valores consignados en esa última columna.
4.—Su división por el efectivo y
5.—La extracción de la raíz cuadrada del cociente.

Obtención del índice de correlación.—Como en todos los casos anteriores, el índice de correlación no es sino la lectura directa de lo que el error de estimación proporciona como lectura inversa. De ahí que su cálculo no represente dificultad adicional.

VI—CORRELACION RECTILINEA CON DATOS CODIFICADOS

CORRELACION RECTILINEA CON DATOS SUJETOS A CODIFICACION

Correlación rectilínea de primer orden obtenida a partir de desviaciones con respecto a una media arbitraria.—Si de cada uno de los valores de una serie x_1 se resta una cantidad constante conveniente, se obtienen valores más sencillos y al proceso se le conoce como "codificación". Frecuentemente la constante elegida es uno de los valores de' la serie misma, por lo que se le suele representar por x'. Asimismo suele convenir que este valor esté tan próximo como sea posible con respecto a una de los promedios de la serie, pues conforme más se aproxima a la media aritmética, por ejemplo, los valores resultantes de restarlo de cada uno de los valores originales resultan más pequeños y, consiguientemente más fáciles de manipular aritméticamente. En este sentido se piensa generalmente en este valor arbitrario como si se tratase de una media, más que determinada mediante el cálculo, propiamente adivinada y, por ello se designa como "media arbitraria o media adivinada". A los resultados obtenidos de restar de cada valor original de la serie (x_1) la media arbitraria (x') se le designa como "desviación" (puesto que, en efecto, es una diferencia estadística y a toda diferencia estadística así se le designa) "con respecto a la media arbitraria". La desviación con respecto a la media arbitraria se representa por d' y, de acuerdo con lo anterior, su fórmula resulta ser:

$$d' = x - x'$$

De acuerdo con la manera en que hemos venido haciendo esta presentación elementalísima de la correlación, el punto de partida de los cálculos lo hemos tenido siempre en la ecuación de regresión. Cuando los datos están representados por sus valores originarios, dicha ecuación es:

$$x_1 = a_0 + a_2 x_2$$

Pero, tanto los valores de la serie x_1 como los valores de la serie x_2 son susceptibles de codificación; o sea, que tanto unos como otros son susceptibles de expresarse como desviaciones con respecto a una media arbitraria. Es decir, que las x_1 pueden trasformarse en d_1' si se les resta x_1' o sea una de las propias x_1 que se suponga que está más o menos próxima de $\bar{x}_1$, la media aritmética. En forma análoga, las x_2 se trasformarán en d_2' si de cada una de ellas se resta una x_2' o valor que se considere cercano de $\bar{x}_2$. Mediante esta trasformación, tendremos una nueva ecuación de regresión, cuyos parámetros tenemos qus suponer, incialmente, diferentes de los de la ecuación previa, por lo que los designaremos por a_0' y por a_2'. En estas condiciones, la ecuación de regresión que podremos obtener a partir de las desviaciones con respecto a las medias arbitrarias de las dos series univariadas constituyentes de la bivariada cuya correlación se estudia, será:

$$d_1' = a_0' + a_2' d_2'$$

¿Qué alteraciones se introducen en el campo de la correlación al tomar desviaciones con respecto a una media arbitraria para cada serie univariada, en vez de tomar los valores originarios? Para estimar dichas alteraciones sustituiremos en la ecuación anterior las d_1' y d_2' por sus equivalentes en términos de la fórmula dada al principio para este tipo de desviaciones, con lo cual tendremos:

$$x_1 - x_1' = a_0' + a_2' (x_2 - x_2')$$

Si se ejecutan las operaciones de dentro del paréntesis, y se pasa x_1' que figura con signo negativo en el primer miembro al segundo con signo positivo, se obtendrá:

$$x_1 = x_1' + a_0' + a_2' x_2 - a_2' x_2'$$

Si se unen todas las constantes en un solo término y se dejan fuera las variables, se tiene:

$$x_1 = (a_0' + x_1' - a_2'x_2') + a_2'x_2$$

Esta ecuación está dada nuevamente en términos de los datos originarios y no de las desviaciones con respecto a las medias arbitrarias. Tiene exactamente la misma estructura que la ecuación de estimación de que se partió, y gracias a esta igualdad de estructuras es posible comparar:

1.—El primer miembro de la primera (x_1) con el primer miembro de la segunda (x_1) para encontrar que son idénticos:

$$x_1 = x_1$$

2.—El primer término del segundo miembro de la primera (a_0) y el primer término del segundo miembro de la segunda ($a_0' + x' - a_2'x_2'$). Gracias a ello, podemos establecer que:

$$a_0 = a_0' + x' - a_2'x_2'$$

O sea, que, para obtener el primer parámetro de la ecuación de estimación tal y como se habría obtenido a partir de datos originarios, es necesario sumar al primer parámetro obtenido a partir de desviaciones con respecto a la media arbitraria, la diferencia entre la media arbitraria de la primera serie y la media arbitraria de la segunda serie por el segundo parámetro obtenido de desviaciones.

3.—El segundo término del segundo miembro de la primera (a_2x_2) con el segundo término del segundo miembro de la segunda ($a_2'x_2$), lo cual permite establecer que

$$a_2 = a_2'$$

Esto equivale a afirmar que el segundo parámetro no se altera, sea que se trabaje con datos originarios, o sea que se utilicen desviaciones con respecto a las medias arbitrarias de las dos series.

Según todo lo anterior, obtenida la ecuación de estimación a partir de los datos codificados, es fácil obtener la ecuación de estimación que da los valores originarios de la variable dependiente (x_1) a partir de los valores asimismo originarios de la independiente (x_2) si al primer parámetro se le agrega $x_1' - a_2' x_2'$ y se recuerda que a_2 permanece inalterada.

Al hacer estas trasformaciones, ¿qué trasformación se determina en el error de estimación? De acuerdo con lo que sabemos, el error de estimación no es sino una media cuadrática de las desviaciones de los valores observados y de los valores calculados de la variable dependiente a partir de la independiente. En tratándose de valores dados en términos de desviaciones respecto a las medias arbitrarias, el error de estimación será la media cuadrática de las desviaciones:

$$d_1' - \hat{d}_{12}'$$

En esa fórmula, el substraendo es la desviación de la primer variable estimada(^) a partir de la desviación de la segunda variable con respecto a su media arbitraria. O sea que:

$$\hat{d}_{12}' = a_0' - a_2' d_2'$$

Sustituido este valor en la fórmula previa, se tiene:

$$d_1' - \hat{d}_{12}' = d_1' - a_0' + a_2' d_2'$$

Como $d_1' = x_1 - x_1'$ y $d_2' = x_2 - x_2'$:

$$= x_1 - x_1' - a_0' + a_2' x_2' - a_2' x_2'$$

Pero $a_0' = -x_1 + a_0 + a_2 x_2$

De acuerdo con esto:

$$= x_1 - x_1' - (- x_1' + a_0 + a_2x_2') + a_2' x_2 - a_2' x_2' =$$

$$= x_1 - x_1' + x_1' - a_0 - a_2x_2' + a_2' x_2 - a_2' x_2' =$$

Al hacer las reducciones pertinentes (recordando que $a_2' = a_2$) se tiene:

$$= x_1 - a_0 - a_2x_2 = x_1 - (a_0 + a_2x_2) =$$

Pero, si se recuerda la ecuación de estimación:

$$x_1 = a_0 + a_2x_2$$

O, más precisamente:

$$\hat{x}_{12} = a_0 + a_2x_2$$

La anterior cadena de igualdades nos conduce a:

$$= x_1 - \hat{x}_{12}$$

El primer miembro de esta cadena de igualdades (que evitamos repetir a todo lo largo del desarrollo por economía tipográfica) era $d' - d_{12}'$. De lo cual resulta que:

$$d_1' - \hat{d}_{12}' = x_1 - \hat{x}_{12}$$

O sea, que las desviaciones de los valores observados de las desviaciones con respecto a la media arbitraria de la primera serie y los de las desviaciones de esa serie con respecto a esa media estimadas a partir de las desviaciones respecto a la media arbitraria de los valores de la segunda serie, es igual a las desviaciones entre los valores originarios observados y estimados de la primer variable.

Si esto es así, las medias cuadráticas de unas y otras desviaciones serán iguales. Y como las medias cuadráticas de este tipo de desviaciones es lo que se ha designado como error de estimación, podemos afirmar que

$$\varepsilon_{12}' = \varepsilon_{12}$$

O sea, que el error de estimación no se altera, sea que se trabaje con los datos originarios, o sea que se trabaje con desviaciones de los valores de cada serie con respecto a una media arbitraria correspondiente a cada serie (medias que, naturalmente no tienen por qué ser iguales en ambas series).

¿Qué alteraciones sufren los índices de correlación al utilizar valores codificados en vez de valores originarios? Como es fácil prever, si se tiene en cuenta que el índice de correlación no es sino la lectura directa de lo que el error de estimación mide en forma inversa, si los errores de estimación son iguales, sus complementos aritméticos (o sean los índices de correlación) tendrán que ser asimismo iguales:

$$r_{12}' = r_{12}$$

La conclusión anterior es válida tanto para los errores de estimación dados en unidades originales como para los dados en unidades sigmáticas y para los índices de correlación consiguientemente puesto que los errores sigmáticos proceden de los no sigmáticos divididos entre las desviaciones cuadráticas medias de las series correspondientes y, conforme se sabe (ver *Técnicas Estadísticas*) la desviación media cuadrática de una serie no se altera si a todos los valores de la serie se les resta una constante (en este caso la media arbitraria x_1').

VII.—CORRELACION CURVILINEA

CORRELACION CURVILINEA

Introducción.—Para hacer el estudio de la correlación curvilínea en sus términos más simples, conviene comenzar por establecer cuál es la diferencia conceptual básica existente entre la correlación rectilínea por una parte, y la correlación curvilínea, por otra.

Hemos dicho que hay correlación rectilínea entre dos fenómenos A y B cuando a incrementos (o decrementos) constantes del fenómeno A, corresponden aumentos (o disminuciones) constantes del fenómeno B. Esta concepción de la correlación rectilínea sigue siendo válida sea que a aumentos de A correspondan aumentos en el valor de B, que a aumentos de A correspondan disminuciones del valor de B o que ocurra lo contrario siempre y cuando siendo constantes los incrementos o decrementos de A sean asimismo constantes los incrementos o los decrementos de B. Como es fácil comprender, en casos muy singulares puede darse el que el incremento de A iguale al incremento de B, pero, en general, no es preciso que esto ocurra para que la correlación sea rectilínea. Basta con que a cada aumento a, constante, del fenómeno A, corresponda un incremento (o decremento) b, asimismo constante, del fenómeno B, para que la correlación sea rectilínea. En otras palabras, sea que el incremento de A se produzca en la zona de valores bajos del fenómeno observado, sea que se produzca en la zona de valores medios de ese mismo fenómeno, o sea que se produzca en la zona de valores altos, el incremento del fenómeno B, será siempre el mismo.

