7.-RELACIONES ENTRE VARIABLES: CORRELACIÓN Y REGRESIÓN

La correlación expresa la relación concomitante entre dos o más variables. Dicha relación puede ser perfecta, imperfecta o nula en función de su intensidad y positiva o negativa según el sentido de la misma. La unión de todos los pares de puntuaciones da como resultado una recta de regresión (cuando es perfecta positiva o negativa), una nube de puntos o diagrama de dispersión cuando es imperfecta positiva o negativa. Cuando las puntuaciones se encuentran en torno a una línea curva se habla de relaciones curvilíneas. Se pueden ver los diagramas siguientes donde se representan gráficamente estos conceptos.

Como se puede apreciar, cuando tenemos una correlación

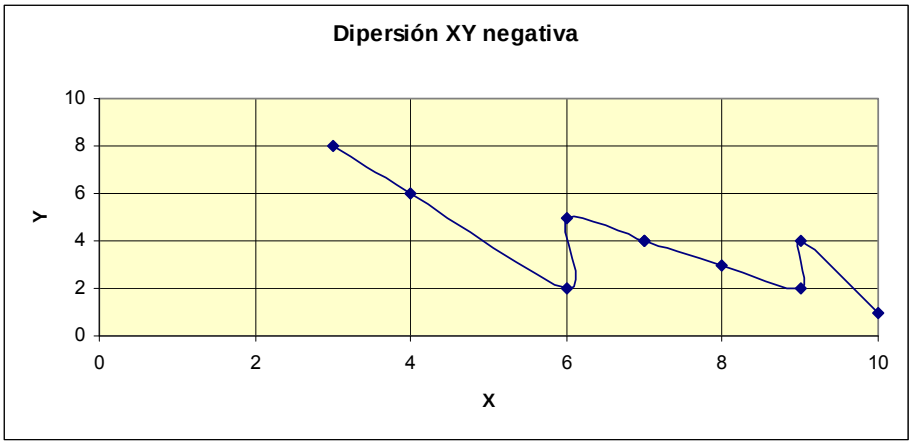
* perfecta y positiva (+1): a mayores puntuaciones de X se corresponden mayores puntuaciones de Y y viceversa,
* perfecta y negativa (-1): a mayores puntuaciones de X le corresponden menores puntuaciones de Y y viceversa,
* nula (0): las variables son independientes y no se puede establecer una relación

El establecimiento de la asociación se realiza en base a la covarianza o variación conjunta de X e Y. La covarianza trabaja con las puntuaciones diferenciales y se define como la media aritmética de los productos x·y de todos los pares de datos en una muestra y se calcula mediante la fórmula

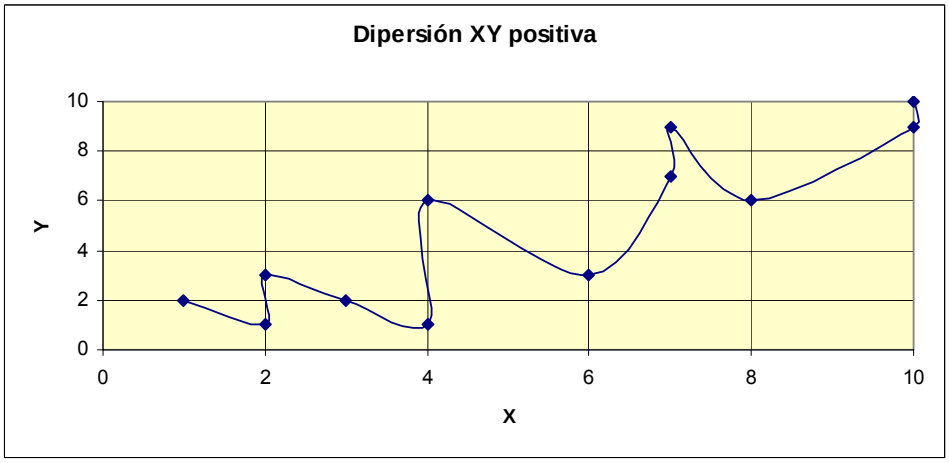
$$\sigma_{xy} = \frac{\sum x \cdot y}{N}$$

La covarianza será positiva, negativa o nula. Sin embargo, no nos indica la intensidad de la relación, ya que la magnitud de la covarianza no tiene límites y es muy difícil saber cuando estamos ante una relación fuerte o débil.

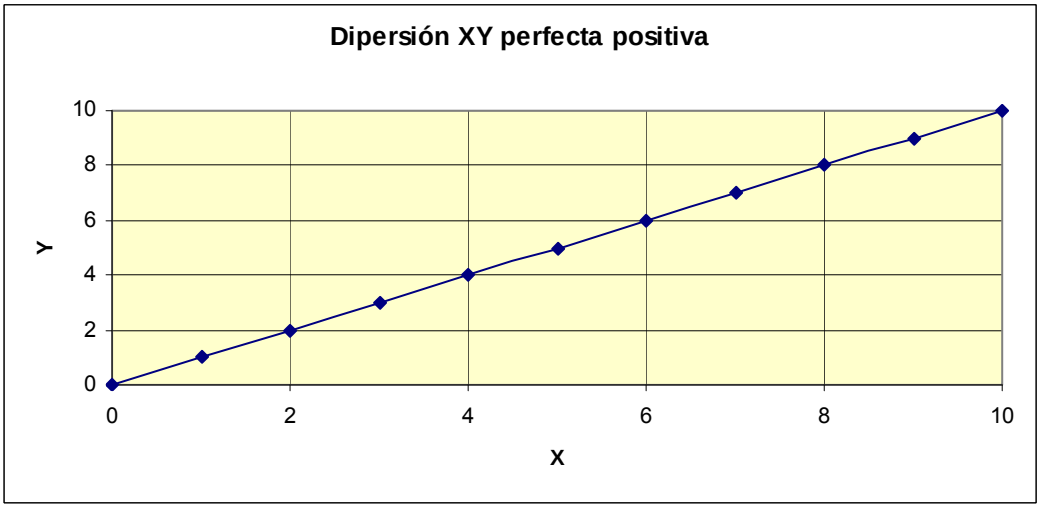
X	Y
3	8
5	6
6	2
6	5
7	4
7	4
8	3
9	2
9	4
10	1
r	-0.83986435
R^2	0.70537212



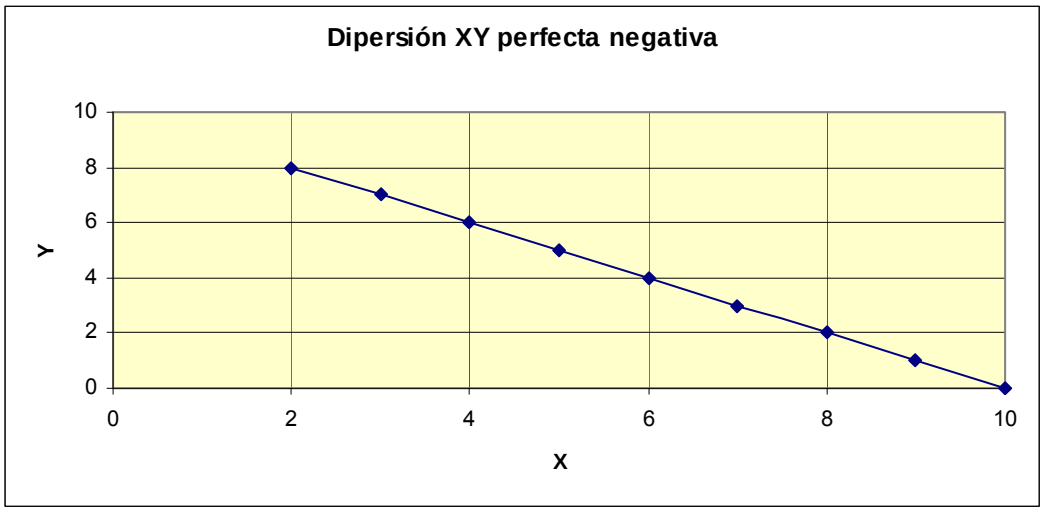
X	Y
1	2
2	1
2	3
3	2
4	1
4	6
6	3
7	7
7	9
8	6
10	9
10	10
r	0.85704604
R^2	0.73452791



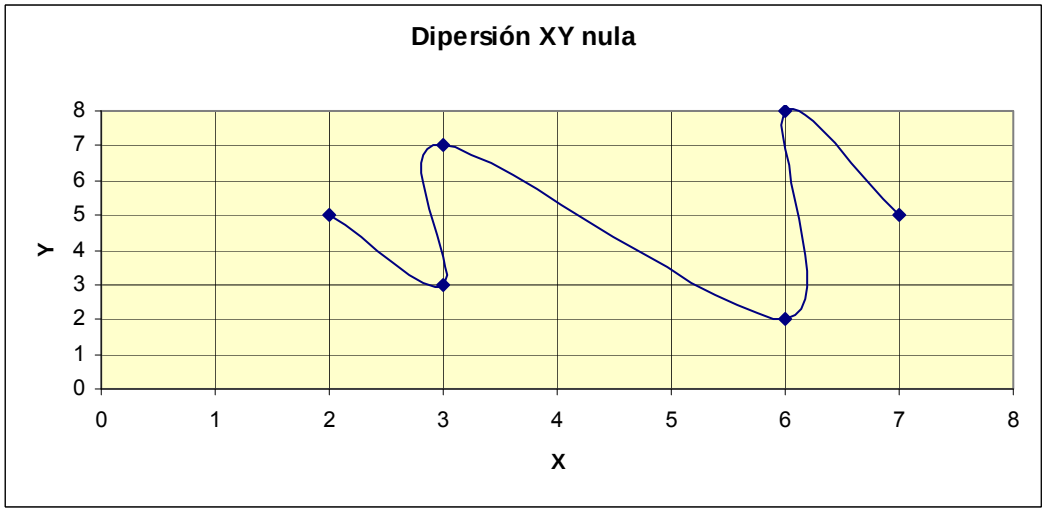
X	Y
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
r	1
R^2	1



