ESTUDIO COMPARATIVO DE LOS COEFICIENTES DE CORRELACION OBTENIDOS EN TABLAS DE CONTINGENCIA DE 2 × 2 (I) I. MUESTRAS DEL MISMC TAMAÑO

por O. LEON y F. J. TEJEDOR,
Universidad Autónoma y Complutense de Madrid

Intentamos ofrecer en este artículo una panorámica general sobre la problemática que entraña la interpretación de los diversos coeficientes de correlación que pueden ser calculados en las tablas de contingencia de 2 × 2.

Bien es sabido que trabajando con variables nominales el investigador de ciencias humanas se preocupa tanto por detectar posibles diferencias significativas entre las frecuencias de las diferentes categorías de las variables estudiadas (lo que puede hacerse con el cálculo e interpretación de ji-cuadrado) como por conocer la dependencia/independencia que puede existir entre las categorías de la variable; dependencia/independencia que podemos expresar mediante diferentes coeficientes de correlación.

La preocupación por conocer esta relación entre las variables no sólo es legítima sino que diríamos que es casi necesaria por aquéllo de que el estadístico ji-cuadrado es muy sensible el tamaño de la muestra: en virtud de la propiedad multiplicativa —si todas las frecuencias empíricas de una tabla de contingencia se multiplican por una constante, el ji-cuadrado queda multiplicado por dicha constante— es directamente proporcional al tamaño de la muestra cuando entre las frecuencias de ambas tablas existe una constante de proporcionalidad.

Los diferentes coeficientes de correlación vienen a paliar, de alguna manera, la influencia tan decisiva del tamaño de la muestra. Y así, para tablas con frecuencias proporcionales se obtendrá un mismo valor de correlación.

Pero el hecho de que para unos mismos datos puedan obtenerse coeficientes de correlación distintos (cuantitativa y cualitativamente) nos hace pensar que al investigador

en ciencias humanas, que en algún momento puede decidir la utilización de esta metodología estadística por propia voluntad o por imposibilidad de sobrepasar el nivel de medición nominal, pueden serle útiles estas reflexiones sobre las peculiaridades y comportamiento funcional de los diferentes coeficientes de correlación.

Los coeficientes sobre los que vamos a trabajar son los siguientes:

1. Coeficiente de correlación fi:

$$\varphi = \sqrt{\frac{X^2}{N}}$$

donde $X^2 = \Sigma \left[\frac{(f_e - f_t)^2}{f_t} \right]$

es la función "ji-cuadrado" utilizada en estadística no paramétrica y no la función de probabilidad "ji-cuadrado" de Pearson definida por $X^2_n = X^2_1 + X^2_2 + \ldots + X^2_n$, con $X = N(0,1)$; aquélla no es más que una aproximación de ésta a la que tiende asintóticamente a medida que la muestra es grande y las observaciones independientes.

En adelante siempre que mencionamos a "ji-cuadrado" nos referimos a la función utilizada en estadística no paramétrica.

En las fórmulas que hemos ofrecido (y en las que ofrezcamos a lo largo de este trabajo):

f_e = frecuencia empírica de cada casilla de la tabla de contingencia.
f_t = frecuencia teórica de cada casilla.
N = suma total de frecuencias de todas las casillas.

2. Coeficiente de contingencia verdadero:

$$C_v = \frac{\text{obtenido}}{C \text{ máximo}}$$

donde, $C_{\text{obtenido}} = \sqrt{\frac{X^2}{X^2 + N}}$

$$C_{\text{máximo}} = \sqrt{\frac{k}{k - 1}} = \sqrt{\frac{1}{2}}$$

(ya que $k = 2$ por tratarse del número de categorías de la variable en una tabla de 2×2, único caso que tratamos e este artículo).

3. Coeficiente O corregido (*):

$$O_c = \frac{O_{obtenido}}{O_{máximo}} = \frac{O_{obtenido}}{1} = O_{obtenido}$$

$$\text{donde, } O_{obtenido} = \frac{\Sigma\, |f_e - f_t|}{N}$$

$$O_{máximo} = 1 \text{ (en tablas } 2 \times 2).$$

4. Coeficiente de asociación:

$$Q = \frac{AD - BC}{AD + BC}$$

correspondiendo las letras a las casillas siguientes:

A	B
C	D

5. Coeficiente de coligación:

$$W = \frac{\sqrt{AD} - \sqrt{BC}}{\sqrt{AD} + \sqrt{BC}}.$$

(*) "O" es un estadístico propuesto por los autores de este artículo como alternativa al uso de ji-cuadrado y que viene definido por

$$O = \frac{\Sigma\, f_e - f_t}{N}$$

es decir, por la suma del valor absoluto de la diferencia entre las frecuencias empíricas y las teóricas de cada una de las casillas de una tabla de contingencia dividida por N.

El estadístico "O" está aún hoy en proceso de estudio: tratamos de conocer su comportamiento en las diferentes situaciones empírico-experimentales y de conocer su distribución muestral.

Las primeras conclusiones referidas a su posible utilización en la investigación con variables nominales se ofrecen en la Tesis Doctoral realizada por O. León y dirigida por F.J. Tejedor: "Tratamiento de variables nominales en psicología. Nuevas aportaciones". Departamento de diagnóstico Psicológico y Medida. Facultad de Filosofía y Letras. Universidad Autónoma de Madrid, septiembre de 1980.

6. Coeficiente coseno de Pearson:

$$r_c = \cos. \frac{\sqrt{BC}}{\sqrt{AD} + \sqrt{BC}} \pi$$

No debe confundirse el r_c con el coeficiente de correlación de Pearson, r_{xy}; si nos hemos referido a r_{xy} es porque, como es sabido, el coeficiente φ no es más que una mera aplicación de r_{xy} para el caso de variables dicotómicas; por tanto, en tablas de 2×2, la igualdad $\varphi = r_{xy}$ se produce siempre; nosotros creemos más apropiado referirnos a φ cuando trabajamos con variables nominales dicotómicas en tablas de contingencia de 2×2.

Algunas observaciones que conviene tener en cuenta y que no vamos más que a enunciar porque su justificación puede encontrarse en cualquier texto elemental de estadística, serían:

—Constancia de r_{xy} (y por tanto de φ) cuando multiplicamos todas las puntuaciones por una constante.

—Constacia de r_{xy} (y por tanto de φ) cuando sumamos una constante a todas las puntuaciones.

—Igualdad $r_{xy} = \varphi$ para puntuaciones 0,1 y, en virtud de las observaciones anteriores, para cualquier tipo de puntuaciones.

—Constancia de los coeficientes de correlación en las tablas de contingencia de 2×2 frente a N.

