Medidas de asociación (variables no métricas)

- Podemos definir la asociación entre dos variables como la intensidad con la que unas categorías de una variable diferencian las frecuencias obtenidas en el cruce con la otra
  - Una primera mediada podría ser la diferencia de porcentajes Para Sánchez Carrión, J. (1995) es la mejor de todos ellas.

	M	V	Total
Opción A	15	35	50
Opción B	35	15	50
Total	50	50	100

En la tabla hay un diferencial de 20% entre Mujeres y Varones entre las opciones A y B

El diferencial porcentual varía entre:
0 <*d*> 100

El problema es que hay que calcularlo para cada casilla, de ahí que se busque un indicador único

  - El Ji-cuadrado además de determinar si son significativas estadísticamente las diferencias ya constituye por si mismo un indicador, su problema es que el valor no es estándar, depende de las frecuencias y del tamaño de la tabla

(a)

30	20	50
20	30	50
50	50	100

(b)

60<	40	50
40	60	50
100	100	200

En ambas tablas existe la misma relación un diferencia porcentual del 10% solo que la b tiene el doble de frecuencia y sus Ji-cuadrados:

$\chi_a^2 = 4{,}0$ $\chi_b^2 = 8{,}0$

- Para evitar estos problemas del Ji-Cuadrado se utilizan algunas modificaciones:
  - El ***«Phi»*** consiste en hacer la raíz cuadrada del Ji-Cuadrada dividida por el número total de casos de la tabla a fin de eliminar el problema de las frecuencias altas

$$\text{Ø} = \sqrt{\frac{x^2}{n}}$$

Su valor oscila entre 0 y 1 y es igual al coeficiente de correlación de Pearson para tablas de 2x2, pero si la tabla es mayor no tiene máximo

  - El ***«Coeficiente de contingencia»*** Intenta solucionar ese problema poniendo en el denominador de la fórmula de Phi la suma de χ^2+n

$$C = \sqrt{\frac{x^2}{x^2 + n}}$$

Plantea a su vez el problema de que nunca llega a valer 1 ni siquiera con asociación perfecta en tablas cuadradas (igual número de filas y columnas *«I»* su valor máximo es:

$$C_{máximo} = \sqrt{(I-1)/I}$$

Por lo que se puede calcular un C ajustado de la siguiente forma: $C_{ajus}=C/C_{max}$

  - El ***«Coeficiente V de Cramer»*** Sustituye en el denominador de ***«Phi»*** el valor mínimo de (I-1) o (J-1)

$$V = \sqrt{x^2 / \text{mínimo de (I-1) o (J-1)}}$$

Asociación. *Indicadores basados en la reducción de error de predicción*

- A diferencia de los anteriores basados en Ji-cuadrado. Estos tratan de ver la relación entre variables intentando predecir como se clasifica un sujeto en la variable «*Y*» a partir de conocer su clasificación en la «*X*»

Coeficiente Lambda de Goodman y Kruskal

- Llamado también «*Coeficiente de predictibilidad de Guttman»* se basa en la reducción proporcional del error en la predicción de la moda, es decir numero de aciertos que proporciona el conocer la distribución dividido por el número de errores sin conocerla.

$$\lambda_{yx} = \frac{(N - M_y) - (N - \sum m_y)}{N - M_y} = \frac{\sum m_y - M_y}{N - M_y}$$

Siendo:
M_y = la frecuencia modal global
$\sum m_y$ = la suma de frecuencias modales
N = Total de casos

El numerador sería pues el número de aciertos cometidos bajo la predicción II (conociendo la distribución de segunda variable) $\sum m_y$ menos los aciertos de la predicción I (sin conocer la distribución) M_y. Al dividir por los errores de la predicción I me debe dar una cifra entre 0 ninguna reducción (independencia total ya que una variable no predice la otra o 1 si la puede predecir de forma total.