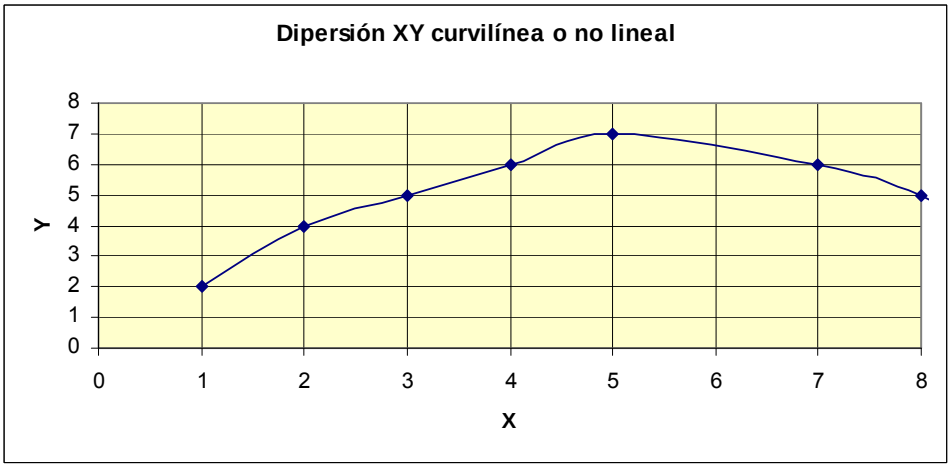
X	Y
2	8
3	7
4	6
5	5
6	4
7	3
8	2
9	1
10	0
r	-1
R^2	1



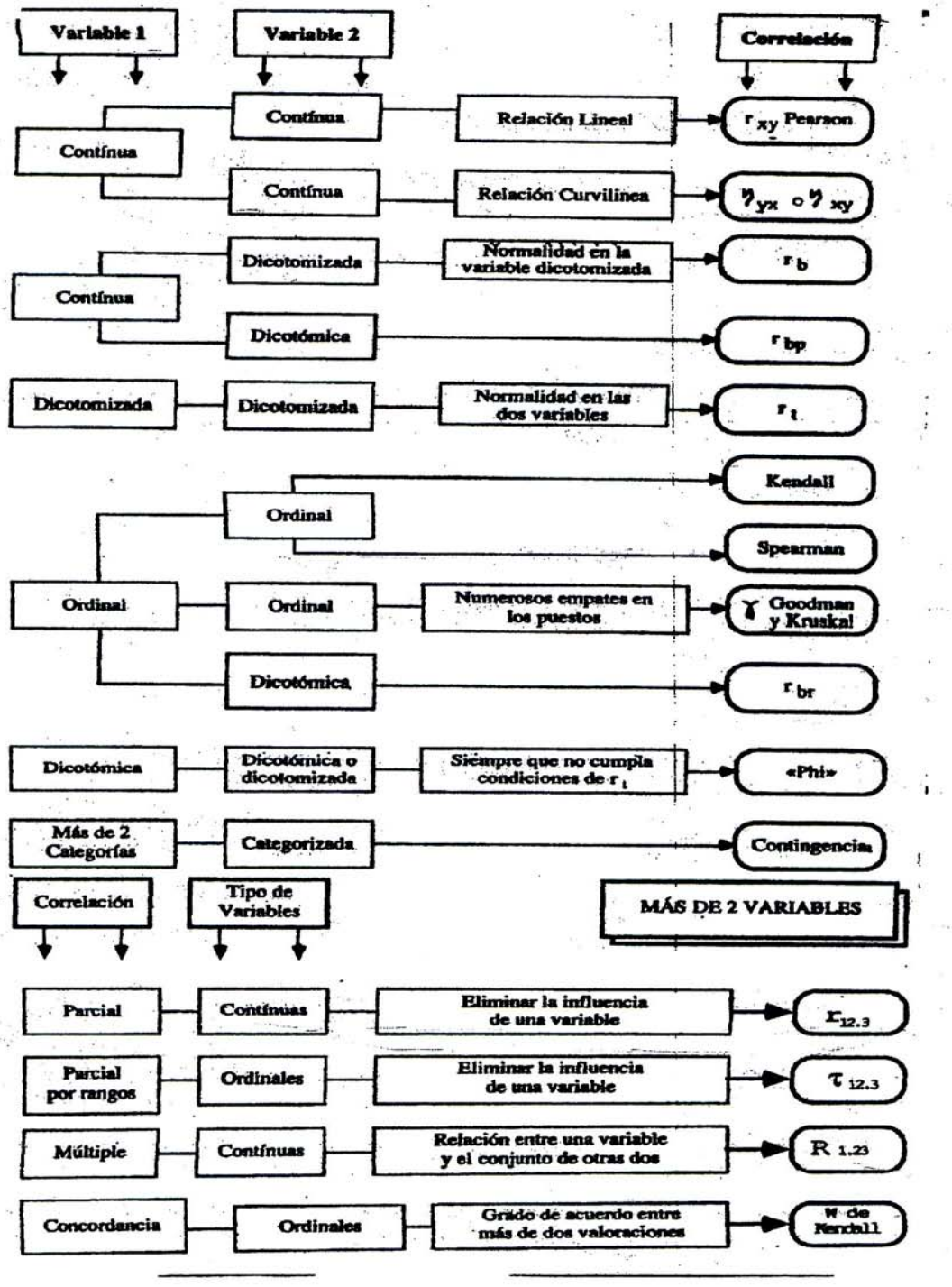
X	Y
2	5
3	3
3	7
6	2
6	8
7	5
r	0
R^2	0



X	Y
1	2
2	4
3	5
4	6
5	7
7	6
8	5
9	3
r	0.2287
R^2	0.0523



Resumen de tipos de coeficientes de correlación



MARTÍN,F. ET AL. (1993): Estadística aplicada. Tratamiento informático con STAT-VIEW 512+. Salamanca: Univ. Pontificia, p. 70-71

		CUALITATIVAS		CUASI-CUANTITATIVAS			CUANTITATIVAS DISCRETAS			CUANTITATIVAS CONTINUAS		
		Dicotómicas o dicotomizadas	+ 2 categorías	Sin categ.	Dicotomiz	+ 2 categ.	Sin categ.	Dicotomiz	+ 2 categ.	Continuas	Dicotomiz	+ 2 categ.
CUALITATIVAS	dicotómicas o dicotomizadas	φ	C	r_{b_r}	φ	C		φ	C	r_{b_p}	φ	C
	+ 2 categorías	C	C		C	C		C	C		C	C
CUASI-CUANTITATIVAS	sin categorías	$\rho\ r\ \gamma$										
	dicotomizadas				φ	C		φ	C	r_{b_p}	φ	C
	+ 2 categorías				C	C		C	C		C	C
CUANTITATIVAS	DISCRETAS Sin categorías											
	DISCRETAS Dicotomizadas							φ	C	r_{b_p}	φ	C
	DISCRETAS + 2 categorías							C	C		C	C
	CONTINUAS continuas (sin categorías)									$r;\eta$	r_b	
	CONTINUAS dicotomizadas									r_b	r_t	C
	CONTINUAS + 2 categorías										C	C

Cuadro con los más importantes coeficientes de correlación, según los distintos tipos de variables

MARTÍN,F. ET AL. (1984): Estadística descriptiva. Salamanca: Univ. Pontificia, p. 136

Interpretación de los coeficientes de correlación

Los valores de una correlación oscilan entre +1 y -1. Sin embargo, hay que tener cuidado a la hora de interpretar un coeficiente ya que esta escala no toma valores constantes o, lo que es lo mismo, 0,80 por ej. no es el doble de 0,40. Para hacernos una idea más atinada del grado de asociación hay que elevar el índice al cuadrado con lo que, en el caso anterior, tendríamos un 64% y un 16% respectivamente (**coeficiente de determinación – ver pág. 80)**.

Las correlaciones no se pueden promediar entre sí directamente. Sí se pueden promediar transformándolas en otro tipo de puntuaciones (Z de Fisher) como veremos más adelante.

Por tanto, a la hora de interpretar un índice concreto, véase el 0,87 obtenido en el ejemplo que presentamos a continuación, hay que tener en cuenta lo siguiente:

* Significatividad estadística: descartar que la relación encontrada no se debe al azar. La significatividad está muy influenciada por el N de la muestra con la que trabajamos -a mayor N mayor significatividad y a menor N menos significatividad-.
* Dirección o sentido de la correlación (+, - o nula): esto nos indicará el sentido de la asociación como ya se ha indicado.
* La cuantía o intensidad:
    1) Depende de la naturaleza de las variables. Por ej. un coeficiente de 0,40 entre peso y CI podría considerarse insospechado y altísimo. Sin embargo, el mismo índice obtenido entre dos tests distintos que miden memoria, sería prácticamente despreciable. Es muy importante la coherencia con los coeficientes obtenidos en estudios que trabajen con variables similares. Nos haremos una idea más adecuada si elevamos dichos índices al cuadrado y así tendremos información sobre el grado de covarianza o varianza común entre las variables correlacionadas.
    2) Depende de la variabilidad del grupo: las correlaciones son siempre más bajas cuanto más homogéneas son las muestras y más altas cuanto más heterogéneas. Remitimos a las representaciones de las nubes de dispersión en las páginas anteriores. Por ejemplo, si realizamos un estudio sobre el rendimiento de los alumnos y CI en una muestra amplia, cabe esperar una correlación más alta que si realizamos el mismo estudio con un grupo más restringido (por ej. con todos los que tienen un CI mayor o igual a 110).