La importancia de esta observación nos aconseja presentarlo para los diferentes coeficientes (hemos elegido sólo dos en la seguridad de que el lector puede generalizar la explicación a los demás); el subíndice 1 lo empleamos para mostrar la modificación de multiplicación de las f_e y de N por una constante k:

$$\varphi = \sqrt{\frac{X^2}{N}}$$

$$\varphi_1 = \sqrt{\frac{k\,X^2}{k\,N}} = \frac{\sqrt{k}\ \sqrt{X^2}}{\sqrt{k}\ \sqrt{N}} = \frac{\sqrt{X^2}}{\sqrt{N}} = \varphi$$

$$Q = \frac{AD - BC}{AD + BC}$$

$$Q_1 = \frac{KAKD - KBKC}{KAKD + KBKC} = \frac{K^2\,(AD - BC)}{K^2\,(AD + BC)} = \frac{AD - BC}{AD + BC} = Q$$

Queda por tanto demostrado que para una relación dada, si aumentamos el tamaño de

la muestra manteniendo dicha relación, el valor de los distintos coeficientes de correlación no varía; podemos por tanto elegir cualquier tamaño de N como base para el estudio de simulación experimental.

Estamos ya en condiciones de presentar nuestros trabajos sobre la variación de los diferentes coeficientes de correlación. Y lo haremos en dos partes: la primera parte, referida al caso en el que las muestras (totales marginales de las filas de la tabla de contingencia) son iguales y la segunda referida al caso en el que las muestras tengan tamaños diferentes.

Por aquello de la adecuación a las exigencias de la revista, publicaremos cada una de las partes en artículos diferenciados. Como es lógico, aquí presentamos la primera parte.

1. Estudio de la variación de los Coeficientes de Correlación en tablas 2 × 2 con muestras del mismo tamaño

Estudiaremos como varían los distintos coeficientes de correlación en una tabla de contingencia de 2 × 2 cuando desde una situación de mínima discrepancia entre las f_e y las f_t hacemos variar progresivamente a aquéllas hasta llegar a la situación de máxima discrepancia.

Hemos elegido N = 80 porque, por una parte, es un tamaño suficientemente grande como para no tener necesidad de aplicar la corrección por cotinuidad de Yates, lo que complicaría innecesariamente los cálculos sin enriquecer el trabajo y, por otra, porque nos proporciona un número suficiente de casos posibles para poder trazar la gráfica de los diferentes valores correlacionales.

Consideraremos para nuestro estudio un ejemplo de dos muestras —hombres, mujeres— interesándonos por conocer su relación con la variable acepto/rechazo de la participación de los padres en la dirección del centro escolar al que acuden sus hijos.

Definiremos el parámetro DTE (diferencia total entre empíricas) como la suma de las diferencias, en valor absoluto, entre las frecuencias empíricas de ambas muestras. Señala este parámetro la discrepancia entre las frecuencias, siendo por tanto el parámetro a variar (de menor a mayor discrepancia).

Ofreceremos para cada situación los valores de X^2 y de los diferentes coeficientes de correlación.

La disposición de los datos en la tabla siempre seguirá las siguientes pautas de distribución:

	aceptación	rechazo
mujeres	A	D
hombres	C	B

N

figurando en cada casilla únicamente la frecuencia empírica; las frecuencias teóricas en una tabla de contingencia suelen colocarse en la correspondiente casilla, debajo de la empírica y

entre paréntesis. Nosotros omitiremos en todo este trabajo la referencia a las teóricas, si bien en todos los casos resultarían de dividir por N el producto de los totales marginales de su fila y de su columna.

Aunque en algún caso, por ser las frecuencias teóricas pequeñas, una mínima prudencia metodológica requeriría aplicar la corrección por continuidad de Yates, e incluso en algún caso la prueba exacta de Fisher, nosotros no tendremos en cuenta esta deficiencia puesto que nos interesa más conocer los valores que va tomando la función ji-cuadrado que la interpretación específica de una hipótesis. Podríamos haber decidido aplicar siempre la corrección, pero ya dijimos que eran más importantes los problemas de cálculo que creaba que las ventajas que suponía.

1.1. Variación uniforme

Por variación uniforme entendemos aquélla que se produce cuando la tabla es contínuamente simétrica respecto de sus diagonales. El primer caso representará una situación en la cual la no decantación de los datos respecto a la variable en estudio es absoluta. Por ejemplo, de 40 mujeres encuestadas sobre la participación de los padres en la dirección del colegio 20 estaban a favor y 20 en contra. La misma situación se observó en los 40 hombres: 20 a favor y 20 en contra. Sin necesidad de calcular ningún coeficiente de correlación estamos en condiciones de establecer que la variable sexo, respecto a la decisión escolar planteada, presenta una correlación nula. ¿Reflejan esto los distintos coeficientes?

La situación de máxima discrepancia será aquélla en que las 40 mujeres estuvieran, por caso, en contra de la participación, mientras que los 40 hombres fueran partidarios de la idea contraria. La dependencia de la opinión con respecto al sexo es ahora total. La correlación debe ser máxima. ¿Lo reflejan así los distintos coeficientes?

Efectivamente, en el primer caso todos valen 0 y en segundo 1. Respecto a estas dos situaciones extremas no hay dudas. Empiezan a surgir cuando la situación es intermedia: 30 mujeres a favor y 10 en contra frente a 30 hombres en contra y 10 a favor. ¿Señalarían ahora los coeficientes una correlación media? Ya veremos que no ocurre así y que en los valores que proporcionan y su forma de variar depende de cada función de correlación.

Presentamos el desarrollo de cuatro casos en la seguridad de que el lector podrá verificar por su cuenta si lo desea el conjunto total de datos que ofrecemos en la tabla 1.
caso 1: mínima discrepancia

20		
20	20	
40	40	80

$DTE = 20 - 20 + 20 - 20 = 0$

$X^2 = \varphi = C_r = Q_c = Q = W = r_c = 0$

caso 2

21	19	40
19	21	40
40	40	80

DTE = 21 — 19 + 19 — 21 = 4
$X^2 = 0,2$ $\varphi = 0,05$
$C_v = 0,0706$ $O_c = 0,05$
$Q = 0,0907$ $W = 0,05$
$r_c = 0,0784$

caso 11

30	10	40
10	30	40
40	40	80

DTE = 40 $\varphi = 0,5$
$X^2 = 20$ $O_c = 0,5$
$C_v = 0,6324$ $W = 0,5$
$Q = 0,80$
$r_c = 0,7071$

caso 21: máxima discrepancia

40	0	40
0	40	40
40	40	80

DTE = 80
$X^2 = 80$
$\varphi = C_v = O_c = Q = W = r_c = 1$

La tabla completa de valores (tabla 1) para el conjunto de casos posibles será:

