MEDIDAS DE ASOCIACIÓN PARA DATOS NOMINALES

El botón *Estadísticos* permite calcular las medidas de asociación más utilizadas para variables nominales y ordinales. Por defecto no hay ninguna opción activada.

Algunos de los estadísticos que recoge este cuadro de diálogo son:

1. *Chi-cuadrado*, con esta opción se obtienen los estadísticos:

- Chi-cuadrado de Pearson: $\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$ (donde *nij* y *eij* son las frecuencias absoluta observada y esperada, respectivamente). Si la tabla es 2x2, el estadístico se puede calcular mediante la expresión abreviada $\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$ (donde *ni.* y *nj.* son las frecuencias marginales por filas y columnas). Este estadístico es fiable únicamente cuando por lo menos el 80% de las frecuencias esperadas son mayores que 5.
- Corrección de continuidad (de Yates): en las tablas 2x2 corrige el error que se comete al calcular el estadístico Chi-cuadrado de Pearson debido a la aproximación de una distribución discreta por una continua. El estadístico corregido es $\chi^2 = \frac{n(|n_{11}n_{22} - n_{12}n_{21}| - 0,5n)^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$.
- Contraste de la razón de verosimilitud: se basa en el estadístico G^2 que se calcula como $G^2 = -2\sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \ln\left(\frac{e_{ij}}{n_{ij}}\right)$.
- Prueba exacta de Fisher: si la tabla es 2x2 y los totales marginales se consideran fijos, la probabilidad exacta de obtener la tabla analizada bajo el supuesto de independencia es:

$$p = \frac{(n_{11} + n_{12})!(n_{21} + n_{22})!(n_{11} + n_{21})!(n_{12} + n_{22})!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}.$$

Si la variable es *Nominal* los estadísticos que se pueden calcular son los siguientes:

2. Coeficiente de Contingencia de Pearson: se define como $C=\sqrt{\frac{\chi^2}{\chi^2+n}}$. Esta medida de asociación no se ve afectada por el tamaño de la muestra y está acotada, $0 \leq C < 1$. Si las variables son independientes se tiene C=0, pero en caso de asociación perfecta nunca alcanza el valor 1. Para tablas cuadradas RxR se puede calcular la cota superior que es $\sqrt{\frac{R-1}{R}}$; en tal caso $0 \leq C \leq \sqrt{\frac{R-1}{R}}$.

3. Phi y V de Cramer:

- El coeficiente Phi es $\Phi = \sqrt{\frac{\chi^2}{n}}$. Este coeficiente toma el valor 0 en caso de independencia, pero su valor máximo depende de las dimensiones de la tabla; para tablas superiores a 2xJ o Ix2 el coeficiente Φ puede tomar valores mayores que 1.

- El coeficiente V de Cramer se calcula también en función de χ^2 :

$V=\sqrt{\frac{\chi^2}{n(k-1)}}$ (donde k = mín (I,J)) y siempre está acotado entre 0 y 1 (sea cual sea la dimensión de la tabla). Para tablas 2xJ o Ix2 el valor de V coincide con el de Φ.

4. *Lambda:* Incluye la λ simétrica y asimétrica y la τ de Goodman y Kruskal. Ambas medidas se basan en la reducción proporcional del error de predicción cuando se utilizan los valores de la variable independiente para predecir los de la dependiente.

- El coeficiente λ de Kruskal se calcula:

- Si la X_1 es la variable dependiente:

$$\lambda_{x1} = \frac{(f_{m1} + f_{m2} + \ldots + f_{mJ}) - f_{m.}}{1 - f_{m.}}$$

donde *fmj* es la frecuencia relativa conjunta máxima correspondiente a la columna j-ésima y $f_{m.}$. es la frecuencia relativa marginal máxima de X_1 El valor de este estadístico, acotado entre 0 y 1, indica en cuanto se reduce la incertidumbre de la variable X_2 cuando se dispone de información sobre el comportamiento de X_1.

- De forma análoga se calcula λ_{x2}.