* Correlación sólo indica asociación lo que no implica que haya relación de causa-efecto necesariamente. Las relaciones causales deberán establecerse por otras vías.

Siempre que se ofrezcan índices de correlación deberá indicarse su significatividad estadística, el N con el que han sido obtenidos y la cuantía de los mismos. Interpretación puramente orientativa (BISQUERRA,R. 1987, p. 189):

>0.80	Muy alta
0.60 - 0.79	Alta
0.40 - 0.59	Moderada
0.20 - 0.39	Baja
<0.20	Muy baja

7.1.- Coeficientes de correlación entre dos variables

7.1.1.- Coeficiente de correlación de Pearson para datos sin agrupar[1]

Ejemplo:

X	Y	x	y	x2	y2	xy	X2	Y2	XY	Zx	Zy	ZxZy
15	8	7	2	49	4	14	225	64	120	1,61	0,57	0,924
13	12	5	6	25	36	30	169	144	156	1,15	1,72	1,981
11	6	3	0	9	0	0	121	36	66	0,69	0	0
10	10	2	4	4	16	8	100	100	100	0,46	1,15	0,528
9	8	1	2	1	4	2	81	64	72	0,23	0,57	0,132
7	6	-1	0	1	0	0	49	36	42	-0,2	0	0
7	5	-1	-1	1	1	1	49	25	35	-0,2	-0,3	0,066
5	3	-3	-3	9	9	9	25	9	15	-0,7	-0,9	0,594
3	2	-5	-4	25	16	20	9	4	6	-1,2	-1,1	1,321
0	0	-8	-6	64	36	48	0	0	0	-1,8	-1,7	3,169
80	60	0	0	188	122	132	828	482	612			8,716

$$\overline{X} = \frac{80}{10} = 8 \qquad \overline{Y} = \frac{60}{10} = 6 \qquad \sigma_x = 4{,}34 \qquad \sigma_y = 3{,}49$$

Fórmulas para el cálculo:

$$1ª \quad r_{xy} = \frac{\sum Z_x \bullet Z_y}{N}$$

[1] Ejemplo tomado de MARTÍN,F. ET AL (1984): Estadística descriptiva. Salamanca: Univ. Pontificia, p. 141.

$$2^{a} \quad r_{xy} = \frac{\sum x \bullet y}{N \bullet \sigma_x \bullet \sigma_y}$$

$$3^{a} \quad r_{xy} = \frac{\sum x \bullet y}{\sqrt{\sum x^2 \bullet \sum y^2}} \quad \text{la que usa Excel}$$

$$4^{a} \quad r_{xy} = \frac{N\sum XY - \sum X \bullet \sum Y}{\sqrt{\left[N\left(\sum X^2\right) - \left(\sum X\right)^2\right] \bullet \left[N\left(\sum Y^2\right) - \left(\sum Y\right)^2\right]}}$$

Soluciones al ejercicio:

$$1^{a} \qquad r_{xy} = \frac{8{,}716}{10} = 0{,}87$$

$$2^{a} \qquad r_{xy} = \frac{132}{10 \bullet 4{,}34 \bullet 3{,}49} = 0{,}87$$

$$3^{a} \qquad r_{xy} = \frac{132}{\sqrt{188 \bullet 122}} = 0{,}87$$

$$4^{a} \qquad r_{xy} = \frac{10 \bullet 612 - 80 \bullet 60}{\sqrt{\left[10 \bullet 828 - 80^2\right]\left[10 \bullet 482 - 60^2\right]}} = 0{,}87$$

Significatividad de la correlación

Mediante los siguientes procedimientos de cálculo podemos establecer si una correlación es distinta de cero o, lo que es lo mismo, si la correlación encontrada no se debe al azar.

Cálculo del error típico:

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}}$$

Cálculo de la razón crítica:

$$R_c = \frac{r}{\sigma_r} = \left(r\sqrt{N}\right) / \left(1 - r^2\right)$$

Cuando el índice obtenido es mayor o igual que el límite crítico establecido por la curva normal (1,96 para un Nc del 5% y 2,58 para el 1%), se puede afirmar que la correlación obtenida es estadísticamente significativa.

Si quisiéramos saber cuantos sujetos necesitaríamos para obtener una correlación significativa podemos aplicar la fórmula siguiente:

$$N = \left(\frac{Nc}{r} \right)^2$$

Ejemplo:

En el caso presentado anteriormente habíamos obtenido una r = 0,87, teníamos un N = 10.

$$\sigma_r = \frac{1-(0{,}87)^2}{\sqrt{10}} = \frac{1-0{,}757}{3{,}1623} = 0{,}0768$$

$$Rc = \frac{0{,}87}{0{,}0768} = 11{,}3281$$

La correlación obtenida es estadísticamente significativa al obtener una Rc mayor que 1,96 o 2,58 que serían los niveles de confianza necesarios para el 5% y 1% respectivamente.

Si quisiéramos saber qué N necesitaríamos para que una correlación (por ejemplo, r = .11) fuese significativa, tendríamos que aplicar la fórmula siguiente:

$$N = \left(\frac{Nc}{r} \right)^2 = \left(\frac{1{,}96}{0{,}11} \right)^2 = 17{,}818^2 = 317$$

Significatividad de diferencias entre correlaciones en muestras relacionadas o en la misma muestra

Si, por ejemplo, tenemos una muestra de 147 casos en los que hemos hallado la correlación entre un test3 y otro test1 (criterio), así como la correlación entre otro test2 y el mismo test1 (criterio), y la correlación entre los test3 y test2. Dichas correlaciones son las siguientes: r_{13} = 0,68 r_{12} = 0,77 r_{23} = 0,84. Deseamos saber si las correlaciones del test1

(criterio) con cada test (r_{13} y r_{12}) son significativamente distintas.

La fórmula a aplicar sería la siguiente:

$$t = \frac{\left|(r_{12} - r_{13})\right|\sqrt{N-3}\sqrt{1+r_{23}}}{\sqrt{2}\sqrt{1-r_{12}^2-r_{13}^2-r_{23}^2+2r_{12}r_{13}r_{23}}}$$

En este caso la t tiene N-3 grados de libertad.

$$t = \frac{\left|(0{,}77-0{,}68\right|\sqrt{147-3}\sqrt{1+0{,}84}}{\sqrt{2}\sqrt{1-0{,}77^2-0{,}68^2-0{,}84^2+2(0{,}68)(0{,}77)(0{,}84)}} = \frac{(0{,}09)(12)(1{,}356)}{(1{,}414)(0{,}3445)} = \frac{1{,}4648}{0{,}48712} = 3$$

En la *Tabla 6 de valores críticos de t*, vemos que para 144 g.l. 2,576 t se corresponde con el Nc. del 1%. Como nosotros hemos obtenido un valor de 3 quiere decir que la diferencia encontrada entre los dos tests con el test 1 (r_{12} - r_{13} =0,09) es significativa a un Nc. superior al 1%.

Media de varias correlaciones

Cuando tenemos varias correlaciones obtenidas entre dos variables en varias muestras independientes de una misma población, podemos hallar la correlación media mediante el uso de la Tabla de z. Para ello seguimos el procedimiento de pasar r a z, utilizar la fórmula siguiente y volver a pasar z a r.

$$\overline{Z} = \frac{z_1(N_1-n)+z_2(N_2-n)+\ldots+z_n(N_n-n)}{N_1+N_2+\ldots+N_n}$$

n= número de correlaciones

En el ejemplo siguiente, hemos obtenido la correlación entre un test de inteligencia y el éxito escolar en tres muestras:

r_1 = 0,64 N_1 = 80

r_2 = 0,72 N_2 = 126

r_3 = 0,64 N_3 = 92

1) Transformamos r en z utilizando la Tabla 8: z_1 = 0,758 z_2 = 0,908 z_3 = 0,758

2) Hallamos la media ponderada de esas z

$$\overline{Z} = \frac{z_1(N_1 - 3) + z_2(N_2 - 3) + z_3(N_3 - 3)}{N_1 + N_2 + N_3}$$

3) Convertimos este valor z a su r correspondiente. r = 0,665

$$\overline{Z} = \frac{(0{,}758)(77) + (0{,}908)(123) + (0{,}758)(89)}{289} = 0{,}8218$$

Esta sería la correlación media o, mejor dicho, la correlación que mejor representa la relación entre las dos variables en la población en función de la información que sobre ella tenemos a partir de las tres correlaciones anteriores.

Regresión y predicción

Galton enunció la ley de regresión universal al analizar la correlación entre estatura y herencia (los hijos de padres altos tendían a ser menos altos que sus padres y los de padres bajos eran más bajos, pero menos, que sus padres). Se producía una regresión hacia la talla media.

En este apartado lo que se trata es de pronosticar o predecir los valores de una variable (dependiente) a partir de la otra (independiente). Para ello se utiliza la recta de regresión construida a partir de la función de relación lineal (y = a + bx) donde x e y son variables y a y b constantes (ordenada en el origen y pendiente de la recta) cuando la pendiente es positiva (relación positiva), negativa (negativa) y nula (recta paralela al eje de abcisas).