Tabla 1

caso	DTE	X^2	φ	C_v	O_c	Q	W	r_c
1	0	0	0	0	0	0	0	0
2	4	0,2	0,05	0,0706	0,05	0,0907	0,05	0,0784
3	8	0,8	0,10	0,1407	0,10	0,1980	0,10	0,1564
4	12	1,8	0,15	0,2097	0,15	0,2934	0,15	0,2334
5	16	3,2	0,20	0,2773	0,20	0,3846	0,20	0,3090
6	20	5	0,25	0,3429	0,25	0,4706	0,25	0,3827
7	24	7,2	0,30	0,4063	0,30	0,5504	0,30	0,4540
8	28	9,8	0,35	0,4671	0,35	0,6238	0,35	0,5225
9	32	12,8	0,40	0,5223	0,40	0,6896	0,40	0,5878
10	36	16,2	0,45	0,5804	0,45	0,7484	0,45	0,6494
11	40	20	0,50	0,6324	0,50	0,80	0,50	0,7071

caso	DTE	X^2	φ	C_v	O_c	Q	W	r_c
12	44	24,2	0,55	0,6815	0,55	0,8445	0,55	0,7604
13	48	28,8	0,60	0,7277	0,60	0,8823	0,60	0,8090
14	52	33,8	0,65	0,7707	0,65	0,9139	0,65	0,8526
15	56	39,2	0,70	0,8111	0,70	0,9396	0,70	0,8910
16	60	45	0,75	0,8485	0,75	0,96	0,75	0,9239
17	64	51,2	0,80	0,8835	0,80	0,9756	0,80	0,9510
18	68	57,8	0,85	0,9158	0,85	0,9869	0,85	0,9724
19	72	64,8	0,90	0,9461	0,90	0,9945	0,90	0,9877
20	76	72,2	0,95	0,9738	0,95	0,9987	0,95	0,9969
21	80	80	1	1	1	1	1	1

Observando la tabla notamos que, en todos los casos, se produce la igualdad $\varphi = Oc = W$. Efectivamente, tiene que producirse esta igualdad ya que, mediante unas sencillas transformaciones en las fórmulas de estos coeficientes podemos comprobar que, en este tipo de tablas, con totales marginales iguales, se produce:

$$r_{xy} = \varphi = Oc = W = \frac{2\ (A - B)}{N}$$

siendo A la f_e mayor y B la menor

La representación gráfica de los valores de la tabla 1 nos proporciona la visión de conjunto que ofrecemos en la gráfica 1.

Como podemos apreciar en el eje de abscisas hemos colocado los sucesivos incrementos de valores DTE y en el eje de ordenadas los valores de correlación.

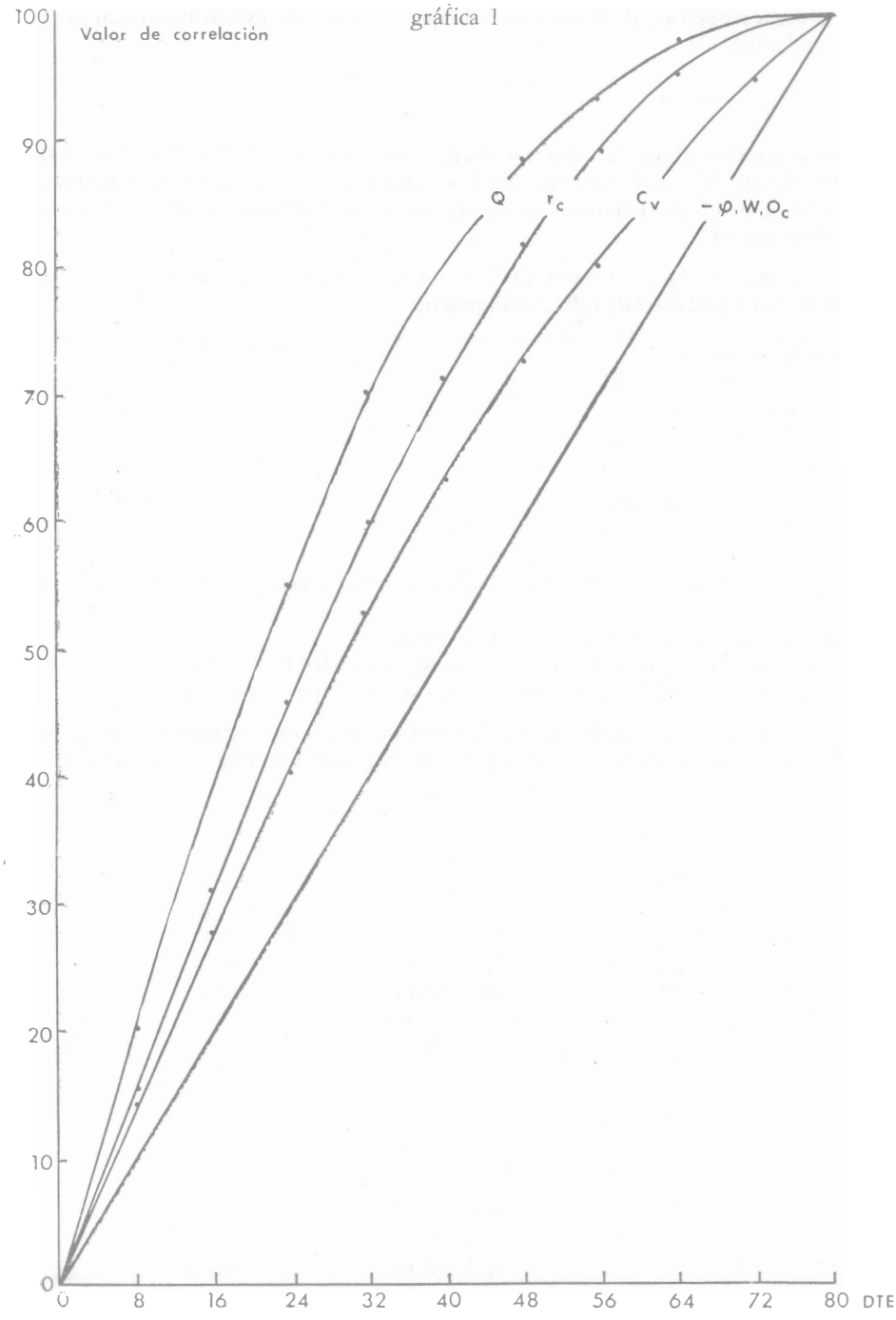
La gráfica descrita por los coeficientes φ, Oc y W corresponde a la ecuación lineal $y = ax$ puesto que para $x = O$, $y = O$ y su gráfica es una recta.

Resulta de interés considerar los valores de las funciones para el caso que habíamos considerado como de correlación media, es decir, en el caso en el que 30 mujeres estuvieran a favor y 10 en contra y los hombres viceversa, 30 en contra y 10 a favor. El valor intuitivamente esperado para la correlación debería ser 0,50. Es el caso 11 de la tabla 1. Podemos comprobar que φ, Oc y W indican el valor esperado mientras que C_v se aleja sensiblemente; r_o y Q lo hacen de una forma tan notable que nos invitan desde ahora a pensar en su falta de fiabilidad.

Tanto C_v como Q y r_c describen curvas por encima de la recta de φ. En el caso de C_v es más inflada en la zona media lo cual supone que si los valores de φ, C_v y W son los mejor considerados, su zona crítica o menos fiable se halla en torno a los valores DTE medios y se aproxima más a los valores hipotéticos en las situaciones de valores DTE extremos.