- Si no se puede determinar cual de las dos variables es la dependiente y cual la independiente se calcula el coeficiente λ simétrica como:

$$\lambda = \frac{(1 - f_{m.})\lambda_{x1} + (1 - f_{.m})\lambda_{x2}}{2 - (f_{m.} + f_{.m})}$$

El valor de λ simétrica está comprendido entre λ_{x1} y λ_{x2}.

Un valor λ igual a 0 indica que la información acerca de la variable independiente no ayuda en absoluto a predecir los valores de la variable dependiente; mientras que un valor λ igual a 1 indica asociación predictiva perfecta entre las variables.

- La τ de Goodman y Kruskal indica la reducción en el error de clasificación de los elementos para una de las variables (dependiente) cuando se tiene información sobre el comportamiento de la otra (independiente). Si X_1 es la variable dependiente, se calcula: $\tau_1 = \frac{TMCSI - TMCCI}{TMCSI}$ siendo TMCSI=total de elementos mal clasificados sin información acerca de X_2. y TMCCI=total de elementos mal clasificados con información acerca de X_2. La τ_2 se define de manera análoga.

5. *Coeficiente de incertidumbre, U*: mide la reducción proporcional del error cuando los valores de una variable se emplean para predecir valores de la otra.

En el cuadro *Nominal por intervalo* se encuentra el estadístico:

6. *Eta:* Este coeficiente es apropiado cuando la variable dependiente se mide en una escala de intervalo, mientras que la independiente es una variable nominal. El programa muestra dos valores de *Eta* según se considere que la variable medida en una escala de intervalo esté en las filas o en las columnas.

MEDIDAS DE ASOCIACIÓN PARA DATOS ORDINALES

En el cuadro de diálogo *Tablas de contingencia*: *Estadísticos* pueden activarse diversas opciones que proporcionan medidas de asociación cuando las variables se miden por lo menos en una escala ordinal; las más utilizadas son:

1. *Correlaciones*: con esta opción se obtienen los estadísticos:

- *Coeficiente de correlación de Pearson*: es una medida de asociación lineal adecuada para variables medidas en escala de intervalo *.

- *Coeficiente de correlación de Spearman*: mide el grado de correspondencia que existe entre los rangos que se asignan a los valores de las variables analizadas. Por ello, este coeficiente se puede calcular con datos ordinales, y se define: $r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$, siendo *d*i la diferencia entre los rangos correspondientes a la observación i-ésima. El coeficiente toma valores entre -1 y +1. Un valor cercano a 0 indica que las variables apenas están relacionadas.

El cuadro *Ordinal* recoge una serie de estadísticos basados en el número de concordancias y discordancias que aparecen al comparar las puntuaciones asignadas a los mismos casos según dos criterios (o jueces) diferentes. Así, por ejemplo, si X_1 recoge las puntuaciones asignadas a los casos según el primer criterio y X_2, según el segundo, para la obtención de concordancias y discordancias que aparecen entre los dos criterios, se procede de la siguiente forma:

- se ordenan los pares de puntuaciones de acuerdo con el orden natural de las puntuaciones asignadas según el primer criterio, X_1.

- se compara cada valor de X_2, con cada uno de los que le siguen, y se registra una concordancia (+1) cuando los dos valores siguen el orden natural, una discordancia (-1) cuando el orden está invertido y un empate (0) cuando coinciden ambas puntuaciones.

- se calculan C total de las concordancias, D total de las discordancias y E el número total de empates.

El número total de comparaciones es $\frac{n(n-1)}{2}$ incluyendo empates.

1. *Gamma*: El estadístico Gamma se define como $\frac{C-D}{C+D}$. Este análisis excluye los casos que presentan la misma puntuación en las dos variables (empates).

2. *Tau-b de Kendall*. Este coeficiente incluye los empates contemplando por separado los que aparecen en la variable $X_1\ (E_{x1})$ y los que aparecen en la variable $X_2\ (E_{x2})$.

Se define como $\tau_b = \frac{C-D}{\sqrt{(C+D+E_{x1})(C+D+E_{x2})}}$.