Ecuaciones de regresión:

Y en función de X	X en función de Y
Y´=a + bX $b = \frac{N\sum XY - \sum X \bullet \sum Y}{N\sum X^2 - (\sum X)^2}$ $A = \overline{Y} - b\,\overline{X}$ Si conocemos la correlación entre X e Y $b = r_{xy}\frac{\sigma_y}{\sigma_x}$	X´=a + bY $b = \frac{N\sum XY - \sum X \bullet \sum Y}{N\sum Y^2 - (\sum Y)^2}$ $A = \overline{X} - b\,\overline{Y}$ Si conocemos la correlación entre X e Y $b = r_{xy}\frac{\sigma_x}{\sigma_y}$
Puntuaciones diferenciales	
Y´= bx $y' = r_{xy}\frac{\sigma_y}{\sigma_x} \bullet x$	x´= by $x' = r_{xy}\frac{\sigma_x}{\sigma_y} \bullet y$
Puntuaciones típicas	
$z'y = bz_x$ $b = r_{xy}$ $z'y = r_{xy} * z_x$	$z'x = bz_x$ $b = r_{xy}$ $z'x = r_{xy} * z_y$

Ejemplo:

¿Cuál es la recta de regresión (Y en función de X) expresada en puntuación directa, diferencial y típica de los datos siguientes?:

Sujetos	X	Y	X^2	X·Y	x	x^2	y	y^2	x·y
A	8	6	64	48	-1	1	-2	4	2
B	10	9	100	90	1	1	1	1	1
C	6	7	36	42	-3	9	-1	1	3
D	12	10	144	120	3	9	2	4	6
Suma	36	32	344	300		20		10	12

Cálculos previos:

$$\overline{X} = \frac{36}{4} = 9$$

$$\overline{Y} = \frac{32}{4} = 8$$

$$\sigma_x = \sqrt{\frac{20}{4}} = 2{,}24$$

$$\sigma_y = \sqrt{\frac{10}{4}} = 1{,}58$$

$$r_{xy} = \frac{\sum xy}{N\sigma_x\sigma_y} = \frac{12}{4 \bullet 2{,}24 \bullet 1{,}58} = 0{,}85$$

Recta de regresión con puntuaciones directas:

$$b = \frac{4 \bullet 300 - 36 \bullet 32}{4 \bullet 344 - 36^2} = 0{,}6$$

a=8-(0,6)(9)=2,6

b=0,85(1,58/2,24)=0,6

Y´=2,6+0,6X

Recta de regresión con puntuaciones diferenciales:

Y´=0,85(1,58/2,24)x=0,6x

Recta de regresión con puntuaciones típicas:

$Z_y' = 0{,}85Z_x$

Podemos calcular los valores pronosticados para el sujeto C, por ejemplo. Sería

Y' = 2,6 + 0,6X = 2,6 +0,6 · 6 = 6,2

y' = 0,6x = 0,6 (-3) = -1,8

Z'y= 0,85 · Zx = 0,85 · ((-3)/2,24) = 0,85 · (-1,34) = - 1,14

Error típico de estimación:

Cuando predecimos mediante ecuaciones de regresión, cometemos errores (véase como en el ejemplo para el sujeto C. Pronosticábamos 6,2 mientras que tiene una Y=7. Hubo, por tanto, un error de estimación de 0,80). El calibre de estos errores viene dado por el error típico de estimación.

Fórmulas para Y en función de X

$$\sigma_{xy} = \sigma_x\sqrt{1 - r_{xy}^2}$$

para X en función de Y

$$\sigma_{yx} = \sigma_y\sqrt{1 - r_{xy}^2}$$

Cuando el N sea menor de 50 sujetos habrá que multiplicar el resultado por la corrección $\sqrt{\frac{N}{N-2}}$ por lo que tendríamos que $\sigma_{yx} = \sigma_y \sqrt{1-r_{xy}^2} \bullet \sqrt{\frac{N}{N-2}}$

$$\sigma_{yx} = 1{,}58\sqrt{1-0{,}85^2} \bullet \sqrt{\frac{4}{2}} = 1{,}58 \bullet 0{,}5268 \bullet 1{,}4142 = 1{,}1771$$

Calculo del error si pronosticamos X en función de Y:

Según el nivel de confianza que tomemos para el pronóstico, la puntuación verdadera se encontraría entre unos márgenes o intervalos de confianza. Así

$Y' \pm \sigma_{y \cdot x} = 68{,}26\%$

$Y' \pm (\sigma_{y \cdot x})(1{,}96) = 95\%$

$Y' \pm (\sigma_{y \cdot x})(2{,}58) = 99\%$

6,2 ± 1,1771 = 7,3771 y 5,0229 la puntuación obtenida por el sujeto en Y en base a la puntuación de X se moverá entre estos márgenes con un nivel de confianza del 68,26%.

6,2 ± (1,1771·1,96) = 8,5071 y 3,8929 la puntuación obtenida por el sujeto en Y en base a la puntuación de X se moverá entre estos márgenes con un nivel de confianza del 95%.

6,2 ± (1,1771·2,58) = 9,2369 y 3,1631 la puntuación obtenida por el sujeto en Y en base a la puntuación de X se moverá entre estos márgenes con un nivel de confianza del 99%.

El error típico varía entre 0 y la desviación típica de la variable dependiente, esto es:
Si $r_{xy} = 0$ ----> $\sigma_{x.y} = \sigma_y$: error máximo
Si $r_{xy} = 1$ ó -1 ----> $\sigma_{x.y} = 0$: no hay error o la predicción es perfecta.

Coeficiente de alienación (K) o % de azar o incertidumbre en el pronóstico. Se multiplica por 100 y se interpreta como porcentaje.

$$K = \sqrt{1-r_{xy}^2}$$

Coeficiente de eficiencia predictiva (E) o seguridad en el pronóstico

$$E = 1 - \sqrt{1-r_{xy}^2}$$

$$E = 1 - K$$

se multiplica por 100 y se interpreta como porcentaje.

Coeficiente de determinación (r_{xy}^2) se multiplica por 100 y se interpreta como el porcentaje de covarianza o varianza asociada, o porcentaje de varianza explicada por X o por Y.

En el ejemplo estudiado tendríamos un coeficiente K = 52,68%, un coeficiente E = 47,32% y un coeficiente de determinación de 72,25%.

7.1.2.- Correlación biserial (r_b) y biserial puntual (r_{bp})

Aunque sean coeficientes distintos, las fórmulas de cálculo son muy semejantes ya que sólo cambia el segundo factor.

Biserial: r_b	biserial puntual: r_{bp}
REQUISITOS	
Dos variables continuas, normales linealmente relacionadas de las que una ha sido dicotomizada (ej. items, sueldos...)	Dos variables una continua y la otra dicotómica o, siendo continua, no se distribuye normalmente
CÁLCULO	
$r_b = \frac{\overline{X}_p - \overline{X}_q}{\sigma_x} \bullet \frac{pq}{y}$	$r_{bp} = \frac{\overline{X}_p - \overline{X}_q}{\sigma_x} \sqrt{pq}$
$r_b = \frac{\overline{X}_p - \overline{X}}{\sigma_x} \bullet \frac{p}{y}$	$r_{bp} = \frac{\overline{X}_p - \overline{X}}{\sigma_x} \sqrt{\frac{p}{q}}$

X es la variable continua

Y la dicotómica o dicotomizada.

p: proporción de casos de una de las dos modalidades de Y
q: 1 - p
y: ordenada de la curva normal que divide las áreas p y q. Ver Tabla 9.

$\overline{X}_p$ media de los casos que en la variable X poseen la característica p

$\overline{X}_q$ media de los casos que en la variable X poseen la característica q

$\overline{X}$ media de todos los casos en la variable X

σ_x desviación típica de la variable X

Interpretación

r_{bp}: biserial puntual

Magnitud: igual que Pearson

Dirección:

+: $\overline{X}_p > \overline{X}_q$ y la diferencia $\overline{X}_p - \overline{X}_q$ es positiva. Esto quiere decir que los sujetos que puntúen alto en X tenderán a pertenecer a la modalidad de Y denominada como p. Los que puntúen bajo en X tenderán a pertenecer a la modalidad q de Y.
-: En caso de ser negativo la interpretación se realiza a la inversa: a puntuación alta en X le corresponderá la modalidad q y, a puntuación baja en X, la modalidad p.

En el ejemplo siguiente se cumple la condición positiva, por lo que los sujetos que tienen una mayor puntuación en la prueba objetiva X, tienden a acertar el ítem Y. O, lo que es lo mismo, los que dominan la materia suelen acertar el ítem Y.

Ejemplo:
Tenemos las puntuaciones totales de 10 sujetos que han contestado a una prueba objetiva (X) y además sabemos si han acertado (1) o fallado (0) un ítem Y. Tenemos, por tanto, una variable cuantitativa y otra dicotómica.