El coeficiente Q presenta sin embargo su zona más inflada en los valores DTE del último tercio; es en esta parte donde su fiabilidad se hace por tanto menor. En cualquier caso la media de discrepancia respecto a los valores de la recta es lo suficientemente alta

gráfica 1


Valor de correlación
100
90
80
70
60
50
40
30
20
10
0
Q
r_c
C_v
$-\varphi, W, O_c$
0
8
16
24
32
40
48
56
64
72
80
DTE

como para desaconsejarlo; al menos eso es lo que se deduce de esta situación empírico-experimental simulada.

1.2. Variación no uniforme

Vamos a estudiar ahora la variación cuando en tablas de 2 × 2 se fijan los totales marginales de las filas (los tamaños de las muestras), se modifican las frecuencias empíricas hasta la situación de máxima discrepancia y se mantiene el valor DTE en una cantidad determinada.

Elegimos, por ejemplo, el valor DTE = 44 (caso número 12 de la tabla 1) que corresponde a la siguiente tabla de contingencia:

31	9	40
9	31	40
40	40	80

Las sucesivas modificaciones sobre la tabla se determinan por las tres condiciones siguientes:

—mantener iguales los totales de las muestras.
—buscar diferencias progresivamente mayores entre las frecuencias, y
—mantener el valor DTE constante (en nuestro ejemplo DTE = 44).

La aplicación de estas condiciones a la tabla elegida como ejemplo (caso 1) nos proporciona diversos casos de los que exponemos alguno recogiendo la totalidad de la información a ellos referida en la tabla 2.

caso 1

31	9	40
9	31	40
40	40	80

DTE = 44
$X^2 = 24{,}2$
$C_c = 0{,}6815$
$Q = 0{,}8445$
$r_c = 0{,}7604$

$\varphi = 0{,}55$
$Oc = 0{,}55$
$W = 0{,}55$

caso 2

8	32	40
30	10	40
38	42	80

DTE = 44
$X^2 = 24{,}2606$
$C_c = 0{,}6822$
$Q = 0{,}8467$
$r_c = 0{,}7624$

$\varphi = 0{,}5507$
$Oc = 0{,}55$
$W = 0{,}5520$

caso 3

29	11	40
7	33	40
36	44	80

DTE = 44
X^2 = 24,4444
C_v = 0,6842
Q = 0,8511
r_c = 0,7685

φ = 0,6159
Oc = 0,55
W = 0,5580

caso 10

22	18	40
0	40	40
22	58	80

DTE = 44
X^2 = 30,3448
C_v = 0,7416
Q = 1
r_c = 1

φ = 0,6159
O_c = 0,55
W = 1

Tabla 2

caso	DTE	X^2	φ	C_v	Oc	Q	W	r_c
1	44	24,2	0,55	0,6815	0,55	0,8445	0,55	0,7604
2	44	24,2606	0,5507	0,6822	0,55	0,8467	0,5520	0,7624
3	44	24,4444	0,5528	0,6842	0,55	0,8511	0,5580	0,7685
4	44	24,7570	0,5563	0,6875	0,55	0,8594	0,5686	0,7791
5	44	25,2083	0,5613	0,6923	0,55	0,8713	0,5844	0,7944
6	44	25,8133	0,5680	0,6985	0,55	0,8871	0,6071	0,8154
7	44	26,5934	0,5766	0,7064	0,55	0,9071	0,6386	0,8431
8	44	27,5783	0,5871	0,7160	0,55	0,9322	0,6845	0,8797
9	44	28,8095	0,60	0,7278	0,55	0,9628	0,7598	0,9286
10	44	30,3448	0,6159	0,7416	0,55	1	1	1

En la gráfica 2 representamos los valores de la tabla 2; sobre la abscisa llevamos los valores de la diferencia entre los totales marginales no iguales (diferencia entre los totales marginales de las columnas de la tabla de contingencia).

Nos parece que lo más interesante a destacar en la gráfica 2 sería:

— φ y C_v describen parábolas ascendentes lentamente, casi paralelas, con diferencias similares de principio a fin de trazado.

—Q, W y r_c presentan un ascenso muy rápido, con su máximo en 1, con distinto punto de arranque por lo que resulta mucho más rápido el ascenso de W (a destacar que se ha perdido la similitud entre W y φ que aparecía en la gráfica 1).

Es la naturaleza de las funciones de Q, W y r_c lo que determina ese ascenso tan rápido: cuando alguna de las frecuencias tiende a cero estos coeficientes tienden rápidamente a 1; así en las tablas

33	33
0	29

66	0
0	29

se produce que $Q = W = r_c = 1$, y sin embargo no parece igual la relación entre las frecuencias de las casillas.

—Es bastante curioso el comportamiento de Oc en la gráfica 2: valor constante. No es fácil la explicación de este suceso (puede verse en las páginas 399-403 del trabajo original mencionado) pero en este momento es suficiente mostrar que Oc es el más parecido a φ, que en principio es el más "fiable" por aquéllo de que es el mismo valor r_{xy}.

1.3. Variación uniforme de las frecuencias de una muestra manteniendo la otra en situación de máxima diferencia.

Mantenemos la muestra de mujeres en situación de máxima diferencia y vamos variando paulatinamente las frecuencias en el grupo de hombres (de menor a mayor diferencia). La tabla base de partida (caso 1) será:

20	20	40
0	40	40
20	60	80

que como vemos refleja la situación de mínima diferencia en el grupo de hombres y de diferencia máxima en el grupo de mujeres. Veamos cómo se irían originando los diferentes casos posibles.