Nada de lo anterior ocurre en el caso de la correlación curvilínea entre los fenómenos C y D. En este caso, a incrementos (o decrementos) constantes del fenómeno C, corresponden incrementos (o decrementos) variables del fenómeno D, e inversamente, los incrementos (o decrementos) constantes del fenómeno D son producidos por incrementos (o decrementos) variables del fenómeno C. Esto quiere decir, en otros términos, que la variación que se produzca en D será distinta —para una misma variación de C— de acuerdo con la zona en que se haya producido la variación de C. Así, por ejemplo, en ciertas correlaciones curvilíneas, los aumentos de D pueden ser mayores en las zonas de valores bajos y en las de valores altos de C, mientras son menores los incrementos de D cuando las variaciones se producen en la zona de valores medios de C. En otras correlaciones curvilíneas puede ocurrir a la inversa; en otras más, puede suceder que las variaciones de D sean más considerables en las zonas de valores bajos, en tanto que son menos considerables en las zonas de valores altos de C, o que las cosas ocurran a la inversa —en el caso de otros tipos de correlación curvilínea— y esto aun cuando las variaciones de C hayan sido siempre de la misma magnitud. En todos estos casos, sin embargo, se postula la posibilidad de descubrir una ley que ligue las variaciones del fenómeno D con las variaciones del fenómeno C. Naturalmente, dicha ley no tendrá forma rectilínea o se expresará como una ecuación de primer grado (en caso de tratarse de dos variables) y no tendrá en general forma o expresión que corresponda al sistema rectilíneo (en caso de más de dos variables), pero, con todo, será una expresión matemática que, a semejanza de las estudiadas en el caso del sistema rectilíneo, merece el nombre de "ecuación de estimación".

Dificultades de tratamiento de la correlación curvilínea cuando se toma como punto de partida la regresión.—En caso de que quisiéramos seguir en un estrecho paralelismo el tratamiento que aplicamos en nuestras primeras páginas al estudio de la correlación rectilínea, comenzaríamos por registrar la ecuación de regresión o la ecuación de estimación de la correlación curvilínea; o sea, aquella expresión matemática que nos permitiría calcular los valores de la variable dependiente a partir de los de la independiente. Algunas de las formas posibles de ecuaciones de estimación (dentro de su apariencia más sencilla), podrían ser:

$$x_1 = a_0 + a_{12}x_2^2$$
$$x_1 = a_0 + a_{12}x^3$$
$$x_1 = a_0 + a_{12} \operatorname{Log} x$$
$$x_2 = a_0 + a_{12} \operatorname{Antilog} x$$

Conocida la ecuación de regresión, teóricamente, y en estrecho paralelismo con lo que ya hemos hecho, se trataría, fundamentalmente de:

1º—Calcular el error de estimación de las x_1 obtenidas de la ecuación con respecto a las x_1 observadas.
2º—Calcular la desviación cuadrática media de las x_1.
3º—Comparar ambas medidas(mediante división) elevadas al cuadrado y tomar el complemento aritmético del resultado, para obtener el índice de correlación.

Esta aparente facilidad, se enfrenta, en realidad, con, por lo menos, dos órdenes de dificultades: unas, dificultades teóricas; otras, dificultades prácticas.

Dificultad teórica.—Si bien el primer paso fundamental no parece susceptible de crítica, el segundo sí lo es, como veremos en seguida.

Si, para que las referencias sean más concretas, tomamos el caso de:

$$x_1 = a_0 + a_{12} x_2^2$$

Obtener, en el caso, la desviación media cuadrática de las x_1 es, desde luego, posible; pero, esa desviación media cuadrática, ¿puede ser la media cuadrática de las desviaciones con respecto a la *media aritmética* de las x_1? O, dicho de otra manera, ¿es justificable que la piedra de toque del error de estimación sea la media cuadrática de los valores que se obtengan de restar de cada x_1 la media aritmética de la x_1 ($\bar{x}_1$)? En el caso de la correlación rectilínea, la respuesta inmediata hubiera sido *sí;* pero, en este caso, es posible dudar de la conveniencia que puede tener el utilizar la media *aritmética.* En efecto, las x_1 son cantidades cuadráticas. ¿Será, por tanto, representativa de una distribución cuadrática una media *aritmética?* En ninguna forma. Si las x_1 son cuadrados (o dependen de variaciones de cuadrados, como ocurre en el caso en que dependen de las variaciones de x_2^2), el promedio más representativo de la serie no será una media aritmética (propia del sistema rectilíneo o apropiado para la representación de distribuciones normales), sino que tendrá que ser un promedio del sistema cuadrático: una media *cuadrática* directa o inversa. Una media cuadrática directa se obtendría elevando al cuadrado las x_1 sumando los cuadrados, dividiendo la suma entre el efectivo de la distribución y extrayendo la raíz cuadrada del cociente. Una media cuadrática inversa o —mejor— "fraccionaria" se obtendría, extrayendo las raíces cuadradas de las x_1, sumando dichas raíces, dividiendo la suma entre el efectivo y elevando el cociente al cuadrado. Como se ve, este cálculo no tendría dificultad, pues se sujetaría a las fórmulas:

$$\bar{x}_{1q} = \left(\frac{\Sigma x_1^2}{N} \right)^{\frac{1}{2}}$$

$$\bar{x}_{1-q} = \left(\frac{\Sigma x_1^{\frac{1}{2}}}{N} \right)^2$$

Fórmulas en las cuales $\bar{x}_{1q}$ y $\bar{x}_{1-q}$ representan respectivamente la media cuadrática directa y la media cuadrática inversa de las x_1.

Pero, si no hay dificultad por esta parte, sí existe una dificultad mayor en cuanto se trata de cálculo de las "desviaciones" de cada x_1 con respecto a su media cuadrática directa o inversa. Porque no se trata simplemente de restar, como en el caso de las desviaciones con respecto a la media aritmética puesto que, en el nivel cuadrático en que hemos pasado a movernos, es indispensable realizar las operaciones que en tal nivel son análogas de las que se realizan en el nivel lineal. Lo que era simple resta en el nivel lineal no es pura resta en el nivel cuadrático.

Para que se entienda más fácilmente lo que queremos decir, cambiaremos de referencia. No tomaremos en este momento la ecuación cuadrática de estimación $x_1 = a_0 + a_{12} x_2^2$ sino que veremos lo que se necesitaría hacer en caso de que se tratase de una ecuación logarítmica de estimación: $x_1 = a_0 + \text{Log}\, x_2$. En este caso, se necesitaría obtener no una media aritmética como en la correlación rectilínea, ni una media cuadrática como en la correlación curvilínea a que nos acabamos de referir, sino una media del sistema logarítmico (una media *geométrica*) directa o inversa (precisarlo no nos interesa por el momento). Calcular una media geométrica no es difícil, pues se trata de tomar los logaritmos de los datos, sumarlos, dividir la suma entre el efectivo y el cociente considerarlo como un logaritmo (o sea, que finalmente hay que tomar su antilogaritmo), de acuerdo con la fórmula:

$$\bar{x}_{1g} = \text{Antil} \frac{\Sigma \text{Log}\, x_1}{N}$$

Pero, una vez obtenida esta media geométrica, el paso equivalente "al cálculo de las desviaciones con respecto a la media aritmética" de la correlación rectilínea, no es, en el caso de la correlación logarítmica un simple cálculo de "desviaciones", pues la resta del nivel rectilíneo, se convierte en división en el nivel logarítmico. En tal virtud, lo que en el fondo se necesitaría calcular serían "razones" y no desviaciones.

Para el caso de la correlación cuadrática que dejamos hace un momento, tampoco sería una resta pura y simple lo que se impondría para calcular lo que —mediante una extensión de significado— seguiremos llamando genéricamente desviaciones; sino que habría que realizar una operación más complicada que equivaliera en el nivel cuadrático con la resta del nivel lineal (esta operación probablemente tendría que ver con el desarrollo de las potencias de binomio).

Como es fácil apreciar, las dificultades teóricas que representa el tratar de resolver el problema de la correlación no lineal a partir de la teoría de la regresión no son insuperables. Sin embargo, superarlas representa enfrentar dificultades prácticas.

Dificultades prácticas.—Si las dificultades prácticas pueden parecer insignificantes en el caso de las correlaciones logarítmicas (en vista de que, en general se cuenta con un buen sistema de equivalencia de las operaciones entre número naturales y las operaciones entre logaritmos, o entre las operaciones entre ciertos números y las operaciones entre sus antilogaritmos), dichas dificultades prácticas crecen considerablemente en el caso de otras formas de correlación curvilínea como las ya indicadas de la correlación cuadrática (o, en general, como en el caso de las series potenciales), puesto que los pasos que hemos indicado, ya de por sí complicados y confusos en una primera aproximación crecen en complicación en cuanto hay que considerar que no sólo hay que calcular el equivalente de las "desviaciones" de la correlación rectilínea, sino que, ulteriormente, hay que promediarlas para obtener el equivalente de la "desviación cuadrática media". Y esto implica realizar operaciones parecidas a las sumas del nivel rectilíneo, a hacer operaciones análogas, pero no idénticas a las divisiones de dicho nivel y, en forma extrema, impone realizar operaciones que son al nivel cuadrático lo que las radicaciones son al nivel lineal.

Como es fácil comprender, a más de necesitarse, como cosa previa, el establecimiento de las equivalencias operacionales entre los diversos niveles lineales, cuadráticos, cúbicos, cuadrático-cúbicos, logarítmicos, antilogarítmicos, etc., se requeriría de la realización de una serie de operaciones que, en la mayoría de los casos resultaría demasiado bromosas para la cantidad de información que podrían proporcionar. Esos resultados, por otra parte, parece factible obtenerlos con menor gasto de energía y de tiempo y sin sacrificio del rigor científico, ya sea recurriendo a las soluciones tradicionales o ya sea tratando de diseñar otras nuevas (en caso de que estas pudieran resultar menos empíricas).

SOLUCION TRADICIONAL DE LAS DIFICULTADES IMPLICITAS EN LA CORRELACION CURVILINEA

LA RAZON DE CORRELACION

Para resolver los problemas que plantea la correlación curvilínea se recurre tradicionalmente a la "razón de correlación", medida análoga, pero que no hay que confundir con el "índice de correlación". Dicha solución, si bien se mantiene en buena parte dentro de las líneas generales de la teoría de la correlación en general, en cierto modo se aparta de ellas y de la vecindad de los principios se aproxima a una cierta vecindad con respecto a los procedimientos más empíricos.

A fin de tratar de encontrar los vínculos que unen a la razón de correlación con el índice de correlación, subsumiendo a ambos en una misma teoría, volveremos a tomar el caso de la correlación rectilínea y trataremos de resolver el problema correspondiente a lo largo de líneas que puedan conducirnos a una forma de solución válida asimismo en el caso de la correlación curvilínea (y que, por otra parte, no se enfrente a las dificultades encontradas en el estudio hecho con base en la regresión).

El índice de correlación obtenido mediante la formación de arreglos en un cuadro de doble entrada.—Tomaremos como punto de partida, una distribución bivariada tal y como la misma aparece consignada en un cuadro de doble entrada. Llamaremos genéricamente x_{1i} a la primer variable (que será la que consideraremos como dependiente) y x_{2j} a la segunda variable (a la que, en forma correspondiente, consideraremos como variable independiente). Los valores que pueda tomar x_{1i} servirán de encabezados a las columnas del cuadro, de modo que x_{11} representará el valor específico de x_{1i} correspondiente a la primer columna, x_{12} el valor específico de x_{1i} para la segunda columna, x_{1n} el valor específico de x_{1i} para la enésima columna. Los valores que pueda tomar x_{2j} servirán de rubros para las filas del cuadro, de modo que x_{21} será el valor específico de la x_{2n} de la primera hilera, x_{22} el de la segunda, x_{2j} el de la enésima fila. Dentro del cuadro de doble entrada, figurarán en las diversas casillas las "frecuencias conjuntas" que quedarán identificadas por la literal f, especificada mediante subíndices; así, f_{11} será la frecuencia de la primera columna y la primera fila; f_{12} la frecuencia de la primera columna y la segunda fila; f_{21} la frecuencia de la segunda columna y la primera fila; f_{st} la frecuencia de la columna s y de la fila t.