- Tras el hundimiento del Titanic de las 1285 personas que viajaban en él perecieron 800 y murieron 485 en función del sexo la distribución fue:

	V	M	Total	%
Mueren	637	163	800	62,3
Sobreviven	138	347	485	37,7
Total	775	510	1285	100
%	60,3	39,7		

Si pretendo acertar el destino de un pasajero cualquiera, sin saber nada más, me aventuraría por decir que murió, ya que fueron mayoría los que perecieron (intervalo modal) y tendría una posibilidad de errar de M_y=485

➡ Sabiendo que es hombre la posibilidad de que fallara mi pronóstico sería m_1= 138 Por el contrario si se que es mujer, la posibilidad de errar es m_2=163. El error al conocer la distribución de la segunda variable es menor que si no la conozco.

Error univariado bajo la predicción de la frecuencia modal global = 485

Error bivariado si es hombre =138
Error bivariado si es mujer =163
} *Error bivariado total =138+163=301*

$E_1 = N - M_y = 1285 - 800 = 485$
$E_2 = N - \sum m_y = 1285 - (637 + 347) = 1.285 - 984 = 301$

$$\lambda_{yx} = \frac{\sum m_y - M_y}{N - M_y} = \frac{984 - 800}{1285 - 800} = \frac{184}{485} = 0{,}37$$

También Lambda = E_1-E_2/E_1=(485-301)/485=184/485=0,379

- Imaginemos la siguiente distribución de familias según tipo de familia y situación del cabeza de familia

	Cabeza familia varón		Cabeza familia mujer		
	Con hijos menores	Sin hijos menores	Con hijos menores	Sin hijos menores	Total
Casado	6.444	4.804	78	50	11.376
Separado	20	126	250	106	502
Divorciado	19	237	284	276	816
Viudo	47	300	236	1.614	2197
Total	6.530	5.467	848	2.046	14.891

Sabiendo que el cabeza de familia es varón con hijos menores el valor modal sería casado, acertaríamos 6444 veces de 6530 es decir fallaríamos en 86 ocasiones. Sabiendo que tiene sería 5467-4804=663 errores. En el caso de ser mujer con hijos la situación modal seria de divorciada esto es 848-284=564 errores y si no tiene hijos sería viuda con 2046-1614=432. Total de errores = 86+663+564+432=1745

$\sum m_y$ = 6.444+4.804+284+1.614 = 13.146 aciertos conociendo la distribución
Aplicando la formula de Lambda obtendríamos una reducción del error de:

Modales parciales
Modal global

$$\lambda_{yx} = \frac{\sum m_y - M_y}{N - M_y} = \frac{13.146 - 11.376}{14.891 - 11.376} = \frac{1.170}{3.515} = 0{,}333$$

- Lambda es un coeficiente asimétrico, eso quiere decir que si en la misma tabla intentamos adivinar la composición familiar sabiendo el estado del cabeza de familia nos daría un resultado diferente

	Cabeza familia varón		Cabeza familia mujer		
	Con hijos menores	Sin hijos menores	Con hijos menores	Sin hijos menores	Total
Casado	6.444	4.804	78	50	11.376
Separado	20	126	250	106	502
Divorciado	19	237	284	276	816
Viudo	47	300	236	1.614	2197
Total	6.530	5.467	848	2.046	14.891

$\sum m_y$ = 6.444+250+284+1.614 = 8592 aciertos conociendo la distribución del estado del cabeza de familia

Modales parciales
Modal global

$$\lambda_{yx} = \frac{\sum m_y - M_y}{N - M_y} = \frac{8.592 - 6.530}{14.891 - 6.530} = \frac{2.062}{8.361} = 0{,}246$$

- La situación familiar permite una reducción del error del 24,6% mientras que el conocer el tipo de familia permitía lo hacía en 33,3%. Lambda permite reconocer la variable más predictora