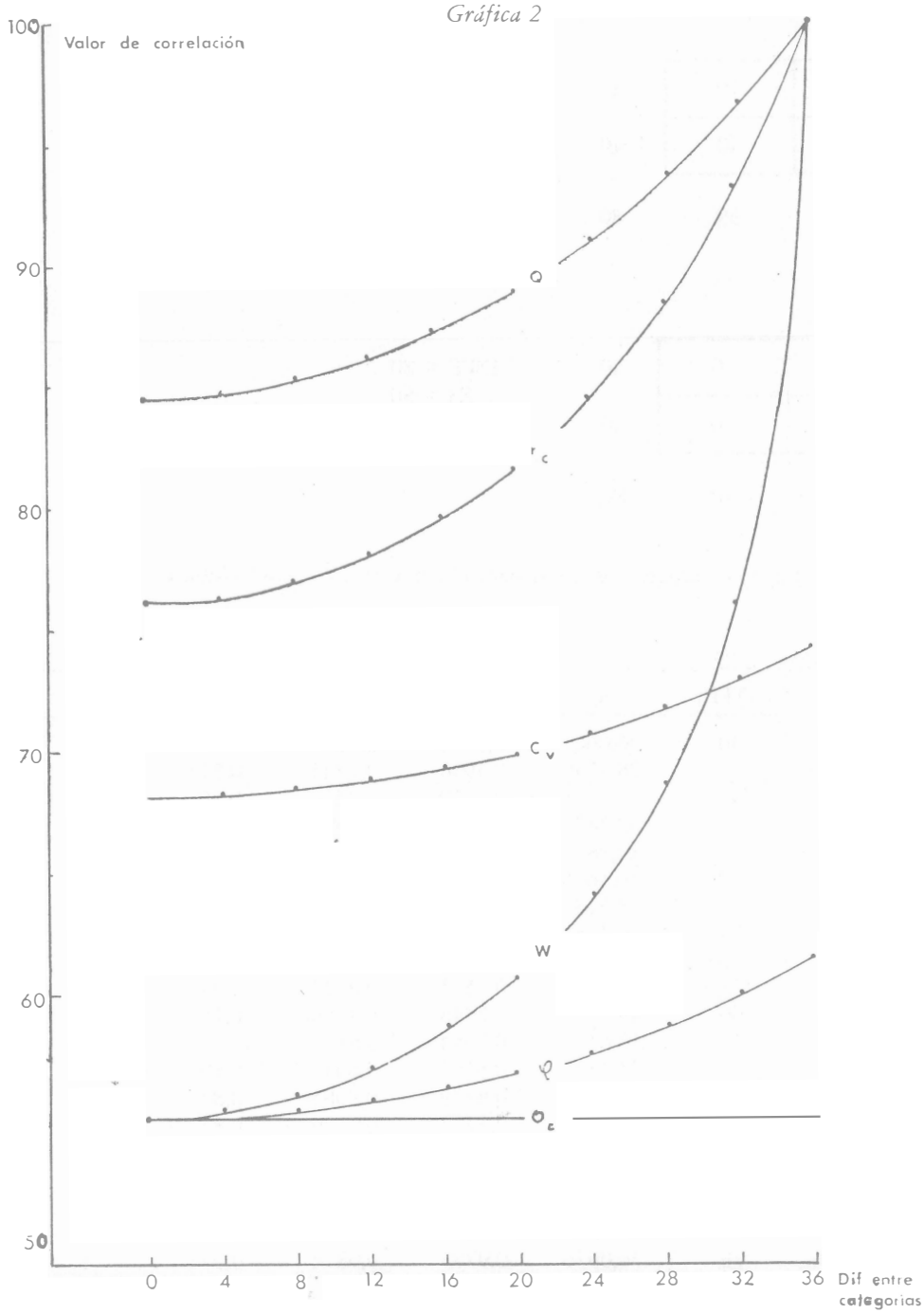
caso 1

20	20	40
0	40	40
20	60	80

DTE = 40
$X^2 = 26,6667$
$C_v = 0,7071$
$Q = W = r_c = 1$

$\varphi = 0,5773$
Oc = 0,50

Gráfica 2


100
90
80
70
60
50
Valor de correlación
Q
r_c
C_v
W
φ
O_c
0
4
8
12
16
20
24
28
32
36
Dif entre categorias

caso 2

21	19	40	DTE = 42
0	40	40	X^2 = 28,4746 φ = 0,5966
			C_v = 0,7245 Oc = 0,525
			Q = W = r_c = 1
21	59	80	

caso 21

40	0	40	DTE = 80
0	40	40	X^2 = 80
			$\varphi = C_v = Oc = Q = W = r_c = 1$
40	40	80	

En la tabla 3 se recogen los datos para el conjunto de casos posibles:

Tabla 3

caso	DTE	X^2	φ	C_v	Oc	Q = W = r_c
1	40	26,6667	0,5773	0,7071	0,50	1
2	42	28,4746	0,5966	0,7245	0,525	1
3	44	30,3448	0,6159	0,7416	0,55	1
4	46	32,2807	0,6352	0,7583	0,575	1
5	48	34,2857	0,6546	0,7746	0,60	1
6	50	36,3636	0,6742	0,7906	0,625	1
7	52	38,5185	0,6939	0,8062	0,65	1
8	54	40,7547	0,7137	0,8216	0,675	1
9	56	43,0769	0,7338	0,8367	0,70	1
10	58	45,4902	0,7541	0,8515	0,725	1
11	60	48	0,7746	0,8660	0,75	1
12	62	50,6122	0,7954	0,8803	0,775	1
13	64	53,3333	0,8165	0,8944	0,80	1
14	66	56,1702	0,8379	0,9085	0,825	1
15	68	59,1304	0,8597	0,9220	0,85	1
16	70	62,2222	0,8818	0,9354	0,875	1
17	72	65,4545	0,9045	0,9487	0,90	1
18	74	68,8372	0,9276	0,9618	0,925	1
19	76	72,3809	0,9512	0,9747	0,95	1
20	78	76,0976	0,9753	0,9874	0,975	1
21	80	80	1	1	1	1

En la gráfica 3 representamos todos los datos de la tabla 3, llevando sobre la abscisa los valores DTE y sobre la ordenada los valores de correlación, a partir de 0,50 ya que es el valor más pequeño que encontramos.

Observaciones a destacar, desde nuestro punto de vista, respecto de la gráfica 3 serían:

—Q, W y r_c no son sensibles a los cambios efectuados.

—φ, C_v y Oc parten de valores diferentes, pareciendo que es Oc el que mejor se adapta a la situación reflejada entre las frecuencias, ya que intuitivamente en el caso

20	20
0	40

la correlación debería estar próxima a 0,50 y ese es el valor de Oc; en el caso

30	10
0	40

el valor intuitivamente esperado para la correlación debería estar próximo a 0,75 y ese es también el valor de Oc.

—C_v describe una curva cóncava respecto al eje de abscisas mientras que φ describe una convexa. Esto casi no se puede apreciar en la gráfica 3, pero sí se observa claramente si se analizan los sucesivos incrementos de las dos funciones: si los incrementos son decrecientes la función es convexa.

La representación de incrementos sucesivos aparece en la gráfica 3.1.

1.4. Variación uniforme de las frecuencias de una muestra manteniendo la otra en situación de mínima diferencia.

Mantenemos la muestra de mujeres en situación de mínima diferencia e iremos variando paulatinamente las frecuencias en el grupo de hombres (de menor a mayor diferencia).