3. *Tau-c de Kendall*. Este estadístico se define como $\tau_c = \frac{2\,k\,(C-D)}{n^2(k-1)}$ siendo k el menor número de casos no empatados que presentan X_1 o X_2.

4. *d de Somers*: A diferencia de los anteriores este estadístico considera que las variables pueden ser simétricas o dependientes. En el primer caso, el estadístico *d de Somers* coincide con la *Tau-b de Kendall*. En el segundo supuesto, se diferencia del estadístico *Gamma* en que incluye los empates de la variable que considera dependiente. Si la variable dependiente es X_1, $d = \frac{C-D}{C+D+E_{x1}}$.

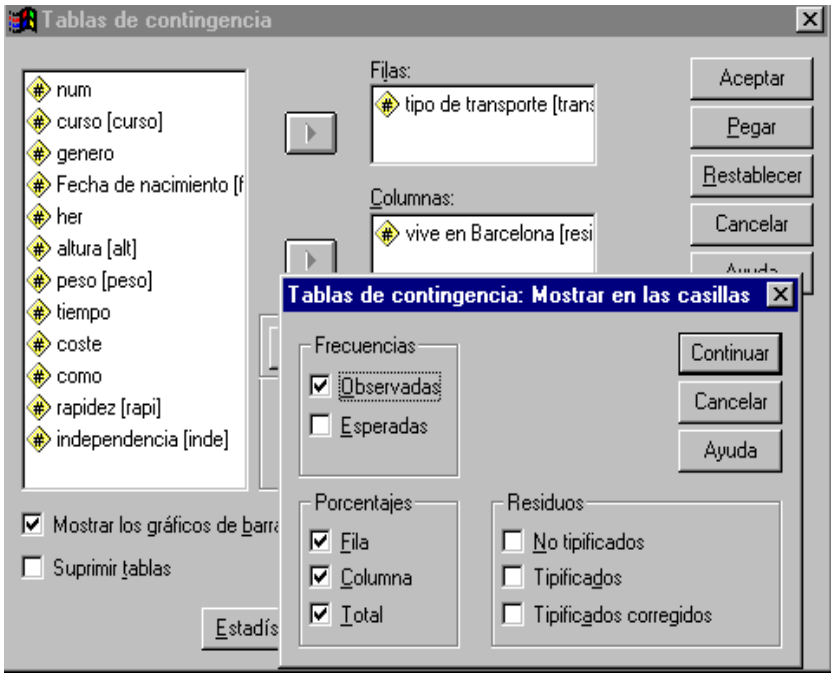
Todas estas medidas toman valores entre -1 y +1, y alcanza los valores extremos cuando existe concordancia o discordancia perfecta. Valores próximos a 0 indican ausencia de asociación.

EJEMPLO

Ejemplo 1.

Obtenga una tabla de contingencia con las frecuencias absolutas, relativas y condicionadas para las variables y el correspondiente diagrama de barras.