Datos:

N	X	Y
1	1	0
2	1	1
3	2	0
4	2	0
5	3	1
6	4	1
7	6	1
8	6	1
9	8	1
10	10	1

$\overline{X}$ =4,3 σ_x =3,093 $\overline{X}_p$ =5,429 $\overline{X}_q$ =1,6667

$$r_{bp} = \frac{\overline{X}_p - \overline{X}}{\sigma_x}\sqrt{\frac{p}{q}} = \frac{5,429 - 4,3}{3,093}\sqrt{\frac{0,7}{0,3}} = 0,3650 \bullet 1,5275 = 0,5575$$

$$r_b = \frac{\overline{X}_p - \overline{X}_q}{\sigma_x} \bullet \frac{pq}{y} = \frac{5{,}429 - 1{,}667}{3{,}093} \bullet \frac{0{,}7 \bullet 0{,}3}{0{,}3477} = 1{,}2164 \bullet 0{,}6040 = 0{,}7347$$

Como se puede apreciar, con los mismos datos, la biserial es mayor que la biserial puntual (un 76% más). Comprobar el valor de y en la Tabla 9.

r_b: biserial

El valor puede superar la unidad y es mayor que r_{bp} para los mismos datos (entre un 25-70% mayor). Es menos fiable que r_{bp} y no se puede comparar con Pearson. Máxima fiabilidad cuando p=q. La dirección o el sentido de la correlación se interpreta como r_{bp}.

Un ejemplo típico de correlación biserial sería el siguiente:[2]

Hemos medido el peso (Y) y la estatura (X) a un grupo de individuos dividiéndolos según el peso en obesos (peso superior a la mediana) y delgados (peso inferior a la mediana) ¿hay alguna relación entre peso y altura así considerados?

X	Y(p)	Y(q)
176	1	
174	1	
172	1	
170	1	
169	1	
168		1
166		1
164		1
160		1
155		1

p=0,5, q=0,5, y=0,3989

$\overline{X}$ =167,4 σ_x =6,40 $\overline{X}_p$ =172,20 $\overline{X}_q$ =162,60

pq/y=0,6267

p/y=1,2530

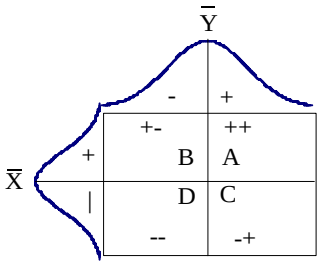
$$r_b = \frac{\overline{X}_p - \overline{X}_q}{\sigma_x} \bullet \frac{pq}{y} = \frac{172{,}2 - 162{,}6}{6{,}4} \bullet \frac{0{,}5 \bullet 0{,}5}{0{,}3989} = 0{,}94$$

$$r_b = \frac{\overline{X}_p - \overline{X}}{\sigma_x} \bullet \frac{p}{y} = \frac{172{,}2 - 167{,}4}{6{,}4} = \frac{0{,}5}{0{,}3989} = 0{,}94$$

Por tanto, existe una alta relación entre tener valores altos en la variable altura y pertenecer al grupo de obesos y obtener valores bajos en altura y pertenecer al grupo de delgados.

7.1.3.- Correlación tetracórica (r_t)

La correlación tetracórica requiere que las dos variables relacionadas sean continuas, normales, relacionadas linealmente y dicotomizadas artificialmente. Además requiere un N ≥100 y todas y cada una de las celdas deberán ser iguales o mayores que el 10% del N total. Cuando una celda es 0, no se calcula.



Cálculo:

Existen diversos métodos para el cálculo de r_t. Nosotros vamos a utilizar el método diagonal de Davidoff y Cohen que consiste en calcular el cociente de las diagonales de la tabla de contingencia (2X2) y ver a qué r_t corresponde en la Tabla 10.

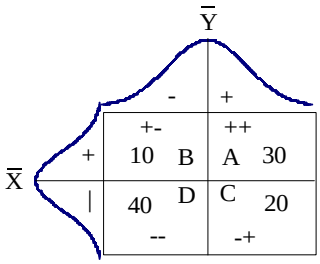
$$r_t = \frac{AD}{BC}$$

Cuando BC>AD se invierte el orden y se estima la r_t con signo negativo

$$r_t = \frac{BC}{AD}$$

Por ejemplo: Hemos sometido a 100 sujetos a dos pruebas y hemos dicotomizado los resultados a partir de la media de modo que nos resulta la siguiente tabla:

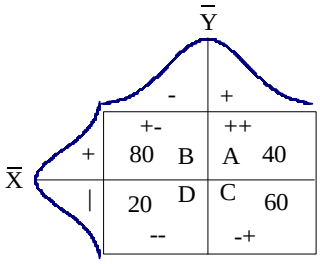
[2] CARRO,J. (1994): Psicoestadística descriptiva. Salamanca: Amaru, p. 201.



$$r_t = \frac{AD}{BC} = \frac{30 \cdot 40}{10 \cdot 20} = 6$$

Consultando la Tabla 10 nos encontramos con un valor de r_t = 0,61

En este otro ejemplo, tenemos datos de 100 sujetos en dos variables X (CI dicotomizada por ±100) e Y (neuroticismo dicotomizada a partir de la media) resultando la tabla de contingencia siguiente:



dado que AD<BC deberemos utilizar la fórmula siguiente:

$$r_t = \frac{BC}{AD} = \frac{80 \cdot 60}{40 \cdot 20} = \frac{4800}{800} = 6$$

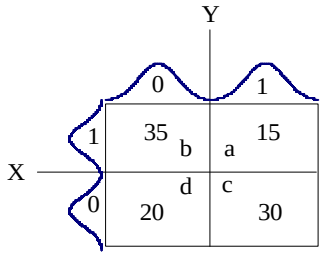
nos encontraremos dicho cociente entre el intervalo 5,81-6,03 con una r_t= - 0,61

Interpretación:

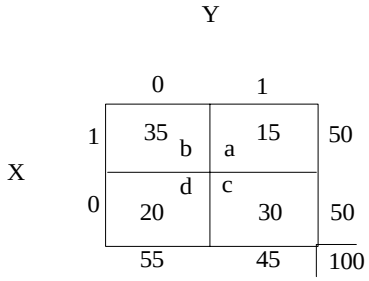
Se interpreta como r. Aunque siempre r es más fiable al usar todos los datos o frecuencias.

7.1.4.- Coeficiente φ

Este coeficiente se utiliza con variables cualitativas, las dos deberán ser realmente dicotómicas (ejs.: hombre-mujer, vivo-muerto...) o una dicotómica y la otra dicotomizada artificialmente. Se requiere que N >100, no debe emplearse ninguna proporción total que sea inferior a 0,05 siendo peligroso emplear proporciones inferiores a 0,10.



Por ejemplo, queremos estudiar la relación existente entre el sexo (X) y la respuesta a una pregunta de un determinado cuestionario (Y). La variable X especifica el sexo (0=hombre, 1=mujer) y la variable Y es clasificada como incorrecta (0) y correcta (1). Con los datos obtenidos en una muestra de 100 sujetos hemos construido la tabla siguiente:



La fórmula más usual para calcular el coeficiente φ es la siguiente:

$$\varphi = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b)(c+d)(b+d)+(a+c)}}$$

en nuestro caso

$$\varphi = \frac{(15 \cdot 20) - (35 \cdot 30)}{\sqrt{(15+35)(20+30)(15+30)+(35+20)}} = -0{,}30$$

el valor es bajo y negativo, lo que quiere decir que la relación entre ser mujer y contestar correctamente (1,1) o entre ser hombre y contestar de manera incorrecta (0,0) es negativa. Esto nos deberá hacer pensar en que el signo del coeficiente deberá interpretarse en función de la organización de las celdas. Cuando es negativo predominan las casillas 1-0, 0-1 y cuando es positivo las casillas (0-0, 1-1).

Algunas características del coeficiente φ

* El valor de φ sólo alcanza la unidad en los dos casos siguientes:

a) Cuando las frecuencias de las celdas a y d son igual a 0, independientemente del número de frecuencias que tengan las casillas b y c

X \ Y	0	1	
1	55 (b)	0 (a)	55
0	0 (d)	45 (c)	45
	55	45	100

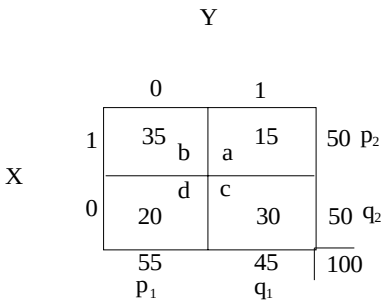
b) Cuando las frecuencias de las celdas b y c son igual a 0, independientemente del número de frecuencias que tengan las casillas a y d.

X \ Y	0	1	
1	0 (b)	35 (a)	35
0	45 (d)	0 (c)	45
	45	35	80

En cualquier otro caso, φ será menor que 1 y para interpretarlo hay que saber cuál sería el φ máximo de esos valores mediante la fórmula:

$$\varphi_{máx} = \sqrt{\frac{p_2}{q_2} \bullet \frac{q_1}{p_1}}$$

siendo p_1 la frecuencia marginal máxima (en proporción) de una variable, p_2 la de la otra variables y q_1 y q_2 las complementarias de ambas.