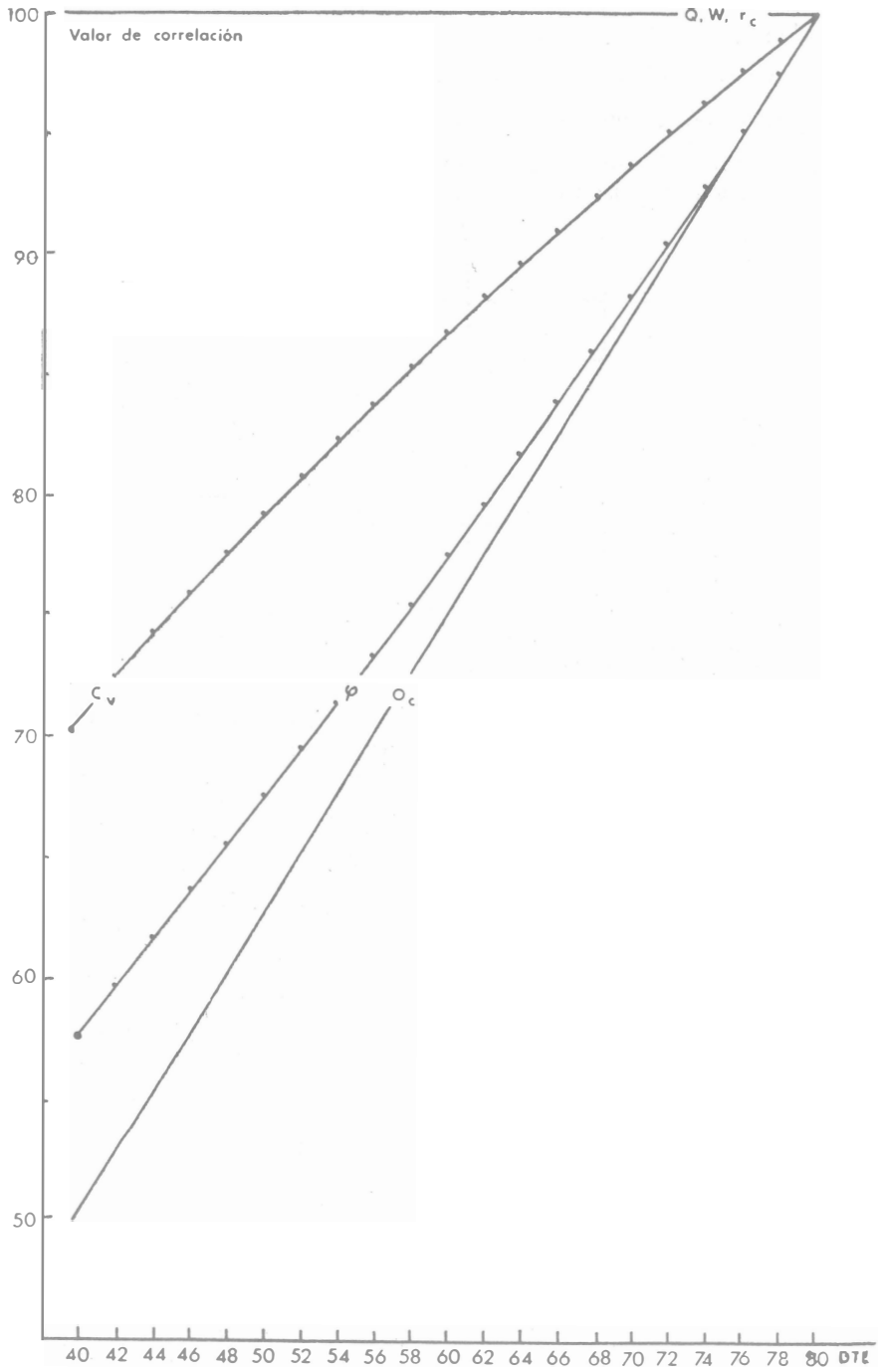
La tabla base de partida que refleja la situación anterior expuesta es la correspondiente al caso 1; los restantes casos se obtendrán, como siempre, por variaciones sucesivas.

caso 1

20	20	40
20	20	40
40	40	80

DTE = 0
$X^2 = 0$.
$\varphi = C_v = Oc = Q = W = r_c = 0$

Gráfica 3


Valor de correlación
100
90
80
70
60
50
Q, W, r_c
C_v
ϕ
O_c
40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 DTE

caso 2

21	19	40
20	20	40
41	39	80

DTE = 2
$X^2 = 0,05$ $\varphi = 0,025$
$C_v = 0,0353$ Oc = 0,025
Q = 0,05 W = 0,025
$r_c = 0,0393$

caso 21

40	0	40
20	20	40
60	20	80

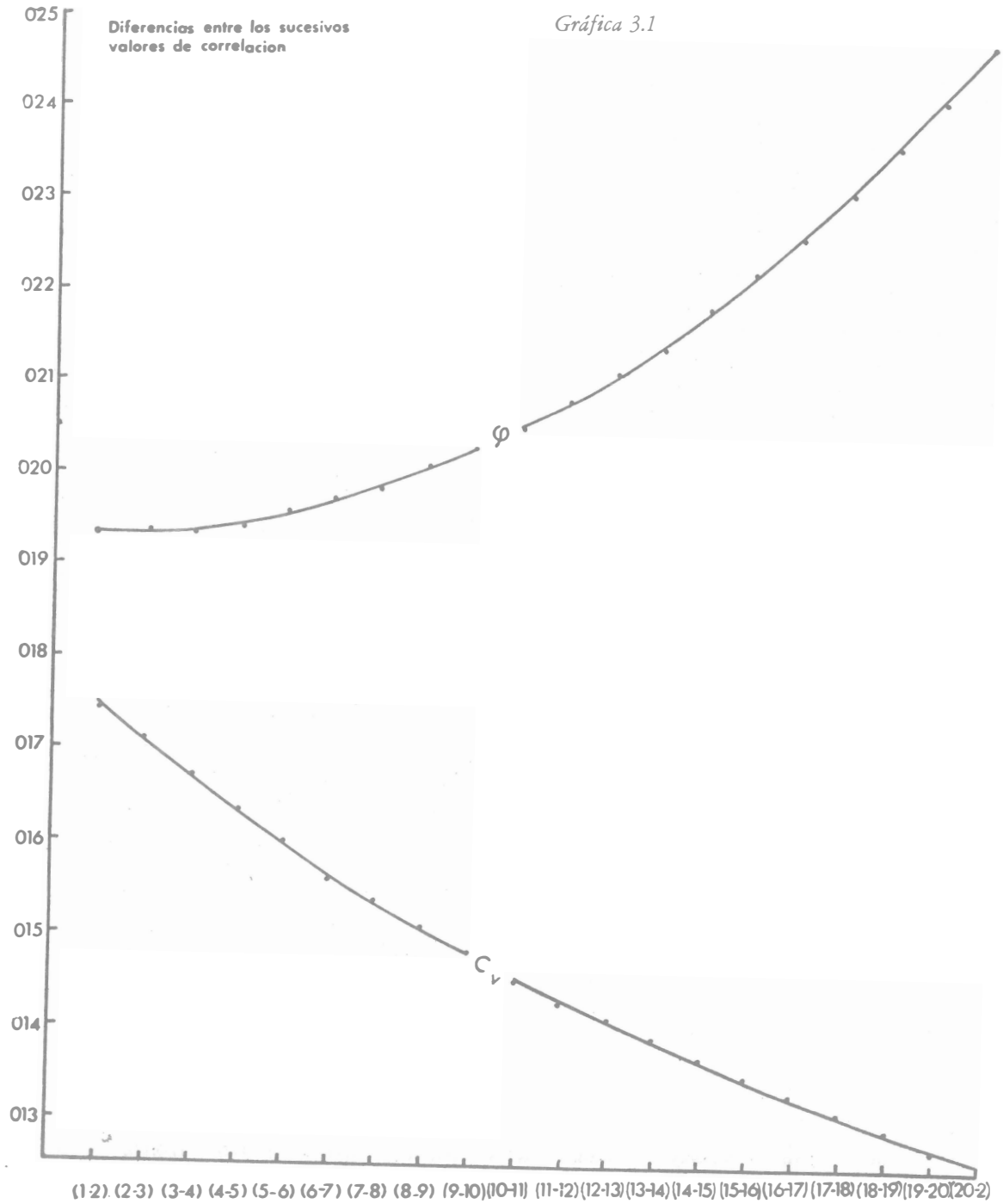
DTE = 40
$X^2 = 26,6667$ $\varphi = 0,5773$
$C_v = 0,7071$ Oc = 0,50
$Q = W = r_c = 1$

En la tabla 4 se recogen los datos para el conjunto de casos posibles.