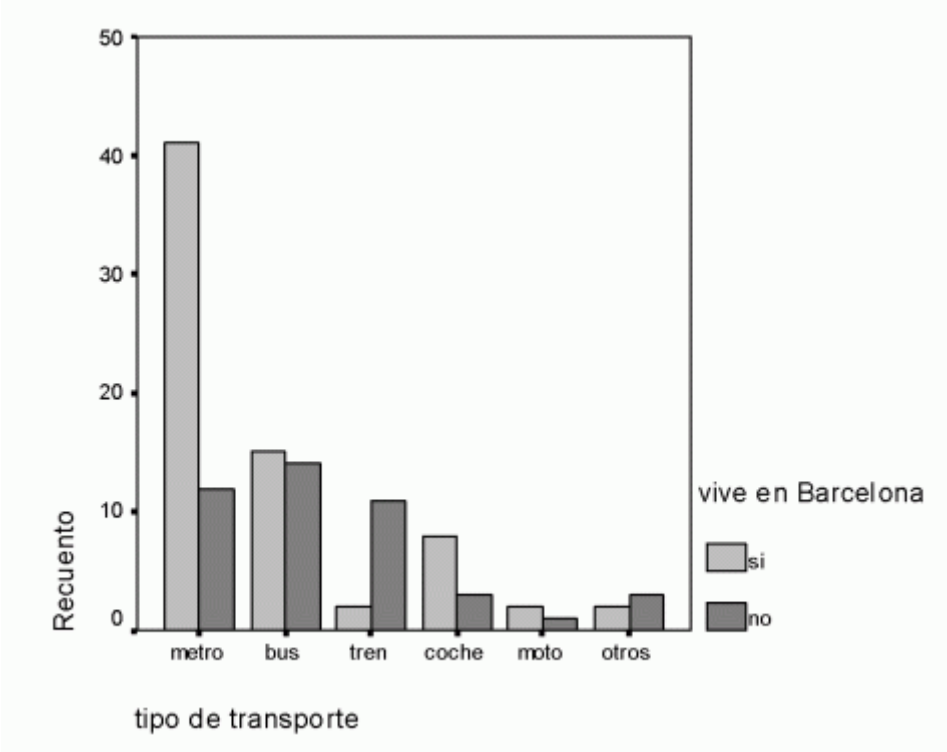
En el cuadro de diálogo *Tablas de contingencia* se activa la opción *Mostrar los gráficos de barras agrupadas*, y en el cuadro de diálogo *Tablas de contingencia* > *Casillas* se activan las opciones *Porcentajes Fila, Columna y Total*.



Los resultados que se obtienen son los siguientes:

Tabla de contingencia tipo de transporte * vive en Barcelona

			vive en Barcelona		Total
			si	no	
tipo de transporte	metro	Recuento	41	12	53
		% de tipo de transporte	77,4%	22,6%	100,0%
		% de vive en Barcelona	58,6%	27,3%	46,5%
		% del total	36,0%	10,5%	46,5%
	bus	Recuento	15	14	29
		% de tipo de transporte	51,7%	48,3%	100,0%
		% de vive en Barcelona	21,4%	31,8%	25,4%
		% del total	13,2%	12,3%	25,4%
	tren	Recuento	2	11	13
		% de tipo de transporte	15,4%	84,6%	100,0%
		% de vive en Barcelona	2,9%	25,0%	11,4%
		% del total	1,8%	9,6%	11,4%
	coche	Recuento	8	3	11
		% de tipo de transporte	72,7%	27,3%	100,0%
		% de vive en Barcelona	11,4%	6,8%	9,6%
		% del total	7,0%	2,6%	9,6%
	moto	Recuento	2	1	3
		% de tipo de transporte	66,7%	33,3%	100,0%
		% de vive en Barcelona	2,9%	2,3%	2,6%
		% del total	1,8%	,9%	2,6%
	otros	Recuento	2	3	5
		% de tipo de transporte	40,0%	60,0%	100,0%
		% de vive en Barcelona	2,9%	6,8%	4,4%
		% del total	1,8%	2,6%	4,4%
Total		Recuento	70	44	114
		% de tipo de transporte	61,4%	38,6%	100,0%
		% de vive en Barcelona	100,0%	100,0%	100,0%
		% del total	61,4%	38,6%	100,0%



En la tabla de contingencia se observan, entre otros resultados, los siguientes:

- Un total de 53 personas utilizan el metro de las cuales 41 viven en Barcelona y 12 no.

- El 36% del total de casos de la muestra utilizan el metro y viven en Barcelona.

- El 58,6% de los que viven en Barcelona utilizan el metro.

- El 77,4% de los que utilizan el metro viven en Barcelona.

- El 10,5% del total de casos utilizan el metro y no viven en Barcelona.

- El 27,3% de los que no viven en Barcelona utilizan el metro.

-El 22,6% de los que utilizan el metro no viven en Barcelona.

Análogamente se interpretan el resto de resultados.

Ejemplo 2.

Obtenga la tabla de contingencia para las variables del ejemplo anterior con frecuencias observadas, esperadas y residuos no tipificados.