Coeficiente Tau-y de Goodman y Kruskal

- Al igual que el Lambda es un coeficiente asimétrico pero a diferencia de éste parte de los errores cometidos al asignar aleatoriamente los casos a las categorías de la variable dependiente.
  - En definitiva supone que en cada categoría se clasificarán erróneamente por puro azar un cierto número de casos, que es igual en cada categoría al número de casos que no pertenecen a la misma. Así en la categoría de casados de los 11.376 casos de un total de 14.891 sujetos, se cometerían 14.891-11.376 = 3.515 errores por lo que si intentásemos designar al azar los 11.376 casos de casados cometeríamos un promedio de errores de:

$$\frac{14.891-11.376}{14.891}\times 11.376 = \frac{3.515}{14.891}\times 11.376 = 2.687,7$$

  - Simbólicamente la formula para las predicciones del tipo I (categorías de la variable dependiente) sin conocer la distribución de la independiente sería:

$$E_1 = \sum_{i=1}^{k}\left[\frac{N-f_i}{N}\times f_i\right]$$

Siendo N el número total de casos, k el número de categorías de la variable e f_i la frecuencia de la categoría i

- Para calcular los errores bajo la predicción I (sin conocer la distribución de la variable independiente) sería:

	Cabeza familia varón		Cabeza familia mujer		
	Con hijos menores	Sin hijos menores	Con hijos menores	Sin hijos menores	Total
Casado	6.444	4.804	78	50	11.376
Separado	20	126	250	106	502
Divorciado	19	237	284	276	816
Viudo	47	300	236	1.614	2197
Total	6.530	5.467	848	2.046	14.891

Errores para la categoría de casados $\frac{14.891-11.376}{14.891}\times 11.376 = 2.685,29$

Errores para la categoría de separados $\frac{14.891-502}{14.891}\times 502 = 485,08$

Errores para la categoría de divorciados $\frac{14.891-816}{14.891}\times 816 = 771,28$

Errores para la categoría de viudos $\frac{14.891-2197}{14.891}\times 2197 = 1872,86$

Total errores del tipo I E_1= 2.685,29 + 485 ,08 + 771,28 + 1.872,86 = 5.814,51

- Para calcular los errores bajo la predicción II (conociendo la distribución de la variable independiente) se utiliza la formula:

$$E_2 = \sum_{i=1}^{c} \sum^{k} \left[\frac{N_i - n_i}{N_i} \times n_i \right]$$

Siendo n_i la frecuencia de la celdilla en la categoría i de la variable dependiente dentro de cada una de las c categorías de la variable independiente y N_i el total parcial de las categorías de la variable independiente

	Cabeza familia varón	
	Con hijos menores	Sin hijos menores
Casado	6.444	4.804
Separado	20	126
Divorciado	19	237
Viudo	47	300
Total	6.530	5.467

⇨ Para la categoría de cabeza de familia varón con hijos sería:

Errores para la categoría de casados $\frac{6.530 - 6.444}{6.530} \times 6.444 = 84,86$

Errores para la categoría de separados $\frac{6530 - 20}{6.530} \times 20 = 19,93$

Errores para la categoría de divorciados $\frac{6.530 - 19}{6.530} \times 19 = 18,84$

Errores para la categoría de viudos $\frac{6530 - 47}{6.530} \times 47 = 46,66$

Errores en ésta categoría E_{21}=84,86 + 19,93 + 18,84 + 46,66 =170,39

⇨ Para la categoría de cabeza de familia varón sin hijos menores sería:

	Cabeza familia varón	
	Con hijos menores	Sin hijos menores
Casado	6.444	4.804
Separado	20	126
Divorciado	19	237
Viudo	47	300
Total	6.530	5.467

Errores en la categoría de casados $\frac{5.467 - 4.804}{5.467} \times 4.804 = 582,60$

Errores en la categoría de separados $\frac{5.467 - 126}{5.467} \times 126 = 123,10$