De acuerdo con esta simbología, ¿qué es un arreglo? Un arreglo es la distribución univariada (y no ya bivariada) que se obtiene asociando los distintos valores de una de las variables de la distribución con las frecuencias correspondientes a un solo valor específico de la otra variable. O sea que, de acuerdo con la simbólica anterior, un primer arreglo estará constituido por todas las x_{1i} asociadas a las correspondientes f_{c1} contenidas en la hilera de la x_{21}. Es decir, el arreglo estará constituido por la siguiente distribución univariada:

x_{1i}	f_{c1}
x_{11}	f_{11}
x_{12}	f_{21}
x_{13}	f_{31}
....	....
x_{1n}	f_{n1}

Un segundo arreglo estará constituido por todas las x_{1i} asociadas a las frecuencias f_{c2} contenidos en la hilera de la x_{22}. Es decir, el arreglo estará formado por la siguiente distribución univariada:

x_{11}	f_{c2}
x_{11}	f_{12}
x_{12}	f_{22}
x_{13}	f_{32}
....	....
x_{1n}	f_{n2}

En general, el emésimo arreglo estará constituido por todas las x_{1i} asociadas a las frecuencias f_{cm} contenidas en la hilera de la x_{m2} Es decir, el arreglo estará formado por la distribución univariada:

x_{1i}	f_{cm}
x_{11}	f_{1m}
x_{12}	f_{2m}
x_{13}	f_{3m}
....	....
x_{1n}	f_{nm}

Los arreglos también pueden formarse no sólo a lo largo, sino a lo ancho de la distribución. La serie de arreglos resultante, será distinta, pero se tratará siempre de la asociación de los distintos valores de una variable con la frecuencias correspondientes a un valor específico de la otra variable.

Así, por ejemplo, en el caso, el primer arreglo estará constituido por todas las x_{2j} asociadas con las frecuencias correspondientes a la x_{11}, o sean las f_{1h}:

x_{2j}	f_{1h}
x_{21}	f_{11}
x_{22}	f_{12}
x_{23}	f_{13}
....	....
x_{2n}	f_{1n}

El arreglo de orden p estará constituido por todas las x_{2j} correspondientes a la x_{1p}

x_{2j}	f_{ph}
x_{21}	f_{p1}
x_{22}	f_{p2}
x_{23}	f_{p3}
....	....
x_{2n}	f_{pn}

Idea central del cálculo del índice de correlación.—En páginas anteriores hemos señalado que, en forma central, el cálculo del índice de correlación respondía a la idea de comparar entre sí:

1.—La bondad de las estimaciones de las valores de una serie cuando las mismas se hacían a partir de una fuente interna (la media aritmética de la serie, generalmente) y
2.—La bondad de las estimaciones de los valores de esa misma serie cuando las mismas se hacían a partir de una fuente exterior (la ecuación de regresión con respecto a otra variable).

En efecto, la correlación sirve, fundamentalmente a una tarea inferencial. Se trata de determinar los valores de un fenómeno desconocido o cuyo conocimiento resulta difícil en un momento dado, al través del conocimiento que se tiene de los valores de un fenómeno cuya asociación con el desconocido se conoce, gracias al conocimiento de las variaciones concomitantes de ambos fenómenos, que se han observado anteriormente. Si siempre pudiésemos conocer un fenómeno A por sí mismo, estudiándolo en forma directa, el estudio de la correlación resultaría superfluo. Como esto no es así, el estudio de la correlación nos permite el que conozcamos en forma indirecta (y dentro de ciertos márgenes de error conocidos) el fenómeno de difícil observación al través de otro asociado con él y que es de fácil observación o registro.

Sea como fuere, para los fines estrictamente estadísticos, importa saber que la idea central del cálculo del índice de correlación es una comparación entre la bondad de las estimaciones obtenidas a partir de una fuente interna y de las obtenidas de una fuente exterior. Conforme más se aproximan las estimaciones obtenidas de una fuente exterior a las obtenidas a partir de una fuente interna, la correlación será mayor: las estimaciones serán mejores.

Aplicación de la idea Central de la correlación al diseño del índice de correlación mediante arreglos.—Si tomamos como brújula la idea anteriormente expresada, el diseño del procedimiento de cálculo del índice de correlación tiene que girar en torno del problema de:

1.—Calcular la bondad de la estimación de la variable x_1 a partir de su media e independientemente de la distribución de frecuencias de la variable x_2.

2.—Calcular la bondad de la estimación de la variable x_1 cuando la misma depende de la distribución de frecuencias de la variable x_2.

La primera parte del problema es la más sencilla. Se trata de construir una distribución univariada de las x_1 que no dependa de las x_2 y que, además, contenga toda la información que puede proporcionar la distribución bivariada en su totalidad. Si se tratara simplemente de llenar la primera condición, bastaría con tomar cualquiera de los arreglos correspondientes a una de las x_{1i} para satisfacer dicha condición. Pero, como, además, es indispensable que contenga la información que proporciona no una de las columnas de cuadro de doble entrada, sino todas sus columnas, la solución tiene que ser otra. En efecto, si se suman al final de cada hilera todas las frecuencias contenidas en ella y dichas frecuencias se asocian con el valor de la x_{2j} que sirve de rubro a la hilera, se tendrá una distribución univariada que muestra las variaciones de x_2 independientemente de x_1. En forma parecida (y es esta última solución la que nos interesa), si se suman al pie de cada columna todas las frecuencias contenidas en la columna y se asocian con las x_{1i} correspondientes, se tendrá una distribución univariada de frecuencias, que mostrará las variaciones de x_1 independientemente de las variaciones de x_2. La distribución univariada siguiente (que hemos entresacado del cuadro de doble entrada para mayor claridad) resuelve esta primera parte del problema.

Distribución univariada de las x_{1i}. Independiente de la distribución de las x_{2j} y sus variaciones.

x_{1i}	Σf_{ih}
x_{11}	Σf_{1h}
x_{12}	Σf_{2h}
x_{13}	Σf_{3h}
....	....
x_{1n}	Σf_{nh}

Esta distribución univariada de la variable dependiente (o de la variable que hemos de considerar como "dependiente" dentro de la correlación, pero que es independiente de cualquier otra forma de variación dentro de esta distribución univariada) permitirá estimar la variabilidad interna de la variable x_1. Como el más sencillo de los medios de medir dicha variabilidad es el cálculo de la desviación cuadrática media con respecto a la media aritmética de la serie, será precisamente esta medida la que servirá de piedra de toque o elemento de comparación para la correlación.

La segunda parte del problema trata de evaluar la bondad de las estimaciones de la misma variable x_1 cuando las mismas se hacen considerando las variaciones que se producen en la otra variable. Para esta evaluación, se necesita que existan:

1.—Variaciones de x_1.
2.—Variaciones de x_2.

Pero, además, se necesita, como condición adicional, que

3.—Dichas variaciones sean fácilmente comparables.

Dentro del cuadro de doble entrada pueden observarse las variaciones en las x_1 y en las x_2, pero dichas variaciones no son fáciles de establecer en virtud de la bi-dimensionalidad. De ahí que se recurra al procedimiento de determinar *dentro* de cada arreglo las desviaciones de todos los valores del arreglo con relación al promedio de dicho arreglo (pues en esta forma se determina la variabilidad de una de las variables). Debe observarse que los promedios con respecto a los cuales se determina la variabilidad dentro de cada arreglo, difieren de un arreglo a otro, o sea, *entre* arreglos (con lo cual se tiene como referencia la variablidad de la otra varaible). Obtenidas las diversas desviaciones posibles, una promediación de las mismas que produzcan una medida análoga a la desviación media cuadrática, permitirá obtener el error de estimación que hemos de comparar con la desviación cuadrática media de la primera variable.

Fórmula para el cálculo de la razón de correlación.—Como en las consideraciones anteriores NO hemos introducido ninguna consideración de rectilinealidad, podemos considerar que la medida que obtengamos a partir de las mismas mide la correlación entre las variables, sea que entre éstas exista una vinculación rectilínea o sea que la relación que entre ellas exista no sea rectilínea. Es decir, que nos encontraremos frente a una medida genérica de la correlación que designaremos como "razón de correlación". De acuerdo con las consideraciones previas, tendremos que:

La razón de correlación es una medida que compara, fundamentalmente, la desviación media cuadrática de los valores de una serie en relación con la serie de *las* medi*as* de los arreglos a que pertenecen y la desviación de los valores de dicha serie con respecto a *la* media# de la serie univariada constituida por la variable dependiente y las frecuencias obtenidas de sumar todas las frecuencias de los arreglos perpendiculares a los considerados previamente.

Esta comparación se obtiene dividiendo el cuadrado de la primera de las desviaciones medias cuadráticas mencionadas (que cumple la función de "error de estimación" sólo que, en este caso no puede decirse que lo sea "con respecto a una línea de regresión" propiamente hablando), entre el cuadrado de la segunda de dichas desviaciones medias.

Para traducir la medida obtenida de un lenguaje negativo (en cuanto es, auténticamente, "medida de falta de correlación") a un lenguaje positivo, se toma el complemento aritmético de la misma, restando el cociente de la unidad.

Finalmente, para reducir las unidades obtenidas del sistema cuadrático al que pertenecen (en cuanto se obtuvieron de un cociente de cuadrados) al sistema lineal, se extrae el cuadrado de dicho complemento. En estas condiciones:

$$\text{Razón de correlación} = 1 - \left\{ \frac{\left(\begin{array}{l}\text{Desviación media cuadrá-}\\\text{tica de los valores de cada}\\\text{arreglo con respecto a la}\end{array}\right)^2 \begin{array}{l}\\\\\\\text{media del arreglo}\end{array}}{\left(\begin{array}{l}\text{Desviación media cuadrá-}\\\text{tica variable dependiente}\\\text{con respecto a la media}\end{array}\right)^2 \begin{array}{l}\\\\\\\text{general}\end{array}} \right\}^{\frac{1}{2}}$$

Criterios para la representación simbólica de los cálculos de la razón de correlación.—Hasta ahora hemos venido considerando, y en lo sucesivo consideraremos principalmente el caso de una distribución bivariada o sea, por lo mismo, de una distribución en la que intervienen dos variables, a más de las frecuencias. A fin de que la simbología sea convencional —como toda simbología—, pero tan consistente internamente como sea posible, deben tenerse en consideración los siguientes criterios:

1.—Las variables se distinguen entre sí por los subíndices que, en este caso conviene que sean dobles.

A.—A la primer variable le corresponden subíndices en los que las cifras significativas figuran en la posición de las unidades.

B.—A la segunda variable le corresponden subíndices en los que las cifras significativas figuran en el sitio de las decenas.

2.—Los diversos valores de una misma variable se distinguen entre sí por el valor de las cifras significativas que se encuentran en el subíndice en el sitio característico de la variable correspondiente. Así, por ejemplo:

A.—Si las unidades del subíndice son n esto significa que se se trata del enésimo valor de la primer variable.

B.—Si las decenas del subíndice son m, esto significa que se trata de emésimo valor de la segunda variable.

3.—Las frecuencias conjuntas se diferencian por un subíndice doble, en el cual

A.—Las unidades del subíndice representan el orden de la columna,

B.—Las decenas del subíndice representan el orden de la hilera a que corresponde una frecuencia dada. Así, por ejemplo, f_{13} representa la frecuencia de la primera columna y la tercera fila; f_{31} la frecuencia de la tercera columna y la primera fila.

4.—Los arreglos se designan por la asociación de dos valores:

1º—Los valores de la variable,

2º—Los valores de las frecuencias.

En los arreglos de la primer variable, los subíndices de la variable son unidades (x_{0c} indicando c la columna de la que se trata). En los arreglos de la segunda variable, los subíndices son decenas (x_{h0} indicando h la hilera de la que se trata).

5.—Para saber el orden del arreglo:

A.—Si el arreglo es de la primer variable, véase cuáles son las decenas del subíndice de f. (x_{0c}, f_{nc}) representa el enésimo arreglo de la primera variable.