En nuestro caso tenemos

$$\varphi_{máx} = \sqrt{\frac{p_2}{q_2} \bullet \frac{q_1}{p_1}} = \sqrt{\frac{0,5}{0,5} \bullet \frac{0,45}{0,55}} = 0,9$$

que supone un -0,33 sobre el $\varphi_{máx}$

$$r^2 = \frac{\varphi}{\varphi_{máx}} = \frac{-0,3}{0,9} = -0,33$$

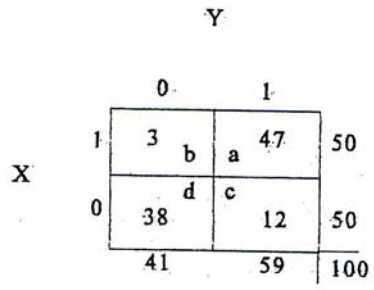
éste sería el verdadero φ, que se puede interpretar como r de Pearson en lo que respecta a la relación que implica.

$$r = \sqrt{\frac{\varphi}{\varphi_{\max}}} = \sqrt{\frac{-0,3}{0,9}} = -0,5773$$

Hay una fórmula directa para llegar a r con estos datos sin pasar por φ y $\varphi_{máx}$ que es la siguiente:

$$r = \frac{ad - bc}{\sqrt{\frac{p_2}{q_2} \cdot \frac{q_1}{p_1}}} = \sqrt{\frac{ad - bc}{p_2 \cdot q_1}}$$

Ejemplo:



$$\varphi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(b+d)(a+c)}}$$

$$\varphi = \frac{(38 \cdot 47) - (12 \cdot 3)}{\sqrt{(47+3)(12+38)(3+38)(47+12)}} = \frac{1750}{2459{,}16} = 0{,}71$$

$$\varphi_{máx} = \sqrt{\frac{p_2}{q_2} \bullet \frac{q_1}{p_1}} = \sqrt{\frac{50}{50} \bullet \frac{41}{59}} = \sqrt{\frac{51}{49}} = \sqrt{0{,}6949} = 0{,}83$$

$$r = \sqrt{\frac{\varphi}{\varphi_{\max}}} = \sqrt{\frac{0{,}71}{0{,}83}} = \sqrt{0{,}8554} = 0{,}92$$

$$r = \sqrt{\frac{(47 \cdot 38) - (3 \cdot 12)}{50 \cdot 41}} = \sqrt{\frac{1750}{2050}} = \sqrt{0{,}8536} = 0{,}92$$

El coeficiente φ está relacionado con χ^2, de modo que, cuando la tabla de contingencia es de 2x2, la relación es la siguiente:

$$\chi^2 = N\varphi^2$$

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

7.1.6.- Coeficiente de contingencia (C) y chi cuadrado (χ^2)

El coeficiente de contingencia (C) es una medida del grado de asociación entre dos conjuntos de categorías pertenecientes a variables. Se utiliza, sobre todo, cuando las variables están divididas en más de dos categorías cada una. Se obtiene en base χ^2 mediante la fórmula siguiente:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

1.- Concepto de χ^2 de Pearson

Permite saber si existe o no asociación entre dos variables. Indica la discrepancia entre unos datos empíricos y otros esperados en función de una hipótesis definida.

χ^2	r
Afirma si existe o no la asociación	Afirma si existe o no relación y el grado de tal relación
No indica el sentido de la asociación	Sí indica el sentido de la relación
Sirve para variables nominales, ordinales y de intervalo	Sólo sirve para variables continuas
No exige una "distribución especial" de las variables	Exige que ambas variables se distribuyan normalmente (homocedasticidad)
No exige función especial entre ambas variables	Exige función rectilínea lineal entre las variables

2.- La lógica del test de χ^2 de Pearson

El χ^2 es el cálculo de la diferencia total que existe entre los resultados obtenidos (empíricos) y los esperados (teóricos) o que se darían teóricamente entre dos variables que no estuvieran asociadas (Ho=hipótesis nula).

Si el χ^2 es tan grande que no puede explicarse por el azar se afirma la asociación entre las variables y se niega Ho. Es decir, si el χ^2 empírico es ≥ que el χ^2 teórico (Tabla 7) a un determinado nivel de confianza (α) y con unos grados de libertad (gl) determinados, no podemos rechazar la hipótesis nula (diferencia igual a 0) y, por tanto, afirmaremos que la diferencia no es significativa. En el caso contrario, es decir, cuando el χ^2 empí-

rico es mayor que el χ^2 teórico (Tabla 7) a un determinado nivel de confianza (α) y con unos grados de libertad (gl) determinados, podemos rechazar la hipótesis nula (diferencia igual a 0) y, por tanto, afirmaremos que la diferencia encontrada es significativa.

Cálculo:

$$\chi^2 = \sum \frac{(f_e - f_t)^2}{f_t}$$

donde f_e corresponde a las frecuencias empíricas, observadas o reales
donde f_t corresponde a las frecuencias teóricas o esperadas que se calculan para cada casilla mediante la fórmula siguiente: $f_t = \frac{\sum K_i \bullet \sum F_j}{N}$ donde K_i es el sumatorio de las f_e de la columna, F_j el de las filas y N el total de casos.
A la diferencia entre ambas en cada casilla o celda se denomina residual o residuo.

3.- Los grados de libertad (gl,df,k)

Cálculo:

gl=(F-1)(K-1) siendo F el número de filas y K el de columnas de la Tabla. Cuando sólo existe una fila,

gl = FK

Los grados de libertad son el número de elementos observados menos el número de relaciones que los ligan. O, dicho de otra forma, el número de grados de libertad en una tabla es la diferencia entre los datos que pueden variar y las condiciones que se imponen externamente (número de sujetos total y total por categorías)

Veamos algunos ejemplos: en una tabla 2x2 sólo tengo un grado de libertad o una única opción para colocar los datos (los de una casilla) ya que, a partir de ahí, todos los demás vienen determinados. Sean las variables del ejemplo (sexo: H y M, nacionalidad: N y E).

	H	M	Marginales filas
N	700	300	1000
E	200	800	1000
Marginales columnas	900	1100	2000

En función de los totales de que disponemos podríamos repartir teóricamente a los hombres (H) de muchas maneras; sin embargo, en cuanto elijamos una, todas las demás casillas serán dependientes. Si en la casilla E, H colocamos 200 nos veremos obligados a poner 700 en N, H y así sucesivamente para que los totales marginales de filas y columnas se mantengan.

Si tuviéramos una tabla 3x2

	H	M	Marginales filas
Licenciados	300	100	400
Bachilleres	400	200	600
Estudios Prim.	200	800	1000
Marginales columnas	900	1100	2000

En teoría, podríamos colocar a los 900 H como quisiéramos, ahora bien, una vez que determinemos las frecuencias de dos celdas el resto se ven forzadas a asumir otras determinadas. Si colocamos 200 H con Estudios primarios, esto nos obliga sólo a colocar 800 M (1 gl), si además (2 gl) colocamos a 300 H Licenciados, nos obliga a poner 400 H Bachilleres, 200 M Bachilleres y 100 M Licenciadas, con lo que hemos completado la Tabla. Por lo que tenemos 2 gl, como se resulta de la fórmula

$$gl=(F-1)(K-1)=(3-1)(2-1)=2$$

Una vez obtenido el valor de χ^2, determinado los grados de libertad y elegido el nivel de confianza, vamos a la Tabla 7 y buscamos el χ^2 donde se cruzan los gl o k y el nivel α establecido. Si el χ^2 empírico es menor o igual que el de la tabla, se considera como no significativo. Si el χ^2 empírico es mayor, se considera significativo estadísticamente. En la citada Tabla con un gl=1, para cada nivel α, hay una equivalencia con la escala z (distribución normal) como sigue $z=\sqrt{\chi^2}$ Por ej.: un χ^2 de 3,841 se corresponde con una

z de 1,9598 ~ 1,96 que incluye al 95% y adopta una p=0,05.

4.- Utilidad de χ^2

Prueba de bondad de ajuste: se trata de comprobar si la distribución empírica se ajusta a una distribución hipotética (teórica) previamente fijada. Mediante la aplicación de la fórmula ya expresada, se comprueba la bondad del ajuste entre las frecuencias teóricas y las empíricas comprobando la significatividad del χ^2 obtenido.

Prueba de independencia: igualmente podemos determinar si dos variables están o no relacionadas, es decir, si son dependientes o independientes. Para ello, lo primero que hay que hacer es calcular las frecuencias teóricas (aquellas que se obtendrían si fueran independientes o no hubiese relación), luego se extrae χ^2 y, finalmente, se comprueba su significatividad.

Vamos a utilizar el ejemplo siguiente que relaciona el tipo de música que le gusta (VD) y nivel de estudios (VI) para revisar todos los aspectos tratados.

Música	Estudios					
	primarios	secundarios	universitarios	Total filas		
Rock	10	30	60	100	fe	empiricas
	16,0	40,0	44,0		ft	teoricas
	10,0	30,0	60,0	100,0	%filas	
	25,0	30,0	54,5	40,0	%columnas	
	-6,0	-10,0	16,0		res	residuos
	-1,5	-1,6	2,4		resT	
	-2,1	-2,6	4,2		resC	
Pop	20	40	30	90	fe	
	14,4	36,0	10,0		ft	
	22,2	44,4	33,3	100,0	%filas	
	50,0	40,0	27,3	36,0	%columnas	
	5,6	4,0	20,0		res	
	1,5	0,7	6,3		resT	
	2,0	1,1	10,2		resC	
Clasico	10	30	20	60	fe	
	9,6	24,0	26,4		ft	
	16,7	50,0	33,3	100,0	%filas	
	25,0	30,0	18,2	24,0	%columnas	
	0,4	6,0	-6,4		res	
	0,1	1,2	-1,2		resT	
	0,2	1,8	-1,9		resC	
Total col	40	100	110	250		

residuos tipificados:resT deberá ser mayor que 1
residuos corregidos o ajustados: resC se interpretan como punt. Típicas +-1,96=95% y +-2,58=99%

g.l.=(F-1)*(C-1)	4	signif.
Chi2	18,586	0,00
C	0,263	
Cmax	0,816	
C--->r	0,610	
C--->r2	37,202	