Errores en la categoría de divorciados $\frac{5.467 - 237}{5.467} \times 237 = 226,73$

Errores en la categoría de viudos $\frac{5.467 - 300}{5.467} \times 300 = 283,54$

Errores en ésta categoría de padres varones sin hijos menores:

E_{22}= 582,60 + 123,10 + 226,73 + 283,54 =1215,96

Para la categoría de cabeza de familia mujer con hijos menores sería:

	Cabeza familia mujer	
	Con hijos menores	Sin hijos menores
Casado	78	50
Separado	250	106
Divorciado	284	276
Viudo	236	1.614
Total	848	2.046

Errores en la categoría de casados $\frac{848-78}{848}\times 78 = 70{,}83$

Errores en la categoría de separados $\frac{848-250}{848}\times 250 = 176{,}30$

Errores en la categoría de divorciados $\frac{848-284}{848}\times 284 = 188{,}89$

Errores en la categoría de viudos $\frac{848-236}{848}\times 236 = 170{,}32$

Errores en ésta categoría de padres varones sin hijos menores:

E_{23}= 70,83 + 176,30 + 188,89 + 170,32 =606,33

Para la categoría de cabeza de familia mujer sin hijos menores sería:

	Cabeza familia mujer	
	Con hijos menores	Sin hijos menores
Casado	78	50
Separado	250	106
Divorciado	284	276
Viudo	236	1.614
Total	848	2.046

Errores en la categoría de casadas $\frac{2.046-50}{2.046}\times 50 = 48{,}78$

Errores en la categoría de separadas $\frac{2.046-106}{2.046}\times 106 = 100{,}51$

Errores en la categoría de divorciadas $\frac{2.046-276}{2.046}\times 276 = 238{,}77$

Errores en la categoría de viudas $\frac{2.046-1.614}{2.046}\times 1.614 = 340{,}79$

Errores en ésta categoría de padres varones sin hijos menores:

E_{24}= 48,78 + 100,51 + 238,77 + 340,79 =728,84

- Los errores del tipo E_2 será igual a la suma de todos los ΣE_{2i}

E2 = 170,39 +1215,96 + 606.33 +728,84 = 2.722

- Conocidos los errores E1 y E2 bajo la predicción del tipo I (sin conocer la distribución de la variable independiente) y tipo II (conociéndola) el coeficiente Tau-y se calcula mediante la fórmula

$$Tau - y = \frac{E_1 - E_2}{E_1}$$

➡ En nuestro caso teniendo en cuenta que $E_1 = 5814{,}51$ y $E_2 = 2722$

$$Tau - y = \frac{5.814{,}51 - 2.722}{5.814{,}51} = 0{,}53$$

- Así, pues, el coeficiente *«Tau-y»* obtenido significa que se ha reducido en un 53% los errores cometidos al predecir la colocación de los casos en la variable dependiente, mediante la información suministrada por la distribución de la independiente

Asociación. *Indicadores para variables ordinales*

- Para variables en escala ordinal son más adecuados otro tipo de indicadores de asociación. En estos casos se trata de saber si el conocer la ordenación de los casos en una variable resulta útil para predecir el orden de la otra
  - ➡ Hablaremos de ***«asociación positiva»*** cuando el tipo de ordenación predice de alguna manera la misma ordenación en la segunda (A mayor edad mayor desconfianza hacia los demás)
  - ➡ Hablaremos de ***«asociación negativa»*** cuando el tipo de ordenación predice de alguna manera una ordenación opuesta en la segunda variable (A mayor edad mayor menor nivel de estudios)
- En las variables ordinales más que buscar la existencia o no de relación, lo que nos interesa conocer es la información sobre el orden en que medida crece o disminuye la dependiente al crecer la independiente y viceversa.