B.—Si el arreglo es de la segunda variable, véase cuáles son las unidades del subíndice de f. (x_{h0}, f_{hm}) representa el emésimo arreglo de la segunda variable).

6.—Las medias de los arreglos de la primer variable tienen subíndices con cifras significativas en el sitio de las unidades ($\bar{x}_{01}$, $\bar{x}_{02}$ son las medias del primero y del segundo arreglos de la primer variable). Las medias de los arreglos de la segunda variable tienen subíndices con cifras significativas en el sitio de las decenas ($\bar{x}_{10}$, $\bar{x}_{20}$ son las medias del primero y del segundo arreglos de la segunda variable).

7.—A fin de diferenciar dichas desviaciones, representaremos:

A.—con des minúsculas las desviaciones de los valores de la primer variable con respecto a las medias de los arreglos de dichas valores

B.—con Des mayúsculas las desviaciones de los valores de la segunda variable con respecto a las medias de los correspondientes arreglos.

8.—Las desviaciones antes mencionadas tienen subíndices dobles:

A.—Las decenas del subíndice indican el orden de la variable que figura como minuendo.

B.—Las unidades del subíndice, el orden del arreglo cuya media sirve de substraendo.

9.—Las desviaciones medias cuadráticas (sigmas minúsculas) se diferencian como sigue:

Si el subíndice es 12 esto indica que se trata de la desviación media cuadrática de todos los valores de la primer variable con respecto a los arreglos de esa misma variable para valores constantes de la segunda variable.

Si el subíndice es 21, esto indica que se trata de la desviación cuadrática media de la segunda variable (obtenida por resta que de cada uno de sus valores se haga de la media de los arreglos de dicha segunda variable).

Si el subíndice es 0a se trata de la desviación cuadrática media de las medias de los arreglos con respecto a la media aritmética de toda la distribución, en caso de tratarse de los arreglos de la primer variable.

Si el subíndice es a0 se trata de la desviación cuadrática media de las medias de los arreglos de la segunda variable con respecto a la media aritmética de toda la serie.

Estos dos últimos valores no son inmediatamente útiles para simbolizar lo que ya hemos dicho; en cambio son indispensables para una ulterior simplificación, tanto de la representación simbólica como del cálculo de la razón de correlación.

Ejemplificación simbológica.—En lo que sigue, se da la traducción en símbolos de lo que previamente se expresa en palabras, de acuerdo con los criterios anteriores.

Primera variable	x_{0i}	Subíndices unidades.
Segunda variable	x_{j0}	Subíndices decenas.

Primer valor de la primera variable	x_{01}
Segundo valor de la primera variable	x_{02}
Enésimo valor de la primera variable	x_{0n}
Primer valor de la segunda variable	x_{20}
Enésimo valor de la segunda variable	x_{m0}

Primer arreglo de la primera variable (x_{0c}, f_{1c})

$$\begin{array}{cc} x_{01} & f_{11} \\ x_{02} & f_{12} \\ x_{03} & f_{13} \\ \hline x_{0n} & f_{1n} \end{array}$$

Enésimo arreglo de la primera variable (x_{0c}, f_{nc})

Primer arreglo de la segunda variable (x_{h0}, f_{h1})

$$\begin{array}{cc} x_{10} & f_{11} \\ x_{20} & f_{21} \\ x_{30} & f_{31} \\ \hline x_{m0} & f_{m1} \end{array}$$

Media del primer arreglo de la primera variable (media de x_{0c}, f_{1c}).

$$\frac{\Sigma x_{0c} f_{1c}}{\Sigma f_{1c}} = \bar{x}_{01}$$

Media del emésimo arreglo de la segunda variable (media de x_{h0}, f_{hm}).

$$\frac{\Sigma x_{h0} f_{hm}}{\Sigma f_{hm}} = \bar{x}_{m0}$$

Desviaciones de la primer variable respecto a las media de los arreglos de esa variable:

$$d_{11} = x_{01} - \bar{x}_{01} \qquad d_{21} = x_{02} - \bar{x}_{01} \qquad d_{31} = x_{03} - \bar{x}_{01}$$

$$d_{32} = x_{03} - \bar{x}_{02}$$

$$d_{33} = x_{03} - \bar{x}_{03}$$

Desviaciones de la segunda variable respecto a las medias de los arreglos de esa variable:

$$D_{11} = x_{10} - \bar{x}_{10} \qquad D_{12} = x_{10} - \bar{x}_{20} \qquad D_{13} = x_{10} - \bar{x}_{30}$$

$$D_{23} = x_{20} - \bar{x}_{30}$$

$$D_{33} = x_{30} - \bar{x}_{30}$$

Desviaciones medias cuadráticas:

σ_{12} — desviación media cuadrática obtenida a partir de las d

σ_{21} — desviación media cuadrática obtenida a partir de las D

Fórmulas de las razones de correlación de una distribución bivariada.—De acuerdo con la simbología anterior, la fórmula que consignamos en palabras en páginas anteriores como equivalente de la razón de correlación, puede escribirse:

$$\eta_{12} = \sqrt{1 - \frac{\sigma_{12}^{2}}{\sigma_{1}^{2}}}$$

$$\eta_{21} = \sqrt{1 - \frac{\sigma_{12}^{2}}{\sigma_{2}^{2}}}$$

Como puede verse, hemos consignado dos fórmulas para sendas razones de correlación de una sola distribución bivariada. En determinados casos, las dos razones pueden confundirse, según veremos más adelante, pero, por regla general, son diferentes entre sí.

Procedimiento de cálculo de la razón de correlación (procedimiento largo).—Para calcular la razón de corelación entre dos variables, puede seguirse el procedimiento que esquematizamos en las líneas subsecuentes:

I.—Calcúlese la desviación media cuadrática de los valores de la variable que se considerará como 'dependiente" con respecto a las medias de los arreglos de dicha variable correspondientes a cada uno de los diferentes valores de la variable que se ha de considerar "independiente". Para ello:

1.—Consígnense los datos de la serie bivariada de frecuencias en un cuadro de doble entrada,

A.—cuyos encabezados columnares serán los valores de una de las variables, x_{01}

B.—cuyos rubros filares serán los valores de la otra variable, y x_{10}

C.—en cuyas casillas figurarán las frecuencias conjuntas f_{ij}

Todo lo anterior indicará:

el número de casos en que un valor de la primer variable (dado por el encabezado de la columna a la que corresponda la casilla)
está asociado con un valor determinado de la segunda variable (especificado por el rubro de la fila a que dicha casilla corresponde)

2.—Obténganse las medias de los arreglos de la variable que se haya de considerar como dependiente. Para esto.

1ª—Abranse tantas columnas a continuación de las del cuadro original como sean las columnas que figuren en éste. En caso de tomarse como independiente la variable que consideramos dependiente, será necesario trazar tantas hileras como hileras figuren en el cuadro original.

2º—Consígnense en esas columnas los productos obtenidos de multiplicar el valor de la variable que figura como rubro de la hilera por la frecuencia conjunta que figura en la casilla correspondiente del cuadro original. En caso de haberse tomado la otra variable como dependiente, se tratará de consignar en las hileras recién abiertas los productos obtenidos de multiplicar el valor del encabezado de la columna por la frecuencia conjunta de la casilla que corresponda a la que se utiliza, dentro del cuadro original.

3º—Súmense al pie de las columnas los productos obtenidos. En caso de haberse tomado la otra variable como dependiente, se sumarán al extremo de las hileras los productos obtenidos.

4º—Súmense al pie de las columnas, en el cuadro original, las frecuencias conjuntas. En caso de tratarse de la otra variable, súmense al extremo de las filas (en ese mismo cuadro original) esas frecuencias conjuntas. De este modo se obtienen los efectivos de cada arreglo (vertical en un caso, horizontal en el otro.)

5º—Divídase la suma de cada columna del cuadro duplicado entre la suma de la columna correspondiente del cuadro original. En caso de tratarse de la otra variable considerada como dependiente, divídase la suma de cada hilera del cuadro duplicado entre la suma de la hilera correspondiente del cuadro original. De este modo se obtiene las medias de cada arreglo (vertical en un caso, horizontal en otro).

3.—Abranse tantas columnas o tantas hileras (según sea la variable que se esté considerando como dependiente) como columnas o hileras figuren en el cuadro original. De cada rubro de cada hilera, réstense todas y cada una de las medias de los arreglos y consígnense las desviaciones a lo largo de la hilera, a modo de que cada desviación se encuentre en la hilera cuyo rubro es el minuendo del que se obtuvo, y en la columna cuyo encabezado sea la media del arreglo que sirvió de substraendo para obtenerla. En caso de haberse tomado como dependiente la otra variable, habrá necesidad de abrir tantas filas como las del cuadro original, restar de cada encabezado columnar, todas y cada una de las medias de los arreglos correspondientes, consignando las desviaciones en forma parecida a como se hizo en el caso anterior, en las casillas correspondientes del triplicado del cuadro. Lo que se ha consignado son desviaciones con respecto a las medias de los arreglos.

4.—Abranse tantas columnas o hileras como haya en el cuadro original para consignar en esta cuadriplicación del cuadro original, en las casillas correspondientes, el cuadrado de cada una de las desviaciones recién consignadas en el triplicado del cuadro.

5.—Abranse tantas columnas o hileras como en el cuadro original para consignar en las casillas correspondientes. el producto que se obtiene de multiplicar cada cuadrado de la desviación por la frecuencia conjunta que figura en la casilla correspondiente del cuadro original. En la práctica, para evitar confusiones, se puede consignar el cuadro original en papel opaco y sobreponer el cuadriplicado en papel transparente a fin de encontrar la correspondencia correcta de cuadrados de desviaciones y frecuencias conjuntas.

6.—Al pie de cada una de las columnas (o al extremo de cada una de las hileras) del quintuplicado del cuadro, deberán obtenerse las sumas. Al extremo de la hilera de sumas (o al pie de la columna de sumas, en el otro caso) se formará una gran suma. Al dividir esta suma entre la suma de todas las frecuencias conjuntas del cuadro original, se obtiene la desviación cuadrática media con respecto a las medias de los arreglos, al cuadrado.

II.—Calcúlese la desviación media cuadrática de los valores de la variable que se ha considerado como "dependiente" con respecto a las media aritmética general. Para ello:

1.—Tómese como punto de partida el cuadro de doble entrada originario.

2.—Súmense al extremo de cada hilera las frecuencias conjuntas. En caso de haberse tomado la otra variable como dependiente, deberán sumarse al pie de cada columna dichas frecuencias conjuntas.

3.—Abrase una columna para contener los productos de los rubros de cada hilera por las frecuencias recién encontradas. En caso de tomar la otra variable como dependiente, deberá de abrirse una nueva hilera para contener los productos de los encabezados de cada columna por las frecuencias recién encontradas al pie de las columnas.

4.—Súmense al pie de la columna dichos productos (o al extremo de la hilera en el otro caso).

5.—Súmense al pie de la columna correspondiente las frecuencias recién encontradas (o al extremo de la hilera, las otras frecuencias recién encontradas). La suma final es el efectivo de la distribución.

6.—Divídase la suma obtenida en 4 entre la obtenida en 5 para obtener la media aritmética de la serie.

7.—Réstese de cada uno de los rubros (en un caso) o de cada uno de los encabezados (en el otro), la media aritmética obtenida, para obtener a su vez desviaciones con respecto a la media, que se consignarán en una columna (en el primer caso), o en una nueva fila (en el segundo caso).

9.—Multiplíquense esos cuadrados por las frecuencias obtenidas en 2.

10.—Súmense los productos así obtenidos.

11.—Divídase la suma de dichos productos entre el efectivo de la distribución.

12.—Extráigase la raíz cuadrada del cociente.