Significatividad: comprobar en tabla o usar función de excel

Proceso:

a) Las f_t de cada celda se calculan a partir de los totales de filas y columnas divididos por el N total. Por ejemplo para la celda A $f_t = \frac{\sum K_i \bullet \sum F_j}{N} = \frac{40 \bullet 100}{250} = 16$ que sería el N que se correspondería con el % de sujetos respecto al total que incluye dicha celda.

b) Se calculan los residuales. En cada casilla (f_e - f_t). Hay que comprobar cuántas celdas hay con f_t menores que 5. Si hay un porcentaje elevado (entorno al 20% o más) los datos no se ajustan al modelo.

c) Se calcula χ^2 que, en este caso, es como sigue:

$$\chi^2 = \sum \frac{(f_e - f_t)^2}{f_t} = \frac{(-6)^2}{16} + \frac{(-10)^2}{40} + \frac{(16)^2}{44} + \frac{(5,6)^2}{14,4} + \frac{(4)^2}{36} + \frac{(9,6)^2}{39,6} + \frac{(0,4)^2}{9,6} + \frac{(6)^2}{24} + \frac{(-6,4)^2}{26,4} = 18,9$$

d) Se determinan los grados de libertad: gl=2·2=4

e) Se establece el nivel de confianza (p=0,01)

f) Se interpreta su significatividad. Dado que el χ^2 obtenido es mayor que el que aparece en las tablas (13,277), podemos rechazar la hipótesis nula (Ho = independencia) e indicar que, al nivel establecido, se puede afirmar que hay dependencia o relación entre las dos variables. Los resultados obtenidos, por tanto, son significativamente distintos de los esperados, por lo que esa diferencia se puede

atribuir a la relación entre las variables o entre algunos de los cruces de sus categorías. También podemos comprobarlo usando la función DISTR.CHI (valor de chi; g.l.) de Excel que nos devuelve la significatividad del coeficiente por lo que dicho valor deberá ser <=0,05 para que podamos considerarlo como significativo. En este caso aparece 0,00 por lo que es significativo.

g) La forma en que están asociadas las variables se puede explicar mediante el **análisis de los residuos tipificados** de cada celda **(Rs= R/raiz(f_t)) y de los residuos corregidos o ajustados Ra= Rs/raiz((1-(tot fila/total)* (1-(tot col/total)).** Dicho análisis nos indicará qué casillas contribuyen en mayor grado al valor del estadístico. "Cuanto mayor es el valor de los residuos mayor es la probabilidad de que una determinada combinación de las variables, es decir, una casilla, sea significativa". (ALVARO,J.L./GARRIDO,A. (1995): Análisis de datos con SPSS/PC+. Madrid: Centro de Investigaciones Sociológicas, p. 68-69). Se considerarán todas aquellas celdas donde los residuos tipificados sean superiores a 1. En el ejemplo, no deberíamos considerar la casilla estudios secundarios/pop y la celda estudios primarios/clásica. Por lo que en estos casos no habría relación entre las variables.

Una vez que los residuos han sido ajustados o corregidos, siguen una distribución normal con media igual a 0 y desviación típica igual a 1 lo que nos permite decidir si el valor de cada uno de los residuos es significativo (por ej. al N.c. del 5% todos los que sean mayores de 1.96 y menores de -1.96). En el ejemplo, vuelven a aparecer niveles inferiores en las mismas casillas que con los residuos tipificados.

Cuando los residuos son positivos quiere decir que la frecuencia observada es mayor que la esperada y, cuando son negativos, que la frecuencia observada es menor que la esperada. Para la interpretación de la tabla suele seguirse la **regla de Zeisel:** los porcentajes se calculan en la dirección de la VI (en nuestro caso) el nivel de estudios) y se interpretan en la dirección de la dependiente (gusto musical). Así vemos que hay más universitarios de los esperados a los que les gusta el rock, más sujetos de los esperados con estudios primarios a los que les gusta la música clásica y no son significativas las diferencias entre lo observado y lo esperado en el resto.

5.- Observaciones

1) Para usarlo se requiere un Ṇ≥50.

2) Para calcularlo en una tabla 2x2 se puede hacer directamente (sin calcular las f_t) mediante la fórmula

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

	+	-	
+	a	b	
-	c	d	
			N

3) Corrección de Yates (tablas 2x2) o corrección de continuidad: gl=1, datos poco numerosos (entre 20 y 50) de modo que en una o más casillas de la tabla las f_t son menores que 5 o incluso menores que 10.

$$\chi^2 = \frac{N(|a\cdot d - b\cdot c| - 0{,}5N)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Si calculamos χ^2 con la fórmula normal, la ecuación de Yates consiste en restar 0,5 a las diferencias entre f_e y f_t de cada casilla en valores absolutos mediante la fórmula sería la siguiente:

$$\chi^2 = \sum \frac{(|f_e - f_t| - 0{,}5)^2}{f_t}$$

4) Cuando χ^2 se extrae de una tabla de más de 30 gl es necesario transformarlo a z ya que la distribución χ^2 a partir de dichos gl comienza a distribuirse normalmente. Una vez transformado, se interpreta como la puntuación típica z.

$$z = \sqrt{2\chi^2} - \sqrt{2gl - 1}$$

5) Cuando hay celdas con f_t (frecuencias teóricas o esperadas) menores que 5 que

superan el 20% del número total de celdas, podemos considerar que el modelo de análisis no se ajusta a la distribución de χ^2 por lo que habría que hacer lo siguiente: agrupar categorías, suprimir categorías o ampliar el número de sujetos de la muestra.

Con el cálculo de χ^2 determinamos si existe o no una asociación significativa entre las variables. Sin embargo, para determinar la cuantía o grado de la asociación necesitamos calcular C mediante la fórmula ya expuesta. En el ejemplo de autoestima y nivel de estudios resultaría lo siguiente:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{18{,}9}{250 + 18{,}9}} = 0{,}26$$

Interpretación de C:

1) Siempre es positivo, como χ^2 y, para interpretar su sentido, hay que ver cómo están organizadas las categorías en la tabla.

2) C se aproxima a 0 cuando no hay relación, pero nunca podrá llegar a la unidad ya que el denominador siempre es mayor que el numerador. El coeficiente máximo que podemos alcanzar viene dado por la siguiente fórmula: $C_{máx} = \sqrt{\frac{K-1}{K}}$ donde K es el nº de categorías (cuando en una tabla el nº de categorías en las variables es distinto se utiliza el menor).

Tabla de $C_{máx}$

K	$C_{máx}$
2	0,707
3	0,816
4	0,868
5	0,894
6	0,913
7	0,926
8	0,935
9	0,943
10	0,949

3) C no es directamente comparable con r. Sin embargo, el cociente $C/C_{máx}$ es un

equivalente de r^2 por lo que $r^2 = \frac{C}{C_{máx}}$ o $r = \sqrt{\frac{C}{C_{máx}}}$ En el ejemplo utilizado ocurriría lo siguiente

$$r^2 = \frac{0,26}{0,816} = 0,318 \text{ ó } r = \sqrt{\frac{0,26}{0,816} = 0,5644}$$

EJEMPLOS DE ANÁLISIS DE TABLAS DE CONTINGENCIA (SPSS)

Vamos a tratar de averiguar si hay asociación y en qué grado entre dos variables. En un caso cada una de ellas tiene 2 dimensiones (tabla 2x2) y en otro 3 (tabla 3x3). Para ello utilizaremos el test de χ^2 como prueba de independencia.

Dos variables son independientes si coinciden las frecuencias esperadas o teóricas (f_t) con las empíricas (f_e) o si la probabilidad de que un caso caiga en una celda es, simplemente, el producto de las probabilidades marginales de las dos categorías que definen la celda. En el ejemplo siguiente se puede comprobar que para los varones que seguirían trabajando si les tocara la lotería **se esperan** 216 casos (35% de los 616 totales), mientras que los **observados o empíricos** son 234 (18 más que los esperados: **residuos**). El cociente entre estos residuos y las frecuencias teóricas proporcionan el índice χ^2 que da cuenta de la independencia o dependencia de las variables.

Ejemplo 1: Tabla 2x2.

Intentamos ver si hay asociación (independencia o dependencia) entre dos variables: el sexo (varón, mujer) y la continuidad laboral caso de que nos tocara la lotería (si, no).

```
CROSSTABS A BY B
/CELLS COUNT EXPECTED RESID
/STAT=CHISQ.
```

A SEXO by B TOCA LOTERIA Y CONTINUA TRABAJANDO

Count Exp Val Residual	SI 1	NO 2	Row Total
VARON 1	234 216.0 18.0	89 107.0 -18.0	323 52.4%
MUJER 2	178 196.0 -18.0	115 97.0 18.0	293 47.6%
Columm Total	412 66.9%	204 33.1%	616 100.0%

Chi_Square	Value	DF	Significance
Pearson	9.48683	1	.00207
Continuity Correction	8.96619	1	.00275
Likelihood Ratio	9.49642	1	.00206
Mantel_Haenszel test for linear association	9.47144	1	.00209
Minimum Expected Frequency _	97.032		

Se puede comprobar el cálculo de las f_t y de los residuos.

Celda 1,1 $$f_t = \frac{\sum K_i \bullet \sum F_j}{N} = \frac{412 \bullet 123}{616} = 216$$

Hemos obtenido un χ^2 = 9,48683 que es significativo estadísticamente (p=0,0207). Lo que quiere decir que, para ese nivel de confianza y gl=1, rechazamos la Ho (independencia de las variables) y admitimos que hay dependencia entre las variables.