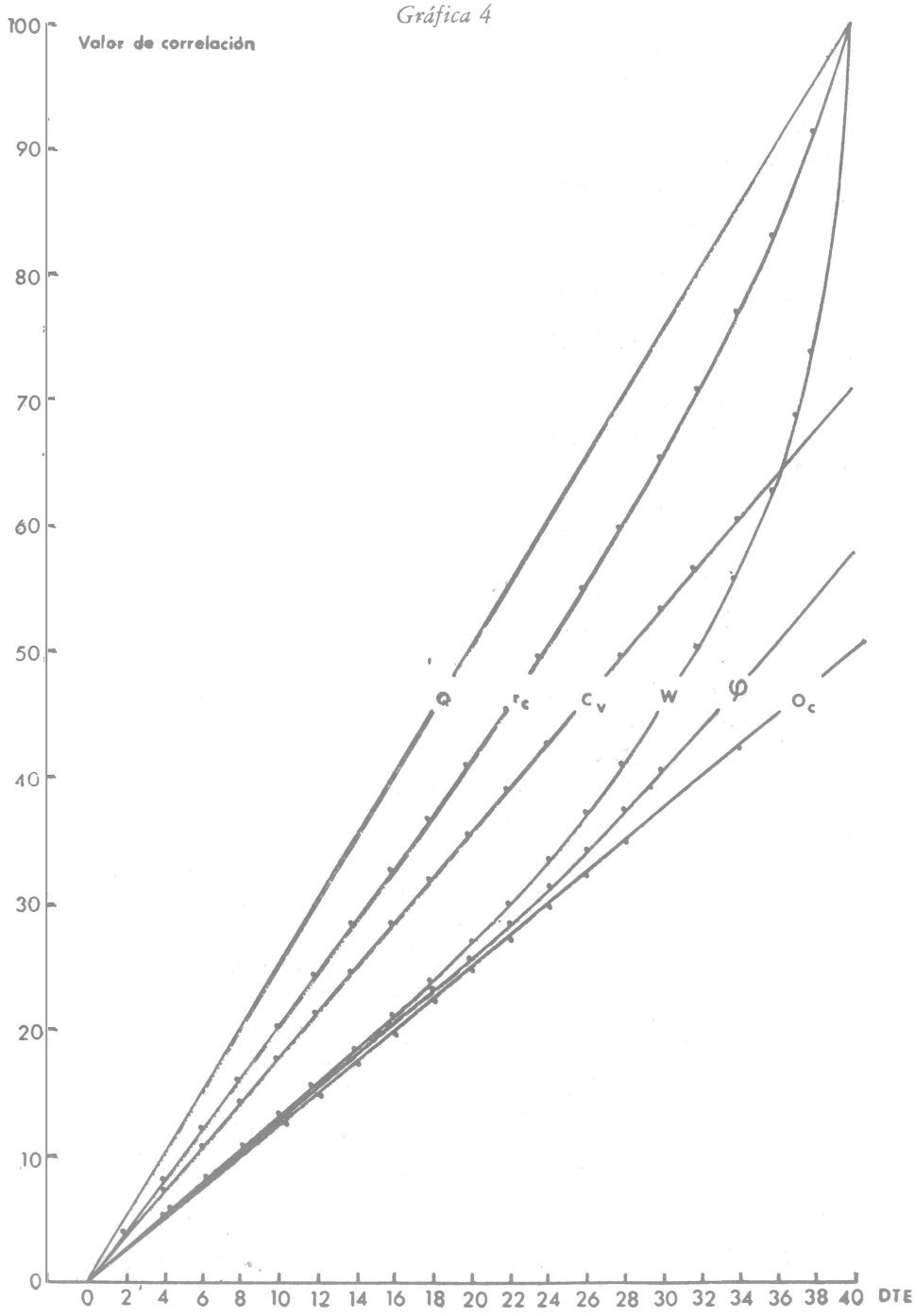
Tabla 4

caso	DTE	X^2	φ	C_v	Oc	Q	W	r_c
1	0	0	0	0	0	0	0	0
2	2	0,05	0,025	0,0353	0,025	0,05	0,025	0,0393
3	4	0,2005	0,0501	0,0707	0,0500	0,10	0,0501	0,0786
4	6	0,4525	0,0752	0,1061	0,075	0,15	0,0754	0,1182
5	8	0,8081	0,1005	0,1414	0,10	0,20	0,1012	0,1580
6	10	1,2698	0,1260	0,1768	0,125	0,25	0,1270	0,1982
7	12	1,8414	0,1517	0,2121	0,15	0,30	0,1535	0,2388
8	14	2,5274	0,1777	0,2475	0,175	0,35	0,1807	0,2801
9	16	3,3333	0,2041	0,2828	0,20	0,40	0,2087	0,3220
10	18	4,2660	0,2309	0,3182	0,225	0,45	0,2377	0,3648
11	20	5,3333	0,2582	0,3535	0,25	0,50	0,2679	0,4086
12	22	6,5450	0,2860	0,3839	0,275	0,55	0,2997	0,4536
13	24	7,9121	0,3145	0,4243	0,30	0,60	0,3333	0,50
14	26	9,4479	0,3436	0,4596	0,325	0,65	0,3693	0,5481
15	28	11,1681	0,3736	0,4950	0,35	0,70	0,4083	0,5984
16	30	13,0909	0,4045	0,5303	0,375	0,75	0,4514	0,6511
17	32	15,2381	0,4364	0,5657	0,40	0,80	0,50	0,7071
18	34	17,4635	0,4672	0,5986	0,425	0,85	0,5567	0,7672
19	36	20,3135	0,5039	0,6364	0,45	0,90	0,6268	0,8330
20	38	23,3091	0,5398	0,6718	0,475	0,95	0,7239	0,9074
21	40	26,6667	0,5773	0,7071	0,50	1	1	1

Gráfica 3.1


Diferencias entre los sucesivos valores de correlacion
025
024
023
022
021
020
019
018
017
016
015
014
013
φ
Cv
(1-2) (2-3) (3-4) (4-5) (5-6) (6-7) (7-8) (8-9) (9-10) (10-11) (11-12) (12-13) (13-14) (14-15) (15-16) (16-17) (17-18) (18-19) (19-20) (20-2)

Gráfica 4


Valor de correlación
100
90
80
70
60
50
40
30
20
10
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
DTE
Q
r_c
c_v
W
φ
o_c

En la gráfica 4 representamos todos los datos de la tabla 4, llevando sobre la abscisa los valores DTE y sobre la ordenada los valores de correlación.