El resultado es la desviación media cuadrática de la variable considerada como dependiente, con respecto a su media.

III.—Obténgase la razón de correlación a partir de las medidas calculadas en I y II.

1.—Dividiendo el resultado obtenido en I-6 entre el resultado obtenido en II-11 (resultado previo a la extracción de la raíz cuadrada).

2.—Restando da la unidad el cociente, y

3.—Extrayendo la raíz cuadrada de la resta.

El resultado, o sea la razón de correlación indica el grado de asociación entre las dos variables, o el grado en que no se comete un error al predecir un valor determinado de la variable dependiente cuando se conoce un valor dado de la independiente.

EJEMPLO DE CALCULO DE LA RAZON DE CORRELACION

	1.—Datos en Cuadro de Doble Entrada.								2.—Medias de los Arreglos de x_{10}							
x_{10} \ x_{01}	2	3	4	5	6	7	8	9	2	3	4	5	6	7	8	9
0	6	6	3	20	13	7			0	0	0	0	0	0		
1	9	8	16	16	9	9			9	8	16	16	9	9		
2	5	5	12	8	24	9			10	10	24	16	48	18		
3		2	4	8	15	12				6	12	24	45	36		
4			1	10	17	26					4	40	69	104		
5				2	2	22	5					10	10	110	25	
6						19	7							114	42	
7						8	6							56	42	
8							6								48	
9							14								126	
10							12								120	
11							10	8							110	88
12							8	16							96	192
13								18								234
14								12								168
15								10								150
SUMAS	20	21	36	64	80	112	68	64	19	24	56	106	180	447	609	832
MEDIAS									0.95	1.14	1.56	1.66	2.25	4	9	13

EJEMPLO DE CALCULO DE LA RAZON DE CORRELACION

3.—Desviaciones de las x con respecto a las medias de los arreglos

x_{10} \ x_{01}	2	3	4	5	6	7	8	9
0	−0.95	−1.14	−1.56	−1.66	−2.25	−4.00		
1	−0.05	−0.14	−0.56	−0.66	−1.25	−3.00		
2	−0.05	0.86	0.44	0.34	−0.25	−2.00		
3		1.86	1.44	1.34	0.75	−1.00		
4			2.44	2.34	1.75	0.00		
5				3.34	2.75	1.00	−4.00	
6						2.00	−3.00	
7						3.00	−2.00	
8							−1.00	
9							0.00	
10							1.00	
11							2.00	−2.00
12							3.00	−1.00
13								0.00
14								1.00
15								2.00

4.—Cuadrados de las desviaciones respecto a las medias de los arreglos.

x_{01}	2	3	4	5	6	7	8	9
0	0.9025	1.2996	2.4336	2.7556	5.0625	16.0000		
1	0.0025	0.0196	0.3136	0.4356	1.5625	9.0000		
2	0.0025	0.7396	0.1936	0.1156	0.0625	4.0000		
3		3.4596	2.0736	1.7956	0.5625	1.0000		
4			5.9536	5.4756	3.0625	0.0000		
5				11.1556	7.5625	1.0000	16.0000	
6						4.0000	9.0000	
7						9.0000	4.0000	
8							1.0000	
9							0.0000	
10							1.0000	
11							4.0000	4.0000
12							9.0000	1.0000
13								0.0000
14								1.0000
15								4.0000

EJEMPLO DE CALCULO DE LA RAZON DE CORRELACION

5.—Ponderación de los cuadrados de las desviaciones respecto de las medias de los arreglos (Factor de ponderación: frecuencias conjuntas)

x_{10} \ x_{01}	2	3	4	5	6	7	8	9
0	5.4150	7.7976	7.3008	55.1120	65.8125	96.0000		
1	0.0225	0.1568	5.0176	6.9696	14.0625	81.0000		
2	0.0125	3.6980	2.3232	0.9248	1.5000	36.0000		
3		6.9192	8.2944	14.3648	8.4375	12.0000		
4			5.9536	54.7560	52.0625	0.0000		
5				22.3112	15.1250	22.0000	80.0000	
6						76.0000	63.0000	
7						72.0000	24.0000	
8							6.0000	
9							0.0000	
10							12.0000	
11							40.0000	38.0000
12							72.0000	16.0000
13								0.0000
14								12.0000
15								40.0000
SUMAS	5.4500	18.5716	28.8896	154.4384	157.0000	395.0000	297.0000	106.0000
SUMA DE SUMAS				1 162.3496				
entre el EFECTIVO (465)				$2.4996 = \sigma_{21}^2$				

EJEMPLO DE CALCULO DE LA RAZON DE CORRELACION

6.—Cálculo de la desviación media cuadrática de la variable x_{10}

x_{10} \ x_{01}	2	3	4	5	6	7	8	9	f	$x_{10}f$	d_{10}	d_{10}^2f
0	6	6	3	20	13	7			55	0	− 5	1375
1	9	8	16	16	9	9			67	67	− 4	1072
2	5	5	12	8	24	9			63	126	− 3	567
3		2	4	8	15	12			41	123	− 2	164
4			1	10	17	26			54	216	− 1	54
5				2	2	22	5		31	155	0	0
6						19	7		26	156	1	26
7						8	6		14	98	2	56
8							6		6	48	3	54
9							14		14	126	4	224
10							12		12	120	5	300
11							10	8	18	198	6	648
12							8	16	24	288	7	1176
13								18	18	234	8	1152
14								12	12	168	9	972
15								10	10	150	10	1000
SUMAS									465	2273		8870 ÷ 465 =
entre el EFECTIVO (465)										4.89		19 = σ_2^2
										$\doteq$ 5.00		

De acuerdo con los cálculos anteriores:

Desviación media cuadrática con respecto a las medias de los arreglos de la segunda variable. Elevada al cuadrado $\sigma_{21}^2 = 2.4996$

Desviación media cuadrática de la segunda variable con respecto a la media general. Elevada al cuadrado $\sigma_2^2 = 19$

Razón de correlación de 2 en 1: $\eta_{21} = \sqrt{1 - \frac{2.4996}{19}} = 0.932$

Simplificación de la fórmula y del procedimiento de cálculo de la razón de correlación.—De acuerdo con lo que hemos asentado anteriormente, la razón de correlación es el complemento aritmético del cociente de los cuadrados de las desviaciones medias cuadráticas de los valores de la variable dependiente con respecto a las medias de los arreglos y con respecto a la media general, sometido a radicación. O sea que:

$$\eta_{12} = \sqrt{1 - \frac{\sigma_{12}^2}{\sigma_1^2}}$$

Si ejecutamos operaciones dentro del radical, tendremos:

$$\eta_{12} = \sqrt{\frac{\sigma_1^2 - \sigma_{12}^2}{\sigma_1^2}}$$

En el numerador tenemos una diferencia entre dos medidas de variabilidad: Una medida de la variabilidad del conjunto con respecto a la media del conjunto, y una medida de la variabilidad del conjunto con respecto a las medias de los arreglos de dicho conjunto. La diferencia entre ambas medidas, ¿qué representa?

En términos generales, puede establecerse que:

$$\left.\begin{matrix}\text{La variación total} & \begin{matrix}\text{ENTRE}\\ \text{Y}\\ \text{DENTRO}\end{matrix}\end{matrix}\right\}\begin{matrix}\text{de los arreglos}\\ \text{es igual a}\end{matrix}\left\{\begin{matrix}\text{La variación entre}\\ \text{arreglos}\\ \text{Y}\\ \text{La variación dentro de}\\ \text{ellos}\end{matrix}\right.$$

O sea, que puede establecerse que:

$$\sigma_1^2 = \sigma_{a0}^2 + \sigma_{12}^2$$

Si pasamos σ_{12}^2, que figura con signo más en el segundo miembro, al primer miembro, con signo menos, obtendremos:

$$\sigma_1^2 - \sigma_{12}^2 = \sigma_{a0}^2$$

Al sustituir este valor en la expresión de la razón de correlación, obtenemos:

$$\eta_{12} = \sqrt{\frac{\sigma_{a0}^2}{\sigma_1^2}} = \frac{\sigma_{0a}}{\sigma_1}$$

En forma parecida, puede establecerse que:

$$\eta_{21} = \frac{\sigma_{a0}}{\sigma_2}$$

Procedimiento corto del cálculo de la razón de correlación.— Las deducciones anteriores son muy importantes, ya que, gracias a ellas se simplifican muy considerablemente los cálculos necesarios para la determinación del valor que tiene la razón de correlación de una serie bivariada de frecuencias.

En efecto, de acuerdo con las fórmulas anteriores, la razón de correlación comienza haciendo honor a su nombre, en cuanto es un verdadero cociente o razón entre dos medidas de variabilidad absoluta: la variabilidad de los arreglos con respecto a la media aritmética general del conjunto, y la variabilidad de los valores no arreglados, con respecto a esa misma media aritmética general. De acuerdo con ello, el procedimiento de cálculo consistirá:

I.—En calcular la desviación cuadrática media de las medias de los arreglos. Para ello será neecsario:

1.—Consignar los datos en un cuadro de doble entrada como el empleado en el procedimiento largo.

2.—Calcular las medias de los arreglos mediante una duplicación de dicho cuadro de doble entrada y consignación en sus casillas de los productos de cada rubro inicial de hilera por todas y cada una de las frecuencias contenidas en las casillas de la hilera (o, en su caso, de los productos de cada encabezado de columna por todas y cada una de las frecuencias contenidas en las casillas de la columna) con suma subsecuente (al pie o al extremo) de dichos productos, y división de cada suma entre la suma de las frecuencias de la columna o de la hilera correspondientes (del cuadro originario).

3.—Calcular la media aritmética del conjunto, para lo cual se sumarán al extremo de las hileras (o al pie de las columnas) todas las frecuencias de una misma hilera (o de una misma columna); se multiplicarán dichas frecuencias por los rubros (o encabezados) de las hileras (o columnas); se sumarán al pie (o al extremo) dichos productos, y se dividirá dicha suma entre la suma de las frecuencias.

4.—Restar de cada media de los arreglos (consignada al pie de la columna correspondiente, o al extremo de la hilera respectiva en el duplicado del cuadro) la media aritmética general. Así se obtendrá una serie de desviaciones de los arreglos respecto de la media general (que figurarán en una hilera en un caso, en una columna en otro).

5.—Elevar al cuadrado cada una de las desviaciones así obtenidas (consignando los cuadrados en otra hilera o en otra columna).

6.—Multiplicar cada cuadrado por la frecuencia que figura al pie de la columna (o de la hilera) correspondiente, en el cuadro originario.

7.—Sumar todos esos productos al extremo de la hilera (o al pie de la columna).

8.—Dividir la suma entre el efectivo de la distribución (o suma de todas las frecuencias).

9.—Extraer la raíz cuadrada del cociente, para obtener la desviación cuadrática media de las medias de los arreglos con respecto a la media general.

II.—Calcular la desviación cuadrática media de la variable considerada como dependiente, de acuerdo con el mismo procedimiento que se siguió en la segunda parte del procedimiento largo, hasta obtener la raíz cuadrada.

III.—Dividir el resultado obtenido en la primera parte del procedimiento entre el resultado obtenido en la segunda parte de este mismo procedimiento. El cociente es la razón de correlación.