➡ Esta tabla tiene dos variables ordinales X e Y con 2 y 3 categorías las casillas las hemos identificado con letras y en cada una se recogen las frecuencias correspondientes. Vamos a considerar las parejas de casillas que podemos formar

Variable Y		*Variable X* 1	2
	1	A=20	B=5
	2	C=15	D=20
	3	E=10	F=15

Parejas de casillas ***«concordantes»*** denominaremos así a parejas como la formada por las casillas *A* y *D* (AD) pues puntúan igual o coincide el signo de su orden en ambas variables

Variable X

Variable Y	1	2
1	A=20	B=5
2	C=15	D=20
3	E=10	F=15

El orden de A es 1 (variable Y) y 1 (variable X)
El orden de D es 2 (variable Y) y 2 (variable X)

La pareja AD la forman sujetos que al crecer A crece también B, son coincidentes en el sentido del orden al estar por encima de los de A en ambas variables lo mismo ocurre con las parejas *AF* y *CF* los sujetos de F están en ambos casos por encima de los de *A* y también con respecto a *C*

Parejas de casillas ***«discordantes»*** denominaremos así a parejas como la formada por las casillas B y C (AD) pues puntúan igual o coincide el signo de su orden en ambas variables

En la pareja *BC* orden de *B* es 1 (variable Y) y 2 (variable X) y el orden de *C* es 2 (variable Y) y 1 (variable X) luego al aumentar en una disminuye en la otra variable siendo discordantes.

Igualmente lo ocurre con la pareja *BE*, y *DE* . En la primera al crecer el valor *X a 2* en para *B* disminuye *1* en *X* para *E*.
En la segunda DE, cuando en la casilla *D* vale en la *X 2* para la casilla *E* vale *1*

Los sujetos de la pareja *A* y *B* están empatados en la variable Y denominamos ***«empatadas»*** a las parejas que coinciden en valor en alguna o ambas variables

Empatadas en *X* serían las parejas C_x= *AB, CD y EF*
Empatadas en Y serían C_y= *AC, AE, CE, BD, BF, DF*

Por tanto, las posibles parejas de sujetos que podemos hacer con cada cruce será igual al producto de sus frecuencias

Variable X

Variable Y	1	2
1	A=20	B=5
2	C=15	D=20
3	E=10	F=15

Parejas concordantes	*Parejas discordantes*
AD 20 x 10 = 200	BC 5 x 15 = 75
AF 20 x 15 = 300	BE 5 x 10 = 50
CF 15 x 15 = 225	CF 10 x 10 = 100
725	225

- Coeficiente ***«Gamma»*** se calcula bajo la fórmula:

$$Gamma = \frac{P-Q}{P+Q} \text{ en nuestro ejemplo } \frac{725-225}{725+225} = \frac{500}{1050} = 0{,}53$$

Donde *P* es el producto de las parejas concordantes y *Q* el de las discordantes

El coeficiente Gamma varia entre –1,0 y +1,0 y se puede interpretar como la reducción proporcional del error cometido al predecir el ordenamiento de los casos de una variable mediante el conocimiento de la ordenación en la otra

- Coeficiente ***«D de Sommers»*** se calcula bajo la fórmula:

$$D_{yx} = \frac{P - Q}{P + Q + T_y} \quad \text{o bien} \quad D_{xy} = \frac{P - Q}{P + Q + T_x}$$

Donde T_x y T_y son las parejas empatadas en X e Y respectivamente. En nuestro ejemplo:

Parejas empatadas en X	
AC 20 x 15 =	300
AE 20 x 10 =	200
CE 15 x 10 =	150
BD 5 x 20 =	100
BF 5 x 15 =	75
DF 20 x 15 =	300
	1125