Las observaciones que consideramos merecen destacarse son:

—Las funciones de Q, W y r_c presentan un crecimiento acelerado, debido a que el último caso tiene una casilla con frecuencia nula.

—El valor intuitivamente esperado para la correlación en el caso

40	0
20	20

vimos que debería estar próximo a 0,50 y ese es el valor que toma Oc, siendo φ el valor más cercano y Q, W y r_c los más alejados.

En el caso

30	10
20	20

la correlación intuitivamente esperada debería estar próxima a 0,25, que es el valor de Oc, siendo igualmente φ el más próximo.

—La representación conjunta de las gráficas 3 y 4 nos proporciona un curioso panorama de la tendencia de los diferentes valores de corrlación; representación conjunta que ofrecemos en la gráfica 4.1.

Existe como vemos una cierta continuidad en las funciones; continuidad que podemos expresar como simetría axial respecto a sendos ejes perpendiculares a la recta que contiene a Oc en el punto DTE = 40 de cada una de las funciones.

—La gráfica 4.1. nos permite alguna observación respecto a C_v y φ que, no olvidemos, son los más utilizados en la investigación en ciencias humanas por estar directamente basados en ji-cuadrado:

- cuando la correlación es perfecta o nula los diversos coeficientes toman efectivamente el valor 1 o cero.
- en los casos de correlación media, φ se aproxima más que C_v a los valores intuitivamente esperados.
- la máxima discrepancia de ambos coeficientes se produce cuando las diferencias de correlación entre los grupos son mayores.

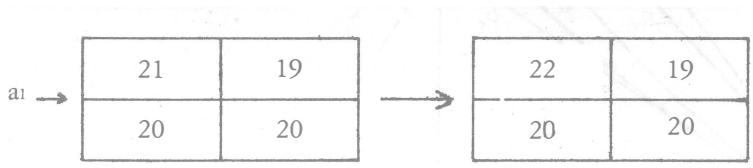
1.5. Conclusiones estadísticas

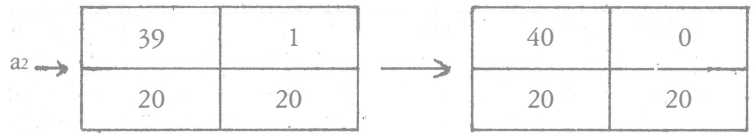
Como epílogo de este apartado (y por tanto de la primera parte del trabajo)

exponemos algunas conclusiones generales de carácter estadístico, a modo de resumen y síntesis de las que hemos venido expresando al plantear las distintas situaciones:

a) Defendemos que una prueba de correlación o de independencia debe reflejar por igual los incrementos de dependencia o correlación independientemente de la cuantía de la diferencia $|f_e - f_t|$. La prueba Oc ha sido pensada para responder a esta situación y como consecuencia se muestra independiente al aumento de tamaño.

La prueba de ji-cuadrado, y φ en la medida en que depende de ella, potencian mucho la diferencia $|f_e - f_t|$ debido a su característica cuadrática. Por ejemplo, entre las parejas de tablas a_1 y a_2

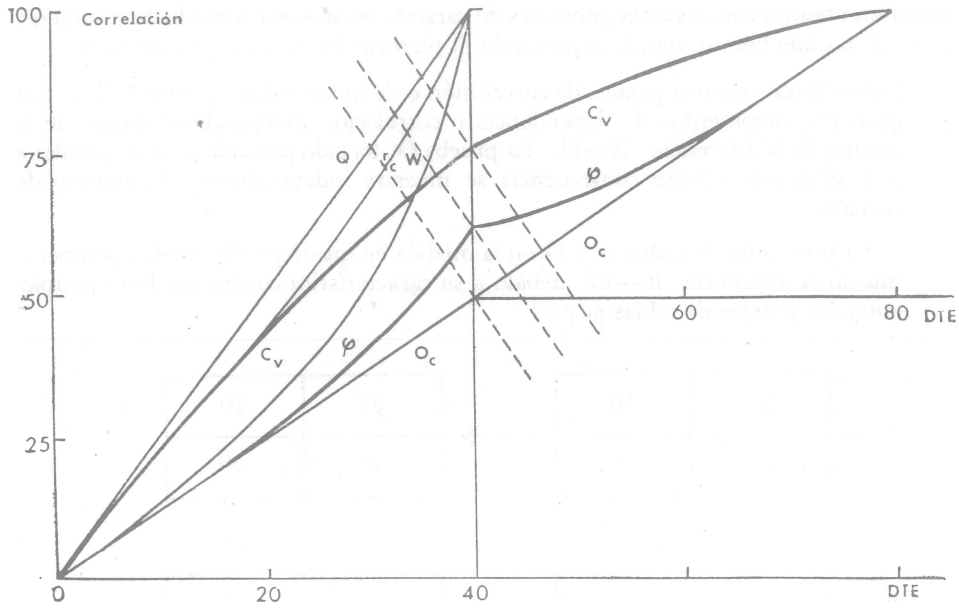




se da incremento similar del valor $|f_e - f_t|$ (el lector puede comprobar que es 0,5 en ambos casos); el valor de ji-cuadrado en a_1 aumenta 0,1505 y en a_2 aumenta 3,3576; es decir, para un mismo incremento $|f_e - f_t|$ el ji-cuadrado aymenta más a medida que los valores $f_e - f_t$ que lo originan son mayores.

b) Desaconsejamos el uso exclusivo de los coeficientes Q, W, r_c (pueden obtenerse como datos complementarios) por su falta notoria de fiabilidad; ponemos en tela de juicio el uso de C_v igualmente por su falta de fiabilidad; recomendamos una cierta precaución al interpretar el coeficiente φ y, por último, creemos que el coeficiente Oc tiene muchas ventajas que esperamos algún día no lejano mostrar rigurosa y probabilísticamente de forma que podamos reafirmarnos en las impresiones que aquí presentamos y a las que hemos llegado por vía empírico-experimental simulada.

Al final del trabajo, en la segunda parte del artículo, expondremos algunas conclusiones de carácter metodológico más general así como la correspondiente información bibliográfica.



Dirección de los autores: O. G. Leon García y F. J. Tejedor Tejedor, c/ Oña, 9, Madrid-34

SUMARIO: Tratamos de ofrecer una visión general de los problemas presentados por la interpretación de los coeficientes de correlación general que pueden ser calculados mediante las tablas de contingencia de 2 × 2.

En este trabajo es presentado el estudio de la dependencia/independencia que puede existir entre las categorías estudiadas —lo cual recomendamos hacer— porque el uso de ji-cuadrado como estadístico para detectar la significación de la diferencia entre frecuencias, pensamos que no es suficiente, y ello es debido a la sensibilidad del estadístico arriba mencionado al tamaño de la muestra.

Hemos revisado la conducta de los diferentes coeficientes de correlación que pueden ser obtenidos mediante tablas de contingencia de 2 × 2; uno de ellos es propuesto por los autores porque puede alternarse con los de uso más común.

Descriptores: Correlation coefficients for a 2×2 table, contingency tables, Chi square, nominal variable, non parameter estimation, Phi coefficient.