EJEMPLO DE CALCULO DE LA RAZON DE CORRELACION (Procedimiento Corto)

x_{01} / x_{10}	2	3	4	5	6	7	8	9	2	4	4	5	6	7	8	9	f	$x_{10}f$
0	6	6	3	20	13	7			0	0	0	0	0	0			55	0
1	9	8	16	16	9	9			9	8	16	16	9	9			67	67
2	5	5	12	8	24	9			10	10	24	16	48	18			63	126
3		2	4	8	15	12				6	12	24	45	36			41	123
4			1	10	17	26					4	40	68	104			54	216
5			2	2	2	22	5					10	10	110	25		31	156
6						19	7							114	42		26	165
7						8	6							156	42		14	98
8							6								48		6	48
9							14								126		14	126
10							12								120		12	120
11							10	8							110	88	18	198
12							8	16							96	192	24	288
13								18								234	18	234
14								12								168	12	168
15								10								150	10	150
SUMAS	20	21	36	64	80	112	68	64	19	24	56	106	180	447	609	832	465	2273
MEDIAS									0.95	1.14	1.56	1.66	2.25	4.00	9.00	13.00		4.89

Desviaciones de las medias de los arreglos −3.94 3.75 −3.33 −3.23 −2.64 −0.89 4.11 8.11 con respecto a la media general (4.89)

Cuadrados de esas desviaciones 15.56 14.06 11.09 10.43 6.97 0.79 16.89 65.77

Ponderación de los cuadrados por las frecuencias del arreglo (hilera de sumas) 310.40 295.26 399.24 667.52 557.6 88.48 1148.52 4209.24

suma	7676.30
entre efectivo (465)	16.5081
	7676.30
raíz cuadrada	4.06

Desviación media cuadrática de las medias de los arreglos respecto a la general.
4.36

Desviación media cuadrática de las x_{10} respecto a su media (para cálculo, véase el ejemplar anterior)

$$\eta_{21} = \frac{4.06}{4.36} = 0.931$$

SUGESTION PARA RESOLVER EL PROBLEMA DE LA CORRELACION CURVILINEA MEDIANTE LA REDUCCION DEL MISMO A UN PROBLEMA DE CORRELACION RECTILINEA

Para principiar con el caso más sencillo, supongamos que se tienen dos variables x_1, x_2 y que dichas variables se encuentran relacionadas al través de la expresión:

$$x_1 = a_{12}\, x_2^2$$

En tal caso, como ya hemos señalado, el establecimiento de un índice de correlación al través de la comparación de las variancias (o los cuadrados de los errores de estimación a partir de la regresión y a partir de la media de la variable) se dificulta en vista de la forma de distribución de la variable dependiente. En cambio, es factible recurrir a una transformación en las variables. Si se extrae la raíz cuadrada de ambos miembros, se obtiene:

$$x_1^{\frac{1}{2}} = a_{12}^{\frac{1}{2}}\, x_2$$

Si designamos la raíz cuadrada de x_1 por su mayúscula X_1 y la raíz cuadrada de a_{12} por la mayúscula A_{12}, tendremos:

$$X_1 = A_{12}\, x_2$$

O sea, que entre la nueva variable X_1 y la variable x_2 existe una relación rectilínea. Por lo mismo, es factible encontrar el índice de correlación rectilínea entre X_1 y x_2 y medir, al través del mismo, en forma indirecta, la correlación entre x_1 y x_2.

En forma semejante, se puede proceder en cuanto se trate de una relación logarítmica, pues la transformación se logrará fácilmente tomando el antilogaritmo de la variable correspondiente y correlacionando de acuerdo con los procedimientos de la correlación rectilínea la nueva variable resultante.

En general, para proceder a la trasformación de las variables correspondientes puede hacerse la interpolación de la curva correspondiente, sustituyendo la antigua variable por la variable independiente de la curva interpolada, la cual pasará a ser variable dependiente que correlacionar con la otra variable antigua.

INVESTIGACION EXPERIMENTAL SOBRE LA CORRELACION ENTRE SERIES CURVILINEAS

En el intento que se haga para correlacionar dos series estadísticas, pueden darse, por lo menos, los siguientes casos:

1.—Correlación entre dos series que son, ambas, del sistema rectilíneo.
2.—Correlación entre una serie rectilínea y una serie de cualquier sistema no-rectilíneo (curvilíneo).
3.—Correlación entre dos series curvilíneas:
   a.—pertenecientes al mismo sistema.
   b.—pertenecientes a diferentes sistemas.

Correlación entre dos series del sistema rectilíneo.—Este caso ha sido tratado (con la amplitud asequible dentro de una presentación tan elemental como lo nuestra) en las páginas anteriores, por lo cual no insistiremos en él.

Correlación entre una serie rectilínea y una serie de un sistema no-rectilíneo.—Tomaremos el caso de una primera serie interpolable por una recta y que deseemos correlacionar con una segunda serie interpolable por una curva exponencial.

Si representamos:

1.—Los valores de la primera serie, por x_1,
2.—Los valores de la segunda serie, por x_2, y
3.—Si suponemos que x_1 y x_2 dependen de una variable y, de acuerdo con dos ecuaciones distintas (rectilínea una, exponencial la otra), podremos escribir:

$$x_1 = b_0 + b_1 y$$

$$x_2 = c_0 c_1{}^y$$

En ambas expresiones, tanto las bes como las ces representan constantes de las ecuaciones respectivas.

Al tomar los logaritmos de los dos miembros de la segunda ecuación (exponencial), obtendremos:

$$Lx_2 = Lc_0 + y\,Lc_1$$

Si Lx_2 lo representamos por X_2 Lc_0 por C_0 y Lc_1 por C_1, tendremos:

$$X_2 = C_0 + C_1 y$$

O sea, que la expresión anterior habrá adquirido el aspecto propio de una ecuación de primer grado. Es decir, que será como si tuviéramos una nueva serie, rectilínea que correlacionar con la primera serie también rectilínea, de acuerdo con procedimientos en los que se ha insistido suficientemente en páginas anteriores.

De este modo, las dos series por correlacionar serán x_1 y X_2 Entre ambas es fácil establecer una correlación rectilínea, como la ya estudiada, pues es posible establecer entre ellas relaciones dadas por una ecuación de estimación del tipo de la siguiente:

$$x_1 = f_0 + f_{12} X_2$$

Tanto la serie de las x_1 como la de las x_2 pueden reducirse a desviaciones respecto de sus correspondientes medias aritméticas. La primera de dichas series de desviaciones no tiene por qué modificarse en su simbología, y puede seguir representándose por d_1; en cuanto a la segunda de dichas series de desviaciones, en cuanto no corresponde a las x_2 originales, sino a las X_2 derivadas de ellas, la representaremos por la interdental correspondiente z_2. La correspondiente ecuación de estimación podrá escribirse, obviamente, como:

$$d_1 = F_{12} z_2$$

Asimismo, es posible reducir las desviaciones respecto a las medias aritméticas de cada serie a desviaciones en unidades sigmáticas. Seguiremos representando la desviación sigmática de la primera serie por delta minúscula índice uno (δ_1); en cambio, la desviación de las nuevas x_2 con respecto a la media aritmética de esas mismas x_2 en unidades de su desviación cuadrática media la representaremos por la interdental correspondiente del alfabeto griego theta o zeta minúscula (θ_2). La nueva ecuación de estimación resultará ser:

$$\delta_1 = \varphi_{12}\, \theta_2$$

En esta última expresión o ecuación sigmática de estimación aparece φ_{12} (phi) en el sitio que correspondía anteriormente a F_{12} de acuerdo con la costumbre que hemos seguido, consistente en representar: los parámetros de la ecuación originaria por minúsculas latinas; los de la ecuación dada en términos de desviación, por mayúsculas latinas, y los de la ecuación sigmática, por las minúsculas griegas correspondientes.

Gracias a la última ecuación, es posible escribir, como fórmula del índice de correlación:

$$r_{12} = \frac{\Sigma\, \delta_1\, \theta_2}{N}$$

Puede decirse que esta medida de la correlación entre las dos series no es una medida directa, sino una medida indirecta de correlación. Sin embargo, cabe recordar que, el sociólogo Auguste Comte, en su curso de filosofía positiva hacía notar la forma en que medir es una operación que consiste, precisamente, en apreciar el valor de una magnitud mediante un proceso, indirecto, de comparación con otra magnitud conocida. No es arriesgado afirmar que si hay correlación entre una serie rectilínea y la serie rectilínea resultante de una transformación de una serie exponencial, existirá asimismo correlación entre la serie rectilínea y la serie exponencial originaria.

Correlación entre dos series curvilíneas.—Sea que las series curvilíneas que traten de correlacionarse pertenezcan a un mismo o a diferentes sistemas de curvas, la diferencia de estos intentos de correlación con los del caso anterior estriba en que, en éste, las transformaciones son necesarias para las dos series, mientras en el previo una de las series permanecía incambiada.

Tomemos las dos series x_1 y x_2, de las cuales la primera es representable por una exponencial y la segunda por una logarítmica de los tipos siguientes:

$$x_1 = c_0\, c_1{}^{y}$$
$$x_2 = f_0 + L\, y$$

Al tomar los logaritmos de los dos miembros de la primera, se obtiene:

$$Lx_1 = Lc_0 + y\, Lc_1$$

O bien:

$$X_1 = C_0 + C_1\, y$$

Al tomar antilogaritmos en la segunda, se tiene:

$$e^{x_2} = e^{(f_0 + Ly)}$$

O sea:

$$e^{x_2} = e^{f_0} e^{Ly}$$

puesto que la exponencial "e elevado a la suma de $f_0 + Ly$" es igual al producto de "e elevado al primer sumando f_0" por "e elevado al segundo sumando Ly". El segundo de estos factores es igual —pura y simplemente— a y, puesto que cuando la base se eleva al logaritmo de una cantidad (e elevada al logaritmo natural de y) se obtiene la propia cantidad (ya que "logaritmo de un número es, por definición, aquel otro número al que hay que elevar una cantidad constante llamada base para obtener el número propuesto").

Consiguientemente, la expresión anterior se convierte en:

$$e^{x_2} = e^{f_0} y$$

Podemos representar al primer miembro (e elevado a x_2) por X_2, y al coeficiente del segundo miembro (e elevado a f_0) por F_0. Así, tendremos una expresión de tipo rectilíneo:

$$X_2 = F_0 \, y$$

Como es fácil comprender, es posible calcular la correlación rectilínea entre x_1 y X_2. La reducción de cada una de las nuevas series a desviaciones respecto a sus medias aritméticas producirá una serie de z_1 y otro serie de z_2. Es posible expresar estas series en unidades de sus desviaciones cuadráticas medias. A estas desviaciones respecto de las medias aritméticas en unidades de las respectivas desviaciones cuadráticas medias las representaremos por θ_1 y θ_2.

De acuerdo con lo anteriores, el índice de correlación rectilínea entre las dos series será:

$$r_{12} = \frac{\Sigma \, \theta_1 \, \theta_2}{N}$$

Especificación del índice de correlación cuando las series son curvilíneas.—A fin de especificar cuál ha sido la transformación indispensable de las series primera y segunda destinada a obtener el índice de correlación, colocaremos, en la misma columna del subíndice que identifica a la primera serie un guarismo indicativo de la transformación efectuada con dicha serie, y en la columna correspondiente al subíndice identificador de la segunda serie, el guarismo que indique qué transformación fue preciso realizar con ella.

Los guarismos que emplearemos serán:

1.—cuando no hay necesidad de transformación por tratarse de rectas,
2.—cuando se trate de parábolas de segundo grado,
3.—cuando se trate de parábolas de tercer grado.
n.—cuando se trate de parábolas de grado n.
O.—cuando se trate de logarítmicas.
e.—cuando se trate de exponenciales.

Así, son las siguientes las interpretaciones que hay que dar a los siguientes símbolos:

$r_{\substack{12\\11}}$ Correlación entre rectas (1 para la 1ª serie, 1 para la segunda).

$r_{\substack{12\\12}}$ Correlación entre recta (1 como subíndice debajo de 1) y parábola (2 como subíndice debajo de 2).

$r_{\substack{12\\00}}$ Correlación entre logarítmicas (ceros debajo de cada uno de los subíndices 1 y 2).

$r_{\substack{12\\ee}}$ Correlación entre exponenciales (e debajo de cada uno de los dos subíndices 1 y 2).