Parejas empatadas en Y	
AB 20 x 5 =	100
CD 15 x 20 =	300
EF 10 x 15 =	150
	550

Variable Y	*Variable X* 1	2
1	A=20	B=5
2	C=15	D=20
3	E=10	F=15

$$D_{yx} = \frac{725 - 225}{725 + 255 + 550} = \frac{500}{1500} = 0{,}33$$

Considerando la variable X como dependiente

$$D_{xy} = \frac{725 - 225}{725 + 255 + 1125} = \frac{500}{2075} = 0{,}24$$

Considerando la variable Y como dependiente

➡ El coeficiente D de Sommer es una medida asimétrica como el coeficiente Lambda, los dos valores que se pueden obtener de una misma tabla dependen de que se tome como independiente la variable X o Y

- Coeficiente ***«Tau b»*** se calcula bajo la fórmula:

$$Tau_B = \frac{P - Q}{\sqrt{(P + Q + T_y) \times (P + Q + T_x)}}$$

Donde T_x y T_y son las parejas empatadas en X e Y respectivamente. En nuestro ejemplo:

Parejas empatadas en X	
AC 20 x 15 =	300
AE 20 x 10 =	200
CE 15 x 10 =	150
BD 5 x 20 =	100
BF 5 x 15 =	75
DF 20 x 15 =	300
	1125

Parejas empatadas en Y	
AB 20 x 5 =	100
CD 15 x 20 =	300
EF 10 x 15 =	150
	550

Variable Y	*Variable X* 1	2
1	A=20	B=5
2	C=15	D=20
3	E=10	F=15

$$Tau_B = \frac{725 - 225}{\sqrt{(725 + 225 + 550) \times (725 + 225 + 1125)}} = \frac{500}{\sqrt{1500 \times 2075}} = 0{,}28$$

➡ El coeficiente Tau B varia entre -1 y +1 según sea el sentido de la asociación, sin embargo cuando la tabla no es cuadrada (no tiene el mismo número de filas y columnas) este coeficiente no puede llegar a valer 1dado que existirán más pares empatados en la variable que tenga más categorías

- Coeficiente ***«Tau C»*** se calcula bajo la fórmula:

$$Tau_C = \frac{2m(P-Q)}{n^2(m-1)}$$

Donde *m* es el mínimo del número de filas o columnas y *n* el tamaño de la muestra. En nuestro ejemplo:

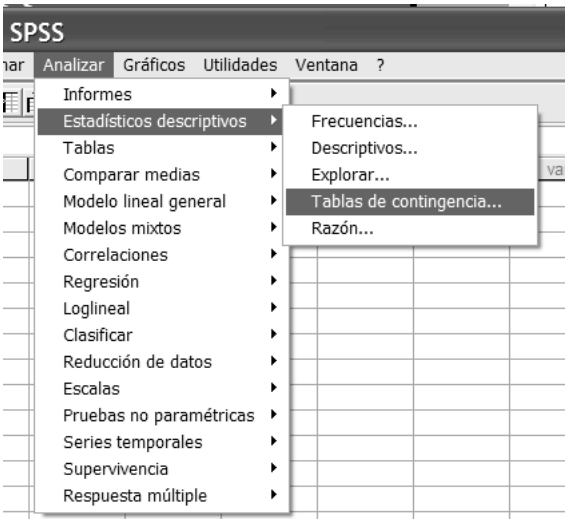
$$Tau_C = \frac{22(725-225)}{75^2(2-1)} = \frac{2000}{7225} = 0{,}35$$

		Variable X		
		1	**2**	
Variable Y	**1**	A=20	B=5	25
	2	C=15	D=20	35
	3	E=10	F=15	25
		45	40	85

➡ El coeficiente Tau C varia entre -1 y +1 según sea el sentido de la asociación, eliminando algunos de los inconvenientes del Tau B

Medidas de asociación en el SPSS

- Para ver los coeficientes de asociación en SPSS debemos entrar en el menú *Analizar →Estadísticos descriptivos→Tablas de contingencia.*



- Una vez que aparezca el menú emergente de tablas pulsar sobre *Estadísticos*

- En este nuevo menú activar aquellos coeficientes que se deseen y sean adecuados al tipo de variables y caso de que se trate