La magnitud del χ^2 obtenido depende no sólo de la bondad de ajuste al modelo de independencia sino también del tamaño de la muestra. Si el tamaño de la muestra de una tabla concreta se incrementa n veces, del igual modo lo haría el valor de χ^2. Así, pueden aparecer valores de chi cuadrado grandes en aplicaciones en las que los residuos son pequeños respecto a las frecuencias teóricas, pero donde el tamaño de la muestra es grande. Por ello se requiere que las f_t sean cuando menos 5 en las celdas. El SPSS nos indica el mínimo valor esperado que aparece y el nº(%) de celdas con f_t<5. En el ejemplo no nos da

el mensaje ya que no hay ninguna celda que reúna esta condición.

En una tabla 2x2, como es el caso, se aplica la **corrección de Yates o Corrección de continuidad** para calibrar mejor el índice χ^2. El SPSS lo realiza descontando 0,5 a los residuos positivos y sumando 0,5 a los negativos antes de su elevación al cuadrado. Compruébese cómo aparece este dato en el ejemplo de tabla 2x2 y no aparece en el siguiente (tabla 3x3). El ordenador nos da un valor de 8,96619 frente al 9,48683 de Pearson.

La **prueba exacta de Fisher** es una prueba alternativa para las tablas 2x2 que calcula las probabilidades exactas de obtener los resultados observados si las dos variables son independientes y los marginales (totales filas, totales columnas, total) están fijados. Es más útil cuando el tamaño de la muestra es pequeño y los valores esperados son pequeños. *SPSS calcula la prueba exacta de Fisher si cualquier valor esperado de las celdas en una tabla 2x2 es menor que 5*. Obsérvense las modificaciones en los datos de la Tabla siguiente:

A SEXO by B TOCA LOTERIA Y CONTINUA TRABAJANDO

Count Exp Val Residual	SI 1	NO 2	Row Total
VARON 1	20 21.6 -1.6	7 5.4 1.6	27 60.0%
MUJER 2	16 14.4 1.6	2 3.6 -1.6	18 40.0%
Columm Total	36 80.0%	9 20.0%	18 100.0%

Chi_Square	Value	DF	Significance
Pearson	1.48148	1	.22354
Continuity Correction	.70023	1	.40271
Likelihood Ratio	1.57511	1	.20947
Mantel_Haenszel test for linear association	1.44856	1	.22876
Fisher's Exact Test:			
One_Tail			.00000
Two_Tail			.00000

Minimum Expected Frequency _ 3.600
Cells with Expected Frequency < 5 _ 1 OF 4 (25.0%)
Number of Missing Observations: 0

Una vez que hemos determinado si existe o no asociación, el paso siguiente es determinar el grado de la misma. Para ello se utilizan diversos índices dependientes de χ^2. Todos ellos intentan minimizar la influencia del tamaño de la muestra, de los grados de libertad y encerrarse en un rango entre 0 y 1.

```
CROSSTABS A BY B
/CELLS COUNT EXPECTED RESID
/STAT=CHISQ PHI CC.
```

A SEXO by B TOCA LOTERIA Y CONTINUA TRABAJANDO

Count	SI 1	NO 2	Row Total
VARON 1	234	89	323 52.4%
MUJER 2	178	115	293 47.6%
Columm Total	412 66.9%	204 33.1%	616 100.0%

Chi_Square	Value	DF	Significance
Pearson	9.48683	1	.00207
Continuity Correction	8.96619	1	.00275
Likelihood Ratio	9.49642	1	.00206
Mantel_Haenszel test for linear association	9.47144	1	.00209
Minimum Expected Frequency _	97.032		

Statistic	Value	ASE1	T_value	Approximate Significance
Phi	.12410			.00207 *1
Cramer's V	.12410			.00207 *1
Contingency Coefficient	.12315			.00207 *1

*1 Pearson chi_square probability
Number of Missing Observations: 0

1) Coeficiente φ :

$$\varphi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{9{,}48683}{616}} == 0{,}12410$$

En una tabla 2x2 φ = r

Para tablas con alguna dimensión mayor de 2 categorías φ no podrá estar entre 0-1 ya que el χ^2 puede ser mayor que el N de la muestra. De ahí que para obtener una medida entre 0 y 1, Pearson sugiriese el uso del coeficiente C.

2) Coeficiente C:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{9{,}48683}{9{,}48683 + 616}} = 0{,}12315$$

Nunca llegará a 1 (ver comentarios sobre $C_{máx}$). Para lograr que el índice pudiera llegar a la unidad, Cramer introdujo la siguiente variante.

3) Coeficiente V de Cramer:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}} = \sqrt{\frac{9{,}48683}{616 \cdot (2-1)}} = 0{,}12410$$

donde k es el nº más pequeño de entre el número de filas y columnas. El índice puede alcanzar un máximo de 1 y, si una de las dimensiones es 2, V y φ serán idénticos como es el caso (0.1241).

La interpretación de los resultados se realiza en función de la significatividad y el sentido o dirección de la relación teniendo presente la tabla de contingencia. En el ejemplo, claramente se ve que en un grado muy pequeño ambos sexos seguirían trabajando y los hombres seguirían trabajando en mayor medida que las mujeres. Cuando la interpretación no está tan clara, asumimos una variable como dependiente y nos fijamos en la categoría de resultados (marginales) que incluye el mayor porcentaje de sujetos (categoría modal) es el caso de los que SI continuarían trabajando (66.9%).

Esto quiere decir que, a la hora de interpretar los índices que nos ofrece la salida de resultados, éste sería el más adecuado para determinar el grado de asociación entre las variables.

Ejemplo 2: Tabla 3x3.

Queremos ver la relación que existe entre nivel de estudios (superiores, medios y primarios) y autoestima (baja, media, alta) en base a la siguiente tabla de resultados:

```
CROSSTABS A BY B
/CELLS COUNT EXPECTED RESID ASRESIS
/STAT=CHISQ.
```

A NIVEL ESTUDIOS by B AUTOESTIMA

Count Exp Val Residual Adj Res	BAJA 1	MEDIA 2	ALTA 3	Row Total
LIC. 1	10 16.0 -6.0 -2.1	30 40.0 -10.0 -2.6	60 44.0 16.0 4.2	100 40.0%
BACH. 2	20 14.4 5.6 2.0	40 36.0 4.0 1.1	30 39.6 -9.6 -2.5	90 36.0%
PRIM 3	10 9.6 .4 .2	30 24.0 6.0 1.8	20 26.4 -6.4 -1.9	60 24.0%
Columm Total	40 16.0%	100 40.0%	110 44.0%	250 100.0%

Chi_Square	Value	DF	Significance
Pearson	18.58586	4	.00095
Likelihood Ratio	18.56849	4	.00096
Mantel_Haenszel test for linear association	10.31135	1	.00132
Minimum Expected Frequency _	9.600		
Number of Missing Observations:	0		

```
CROSSTABS A BY B
/STAT=PHI CC.
```

A NIVEL ESTUDIOS by B AUTOESTIMA

Count	BAJA 1	MEDIA 2	ALTA 3	Row Total
LIC. 1	10	30	60	100 40.0%
BACH. 2	20	40	30	90 36.0%
PRIM 3	10	30	20	60 24.0%
Columm Total	40 16.0%	100 40.0%	110 44.0%	250 100.0%

Approximate Statistic	Value	ASE1	T_value	Significance
Phi	.27266			.00095 *1
Cramer's V	.19280			.00095 *1
Contingency Coefficient	.26306			.00095 *1

*1 Pearson chi_square probability
Number of Missing Observations: 0

Interpretación:

El índice de Pearson obtenido (18,58586) es estadísticamente significativo (p=0,00095) por lo que podemos decir que hay dependencia entre nivel de estudios y autoestima. El valor esperado mínimo es 9,6 (casilla 3,1) por lo que no hay ninguno que sea menor que 5. No hay corrección de continuidad al estar trabajando con una tabla 3x3. Igualmente V y φ no coinciden por el mismo motivo.

El grado de la asociación también es significativo y nos viene determinado por C (0,26306), por φ (0,27266) y por V (0,19280). Dada la estructura de la tabla, a mayor nivel de estudios mayor autoestima y viceversa.

Sin embargo, esta relación se puede matizar por el análisis de los residuos estandarizados del SPSS. En el ejemplo concreto, vemos que las diferencias entre valores observados y esperados es positiva y significativa (>1.96) en los Lic. con autoestima alta . Es igualmente significativa, aunque negativa (valores observados menores que los esperados), en los Lic. con autoestima baja y media. En el caso de los Bach. aparecen

residuos significativos en las categorías baja y alta (negativo) y con un sentido inverso al caso de los Lic. Finalmente, en el caso de los que tienen estudios Prim. no aparecen diferencias significativas.

De todo este análisis se podría concluir que a mayor nivel de estudios mayor nivel de autoestima (concretamente en el caso de los que tienen el nivel más elevado), invirtiéndose la tendencia en los que tienen un nivel de Bach. y no siendo estadísticamente significativa en los que tienen menor nivel de estudios.

Si quisiéramos aproximarnos a la interpretación de C como una r de Pearson, hallaríamos el cociente $C/C_{máx}$ como un equivalente de r^2 por lo que $r^2 = \frac{C}{C_{máx}}$ o $r = \sqrt{\frac{C}{C_{máx}}}$

En el ejemplo utilizado el resultado sería el siguiente

$$r^2 = \frac{0{,}26306}{0{,}816} = 0{,}3222 \qquad \text{ó} \qquad r = \sqrt{\frac{0{,}26306}{0{,}816} = 0{,}5676}$$

por lo que estaríamos hablando de un 32,22% de grado de asociación entre las variables.