$r_{\substack{12\\0e}}$ Correlación entre logarítmicas (cero para la primera serie) y exponencial (e para la segunda serie).

EJEMPLO PEDAGOGICO DE CALCULO DE UNA CORRELACION ENTRE UNA SERIE RECTILINEA Y UNA SERIE EXPONENCIAL

	Datos		Transformación	Elaboraciones			
	x_1	x_2	$\log x_2 = x_2$	d_1	D_2	d_1^2	D_2^2
	3	1	0.00000	−13.5	−1.35463	182.25	1.8350224369
	6	2	0.30103	−10.5	−1.05360	110.25	1.1100729600
	9	4	0.60206	− 7.5	−0.75257	56.25	0.5663616040
	12	8	0.90309	− 4.5	−0.45154	20.25	0.2038883716
	15	16	1.20412	− 1.5	−0.15052	2.25	0.0226562704
	18	32	1.50515	1.5	0.15062	2.25	0.0226562704
	21	64	1.80618	4.5	0.45154	20.25	0.2038883716
	24	128	2.10721	7.5	0.75257	56.25	0.5663616040
	27	256	2.40824	10.5	1.05360	110.25	1.1100729600
	30	512	2.70927	13.5	1.35463	182.25	1.8350224369
SUMAS	165		13.54635			742.50	7.476003286
MEDIAS	16.5		1.354635			74.25	0.7476003286
RAÍCES						8.6	0. 8 6 4 6 3
o desviaciones medias cuadráticas:						σ_1	σ_2

E L A B O R A C I O N E S

	$\delta_1 = d_1/\sigma_2$	$\theta_2 = D_2/\sigma_2$	$\delta_1\,\theta_2$	θ_2^2
	−156.	−1.56	2.4336	2.4336
	−1.22	−1.22	1.4884	1.4884
	−0.87	−0.87	0.7569	0.7569
	−0.52	−0.52	0.2704	0.2704
	−0.17	−0.17	0.0289	0.0289
	0.17	0.17	0.0289	0.0289
	0.52	0.52	0.2704	0.2704
	0.87	0.87	0.7569	0.7569
	1.22	1.22	1.4884	1.4884
	1.56	1.56	2.4336	2.4336
			9.9564	9.9564
			Núm.	Den.

Fórmula:

$$r_{\substack{12\\1e}} = \frac{\Sigma\,\delta_1\,\theta_2}{\Sigma\,\theta_2^2}$$

$$r_{\substack{12\\1e}} = \frac{9.9564}{9.9564} = 1$$

$r_{\substack{12\\1e}} = 1$

Forma explícita de la ecuación de estimación.—A partir de la forma más compacta de la ecuación de estimación de la primer variable, que toma como base de cálculo a la transformada de la segunda variable,

$$\delta_1 = \underset{1e}{r_{12}}\, \theta_2$$

es posible obtener, mediante la substitución de δ_1 y de θ_2 por sus equivalentes, lo siguiente:

$$\frac{d_1}{\sigma_1} = \underset{1e}{r_{12}} \frac{D_2}{\sigma_{2'}}$$

O, en la sustitución de d_1 y D_2 por los suyos:

$$\frac{x_1 - \bar{x}_1}{\sigma_1} = \underset{1e}{r_{12}} \frac{x_2 - \bar{x}_2}{\sigma_{2'}}$$

Como x_2 es igual a Lx_2, $\bar{x}_2$ es igual a la media de las x_2; o sea, que $\bar{x}_2$ también es igual a la media de los $L\,x_2$. Es decir:

$$\bar{x}_2 = \overline{Lx_2}$$

Pero, la media de los logaritmos de x_2 es igual a la suma de dichos logaritmos sobre el efectivo (N).

$$\bar{x}_2 = \frac{\Sigma Lx_2}{N}$$

Como en el numerador aparece una suma de logaritmos, dicha suma se puede substituir por el logaritmo de un producto (que representaremos mediante el operador pi mayúscula, o "multiplicador", Π).

$$\bar{x}_2 = \frac{L\Pi x_2}{N}$$

Pero Πx_2 es la parte fundamental del cálculo de la media geométrica de las x_2; en efecto, la media geométrica de las x_2 $(\underset{0}{\bar{x}_2})$ es:

$$\underset{0}{\bar{x}_2} = (\Pi_2)^{1/N}$$

$$\therefore \quad x_2 = (\underset{0}{\bar{x}_2})^N$$

Sustituido x_2 por su equivalente, se tiene:

$$\bar{x}_2 = \frac{L\,(\underset{0}{\bar{x}_2})^N}{N}$$

Sin embargo, en el numerador aparece el logaritmo de una potencia, y éste es igual al exponente tomado como coeficiente del logaritmo de la cantidad:

$$L(\underset{0}{\bar{x}_2})^N = N\,L\,(\underset{0}{\bar{x}_2})$$

Por tanto:

$$\bar{x}_2 = \frac{N\,L\,(\underset{0}{\bar{x}_2})}{N} = L(\underset{0}{\bar{x}_2})$$

Resultado ya conocido, según el cual: la media aritmética de los logaritmos de una serie $(\bar{x}_2)$ es igual al logaritmo de la media geometría de la serie (o sea: $L\,(\underset{0}{\bar{x}_2})$).

De acuerdo con lo anterior, la sustitución de $\bar{x}_2$ por su equivalente en la ecuación de estimación,

$$\frac{x_1 - \bar{x}_1}{\sigma_1} = \underset{1e}{r_{12}} \frac{x_2 - \bar{x}_2}{\sigma_{2'}}$$

produce:

$$\frac{x_1 - \bar{x}_1}{\sigma_1} = r_{\underset{1e}{12}} \frac{Lx_2 - \bar{x}_{\underset{0}{2}}}{\sigma_{2'}}$$

Por su parte, σ_2, es igual a:

$$\sqrt{\frac{\Sigma D_2^2}{N}} = \sqrt{\frac{\Sigma(Lx_2 - L\bar{x}_{\underset{0}{2}})^2}{N}}$$

La sustitución de este equivalente, nos permite establecer, como ecuación de estimación, la siguiente:

$$\frac{x_1 - \bar{x}_1}{\sigma_1} = r_{\underset{1e}{12}} \frac{\sqrt{N}\,[\,Lx_2 - L(\bar{x}_{\underset{0}{2}}\,]}{\sqrt{\Sigma(Lx_2 - L\bar{x}_2)^2}}$$

Aun cuando pueden hacerse ulteriores reducciones, ésta parece ser una forma conveniente de la ecuación de estimación para la correlación entre una recta (x_1) y una exponencial (x_2).

Momentos "logarítmicos" de una distribución.—Lo anterior nos pone en camino de establecer medidas características adicionales para las distribuciones estadísticas.

En estrecho paralelismo con $d_1 = x_1 - \bar{x}_1$ ha aparecido una desviación entre logaritmos $Lx_2 - L\,(\bar{x}_{\underset{0}{2}})$ que, como toda desviación, es diferencia ($-$) entre una variable (x_2), pero tomada ahora en forma logarítmica ($L\,x_2$), y una media (en el caso la media geométrica que hemos representado por $\bar{x}$ afectada del subíndice doble $\underset{0}{2}$) tomada también en forma logarítmica.

Por otra parte, esta desviación entre logaritmos puede considerarse como el logaritmo de un cociente entre cantidades alogaritmadas (x_2 y $\bar{x}_{\underset{0}{2}}$):

$$Lx_2 - L(\bar{x}_{\underset{0}{2}}) = L\left(\frac{x_2}{\bar{x}_{\underset{0}{2}}}\right)$$

La cantidad de dentro del paréntesis podemos definirla, en términos generales, como una "razón" estadística, reflejo claro, en el nivel logarítmico o geométrico, de la "desviación" estadística del nivel natural o aritmético:

$$L\left(\frac{x_2}{\bar{x}_{\underset{0}{2}}}\right) = L\, r_2 = R_2$$

Si $L\, r_2$ es una especie de desviación, puede pensarse en constituir una serie de momentos análogos a los constituidos cuando se trabajaba con desviaciones ordinarias. Es decir, puede pensarse en una serie de sumas de potencias de las R_2 o $L\, r_2$. Designaremos tales momentos por mus minúsculas griegas (μ) afectadas de un subíndice que señale el orden del momento, o sea el exponente al que hay que elevar las R, y de otro subíndice 0 alineado con él en la misma columna, que indique que se trata de momentos del nivel logarítmico (o sea: momentos de los logaritmos de las razones de la variable respecto de su media geométrica).

$$\mu_{\underset{0}{n}} = \frac{\Sigma R_2{}^n}{N} = \frac{\Sigma L^n r_2}{N}$$

De acuerdo con lo anterior, existe un segundo momento de los logaritmos de las razones:

$$\mu_{\underset{0}{2}} = \frac{\Sigma L^2 r_2}{N}$$

Al que corresponde una raíz cuadrada que no es sino la desviación media cuadrática logarítmica o desviación geocuadrática de la serie:

$$\sqrt{\frac{\Sigma L^2 r_2}{N}} = \underset{0}{\sigma_2}$$

Esta es precisamente la desviación estadística media con la que hemos tenido que trabajar al correlacionar recta y exponencial.

A más de momentos como éstos, con respecto a la media geométrica, existen momentos respecto al origen que, en este caso es 1, como correspondiente al logaritmo de 0. Estos momentos los representaremos por nus griegas minúsculas (ν), afectadas de los subíndices dobles correspondientes, de acuerdo con la siguiente fórmula:

$$\underset{0}{\nu_n} = \frac{\Sigma L^n x_2}{N}$$

En el caso particular del primer momento, se tiene:

$$\underset{0}{\nu_1} = \frac{\Sigma L x_2}{N} = \overline{(L\, x_2)} = L\, \underset{0}{\bar{x}_2}$$

Es decir, que el primer momento de los logaritmos, con respecto al origen, no es sino el logaritmo de la media geométrica de la serie. De este modo, puede escribirse:

$$\frac{x_1 - \bar{x}_1}{\sigma_1} = \underset{1e}{r_{12}} \frac{\underset{0}{x_2} - \underset{0}{\nu_{01}}}{\underset{0}{\sigma_2}}$$

Para establecer una correcta correspondencia simbológica, podemos sustituir $\bar{x}_1$ por su equivalente en términos de los momentos con respecto al origen (primer momento natural con respecto al origen ν_1, que, en esta nueva simbología se convierte en $\underset{1}{\nu_1}$, o, más precisamente, por tratarse del primer momento de la primer variable, en $\underset{1}{\nu_{10}}$).

$$\frac{\underset{0}{x_1} - \underset{1}{\nu_{10}}}{\underset{1}{\sigma_1}} = \underset{1e}{r_{12}} \frac{\underset{0}{x_2} - \underset{1}{\nu_{01}}}{\underset{0}{\sigma_2}}$$

O bien:

$$\frac{\underset{1}{x_1} - \underset{1}{\nu_{10}}}{\underset{1}{\mu_{20}}} = \underset{1e}{r_{12}} \frac{\underset{0}{x_2} - \underset{0}{\nu_{01}}}{\underset{0}{\mu_{02}}}$$

Si se toma como base esta formulación, puede pensarse en presentar en forma generalizada la estructura de la ecuación de estimación. Ecuación de estimación de una variable a partir de otra, sea cual fuere el sistema al que cada una de ellas corresponda.

Fórmula generalizada de la ecuación de estimación.—La ecuación de estimación consta:

1.—De dos miembros, de los cuales:

a.—el primero contiene las estadísticas y parámetros de la primera serie, y

b.—el segundo contiene los de la segunda serie, multiplicados por el índice de correlación.

2.—Para cada serie aparece,
de acuerdo con el sistema al que partenezca:

a.—la variable (o una función de la variable que la lleve de su propio sistema al sistema rectilíneo): aparece la variable x_1 en el caso de la recta; la variable alogaritmada L x_2 en el de la exponencial.

b.—menos el primer momento de la serie que corresponda al sistema al que pertenece la serie (primer momento aritmético o natural para la recta; primer momento geométrico o logarítmico para la exponencial).

c.—dividido el residuo entre la raíz cuadrada del segundo momento del sistema al que corresponda cada serie.

O sea que, en la forma más amplia, puede establecerse:

$$\frac{\underset{A}{x_1} - \underset{A}{\nu_1}}{\sqrt{\underset{B}{\mu_{20}}}} = \underset{\overline{AB}}{r_{12}} \frac{\underset{B}{x_2} - \underset{B}{\nu_1}}{\sqrt{\underset{A}{\mu_{02}}}}$$

VIII.—COORDENACION ESTADISTICA

COORDENACION ESTADISTICA

La que ordinariamente se conoce como "correlación estadística 'por rangos' " —y que nosotros preferimos denominar coordenación estadística, para evitar el galicismo o el anglicismo "rango"— es un método de correlación que no se basa, como los previos, en caracteres de los fenómenos estudiados que sean expresables por medio de cardinales, sino en caracteres que se expresan mediante ordinales.

Lo que es para la estadística ordinaria o de cardinales una "serie" o arreglo de valores, lo es para la estadística ordinal o de jerarquías lo que llamaremos una "ordenación".

Una "ordenación" es un arreglo de los individuos de un conjunto, universo o población de acuerdo con la secuela que permite establecer entre ellos la magnitud más o menos grande con que se presenta en cada uno de ellos una característica determinada. Así, si los individuos designados por A, B, C, D tienen estaturas de 1.53, 1.90, 1.73, 162, puede formarse una ordenación de ellos, asignando a A el primer lugar, a D el segundo, a C el tercero y a B el cuarto:

Individuos	Medida de la característica que sirve de base a la ordenación	Ordenes
A	1.53	1º
B	1.90	4º
C	1.73	3º
D	1.62	2º

Convencionalmente consideraremos como positiva o directa la ordenación que se haga en el sentido de los valores crecientes de la característica que sirve como base a la ordenación, y como negativa la que se hace en el sentido de los valores decrecientes de la característica.

Ordenación Negativa o Inversa

Individuos	Medidas	Ordenes
A	1.53	4º
B	1.90	1º
C	1.73	2º
D	1.62	3º

Sin embargo, una "ordenación" puede no ser simplemente el resultado de dar un orden a los individuos de un conjunto de acuerdo con una única característica, sino que puede ser el resultado que se obtenga de arreglarlos tomando como base la evaluación de las posiciones de esos individuos, hecha por un juez que tome en consideración no una característica mensurable única, sino:

1.—un complejo de características mensurables o
2.—una cualidad o un complejo de cualidades que no se prestan a medición precisa, sino a simple evaluación.

Así, a un juez puede pedírsele que ordene a determinados individuos que practican ocupaciones distintas entre sí (y que incluso pueden tener unos mismos ingresos, un mismo número de años de instrucción y otros varios caracteres mensurables) de acuerdo con la posición más o menos alta que, a su juicio, les corresponda en la sociedad. Esto constituye una de las prácticas más generalizadas en la construcción de las llamadas escalas sociométricas. Análogamente, también suelen construirse escalas psicométricas parecidas a partir del mismo procedimiento.

De este modo, puede obtenerse una ordenación como la siguiente:

Ordenación hecha por el juez K de las posiciones de cuatro individuos (W, X, Y, Z)

Individuos	Posiciones
W	4ª
X	2ª
Y	3ª
Z	1ª

Una "coordenación" estadística busca establecer el grado de acuerdo (o desacuerdo) que existe entre dos ordenaciones estadísticas hechas:

1.—Tomando como base distintos criterios de ordenación (o sea, basadas en diferentes características de los individuos del conjunto),

2.—Por diferentes jueces.

Así, por ejemplo, una "coordenación" estadística puede buscar el grado de acuerdo entre las dos ordenaciones estadísticas siguientes:

Ordenación A			*Ordenación B*	
Individuos	Trabajos publicados	Orden	Individuos	Orden de los puestos ocupados por ellos
K	7	3º	K	4º
L	8	4º	L	2º
M	4	1º	M	1º
N	5	2º	N	3º

Tal coordenación trataría de establecer hasta qué punto, en el centro de trabajo de que se tratara, la posición ocupacional de los individuos estaba de acuerdo con el *volumen* de la producción o en el que este volumen de trabajos publicados respondería a la posición ocupacional de quien los hubiese publicado.

Otra "coordenación" estadística podría buscar el grado de acuerdo entre diferentes jueces a los que se les hubiese pedido que clasificaran a un conjunto de individuos dentro de una escala de posiciones sociales.

Ejemplificativamente, pudiera ser que se tuvieran las siguientes ordenaciones de las posiciones ocupadas por los individuos D, F. G. H. de un conjunto, de acuerdo con el criterio de los jueces P, Q, R.

Individuos	Ordenación según el juez P	Ordenación según el juez Q	Ordenación según el juez R
D	3º	3º	2º
F	1º	2º	4º
G	2º	1º	1º
H	4º	4º	3º

Indice de coordenación.—Para diseñar un índice de coordenación muy sencillo, tendremos en consideración lo siguiente:

1.—Las ordenaciones pueden ser de dos o más objetos,
2.—Las ordenaciones pueden hacerse de acuerdo con: dos o más criterios, o sea, considerando dos o más características de cada objeto (una característica para cada ordenación), o los criterios de dos o más jueces.

La existencia de estas posibilidades permite hablar siguiendo un paralelismo estrecho con las líneas seguidas en el estudio de la correlación, de:

1.—Coordenaciones simples (o sean coordenaciones entre ordenaciones de dos o más objetos, pero hechas de acuerdo con *sólo dos criterios* o por sólo dos jueces),

Teóricamente, al menos, pueden pensarse no sólo en una coordenación parcial, sino, incluso, en dos tipos de coordenación parcial:

3.—Coordenación parcial,
con eliminación de la influencia que tiene la presencia de un objeto en la ordenación de los restantes, o

4.—Coordenación parcial,
con eliminación de la influencia que un criterio de ordenación o un juez ordenados tienen sobre las ordenaciones hechas de acuerdo con otros criterios o por otros jueces.

En ésta, que no es sino una iniciación a los métodos de correlación (a la que en forma apendicular hemos agregado estas notas introductorias de los procedimientos de coordenación) trataremos del problema central de la coordenación múltiple.

Coordenación múltiple.—Tomemos el caso de dos objetos que han de clasificar cuatro jueces. Designemos por A y B los dos objetos y por P, Q, R, S a los jueces ordenadores. Las situaciones típicas que pueden producirse son las siguientes:

	A	B	A	B	A	B
P	1	2	1	2	2	1
Q	1	2	1	2	2	1
R	1	2	2	1	2	1
S	1	2	2	1	2	1
	Acuerdo positivo $+1$		Bipartición		Acuerdo negativo -1	

	A	B	A	B
P	1	2	2	1
Q	1	2	2	1
R	1	2	2	1
S	2	1	1	2
	+3/4		−3/4	

Es fácil ver que hemos llamado acuerdo positivo aquel en el cual el acuerdo se establece sobre la base de una forma de ordenación que —por cualquier motivo— puede considerarse como directa o positiva, y que llamamos acuerdo negativo aquel en el cual si bien todos los jueces están de acuerdo en el rango que hay que asignar a cada uno de los dos objetos, estos rangos guardan entre sí una relación que —por cualquier motivo específico— pueda considerarse como inversa de la natural.

Supongamos, ahora, que contamos no con dos, sino con tres objetos que deben ordenar los cuatro jueces de nuestro ejemplo, y que designamos tales objetos por A, B, C. En estas condiciones, es posible descomponer la ordenación de los tres objetos en tres ordenaciones de dos objetos:

A B   A C   B C

En cada una de estas tres coordenaciones, pueden darse las situaciones típicas de acuerdo positivo o negativo o de bipartición que hemos señalado anteriormente, y, por supuesto el acuerdo positivo o negativo puede alcanzar diferentes grados (en el caso de cuatro jueces $(1 = 4/4$ y $3/4$; $-1 = -4/4$ y $-3/4)$.

Las posibilidades de combinación de los acuerdos, biparticiones y desacuerdos de estas tres coordenaciones bipartitas quedan esquematizadas, en forma típica en la tabulacioncita siguiente, en la cual A indica acuerdo positivo, B bipartición y N acuerdo negativo. Para mayor facilidad, sólo se han tomado en consideración los valores máximos de acuerdo (1) y de acuerdo negativo (−1), pero, por supuesto, pueden usarse los valores intermedios (en el caso 3/4 y — 3/4) a los que, naturalmente corresponderán cifras totales intermedias entre las registradas en la última columna del cuadro.

Los valores de esa última columna del cuadro se obtienen sumando los índices correspondientes a cada arreglo bipartita y dividiendo el resultado entre el número de objetos (3 en este caso).

El cuadro es el siguiente:

Posibilidades	Primera coordenación bipartita	Segunda coordenación bipartita	Tercera coordenación bipartita	Indice de coordenación tripartita
1ª posibilidad típica	A(1)	A(1)	A(1)	3/3
2ª posibilidad típica	A(1)	A(1)	B(0)	2/3
3ª posibilidad típica	A(1)	B(0)	B(0)	1/3
4ª posibilidad típica	B(0)	B(0)	B(0)	0/3
5ª posibilidad típica	B(0)	B(0)	N(—1)	—1/3
6ª posibilidad típica	B(0)	N(—1)	N(—1)	—2/3
7ª posibilidad típica	N(—1)	N(—1)	N(—1)	—3/3

A fin de formular simbólicamente la estructura de este índice de coordenación de máxima simplicidad, recurriremos a las siguientes representaciones:

Número de jueces que sostienen el orden directo a natural	N
Número de jueces que sostienen el orden inverso o contrario	C
Número total de jueces	J

$$N + C = J$$

El grado de acuerdo bipartita estará dado por:

$$\frac{(1)\,N + C(-1)}{J}$$

Número de pares de objetos: P

Grado de coordenación:

$$\frac{\dfrac{N_1(1) + C_1(-1)}{J} + \dfrac{N_2(1) + C(-1)}{J} + \dfrac{N_3(1) + C_3(-1)}{J} + \ldots}{P}$$

O bien:

$$\frac{(N_1 + N_2 + N_3 + \ldots) + C_1 + C_2 + C_3)\ (-1)}{JP}$$

O sea, que, finalmente, el índice de concordancia estará dado por:

$$\frac{\Sigma N_i + \Sigma\, C_i\ (-1)}{JP}$$

I N D I C E

SE TERMINO DE IMPRIMIR ESTE LIBRO EL DIA 21 DE NOVIEMBRE DE 1962, EN LOS TALLERES DE LA EDITORIAL CVLTVRA, T. G., S. A., AV. REPUBLICA DE GUATEMALA NUM. 96, DE LA CIUDAD DE MEXICO, D. F. SIENDO SU TIRAJE DE 1.000 EJEMPLARES.

EL CUIDADO DE LA EDICION ESTUVO A CARGO DE LOS SRES. JOSE MARIA AVILES Y DEL AUTOR. INTERVINIERON EN SU COMPOSICION: LINOTIPISTA, MIGUEL SARDANETA, JR.; CAJISTA, MIGUEL ANGELES A. DIRECCION TECNICA: PORFIRIO LOERA Y CHAVEZ.

URIBE VILLEGAS

EL A. B C. DE LA
CORRELACION
Y SUS
APLICACIONES
SOCIALES