Tabla de contingencia tipo de transporte * vive en Barcelona

			vive en Barcelona		
			si	no	Total
tipo de transporte	metro	Recuento	41	12	53
		Frecuencia esperada	32,5	20,5	53,0
		Residual	8,5	-8,5	
	bus	Recuento	15	14	29
		Frecuencia esperada	17,8	11,2	29,0
		Residual	-2,8	2,8	
	tren	Recuento	2	11	13
		Frecuencia esperada	8,0	5,0	13,0
		Residual	-6,0	6,0	
	coche	Recuento	8	3	11
		Frecuencia esperada	6,8	4,2	11,0
		Residual	1,2	-1,2	
	moto	Recuento	2	1	3
		Frecuencia esperada	1,8	1,2	3,0
		Residual	,2	-,2	
	otros	Recuento	2	3	5
		Frecuencia esperada	3,1	1,9	5,0
		Residual	-1,1	1,1	
Total		Recuento	70	44	114
		Frecuencia esperada	70,0	44,0	114,0

La observación de los residuos permite tener una primera aproximación sobre la existencia de asociación entre las variables. Si los residuos en valor absoluto son próximos a 0 se espera que la hipótesis de independencia entre las variables no se pueda rechazar. Por el contrario, cuanto mayores sean los valores absolutos de los residuos se tendrán más indicios sobre la existencia de asociación. En cualquier caso, la confirmación de la existencia o no de asociación entre las variables se obtiene a partir de los estadísticos correspondientes. En este ejemplo se observan unos valores de los residuos, en general, elevados, lo que nos lleva a pensar en la existencia de asociación entre el tipo de transporte y el lugar de residencia.

Ejemplo 3.

Obtenga el coeficiente Chi-cuadrado, coeficiente de contingencia, Phi y V de Cramer e interprete los resultados.

En el cuadro de diálogo *Tablas de contingencia*: *Estadísticos* se activan las opciones correspondientes. Los resultados aparecen en los siguientes cuadros.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asint. (bilateral)
Chi-cuadrado de Pearson	20,052[a]	5	,001
Razón de verosimilitud	20,584	5	,001
Asociación lineal por lineal	4,558	1	,033
N de casos válidos	114		

a. 5 casillas (41,7%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 1,16.

Como se ha visto en el ejemplo anterior, la tabla de contingencia de las variables Trans y Resi presenta 5 casillas con frecuencia esperada inferior a 5, lo que representa un 41,7% de las casillas. En estas condiciones, los resultados del contraste Chi-cuadrado no son fiables. Para resolver este problema se agrupan clases hasta obtener frecuencias esperadas superiores a 5. Agrupando Moto y Otros se obtiene una clase que tiene frecuencia esperada prácticamente igual a 5. Para agruparlas se recodifica la variable Trans en una nueva variable (Trans2) manteniendo todos los valores de Trans excepto los correspondientes a Moto y Otros, a los que se les asigna un único valor.

Repitiendo el análisis se obtiene:

Tabla de contingencia TRANS2 * vive en Barcelona

			vive en Barcelona		
			si	no	Total
TRANS2	Metro	Recuento	41	12	53
		Frecuencia esperada	32,5	20,5	53,0
	Bus	Recuento	15	14	29
		Frecuencia esperada	17,8	11,2	29,0
	Tren	Recuento	2	11	13
		Frecuencia esperada	8,0	5,0	13,0
	Coche	Recuento	8	3	11
		Frecuencia esperada	6,8	4,2	11,0
	Otros	Recuento	4	4	8
		Frecuencia esperada	4,9	3,1	8,0
Total		Recuento	70	44	114
		Frecuencia esperada	70,0	44,0	114,0

Pruebas de chi-cuadrado

	Valor	gl	Sig. asint. (bilateral)
Chi-cuadrado de Pearson	19,490[a]	4	,001
Razón de verosimilitud	20,043	4	,000
Asociación lineal por lineal	3,728	1	,054
N de casos válidos	114		

a. 3 casillas (30,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 3,09.

Como puede observarse en la tabla de contingencia hay tres frecuencias esperadas menores que 5, es decir, el 30% del total; no obstante, una de dichas frecuencias es igual a 4,9 que a efectos prácticos puede considerarse igual a 5. Por tanto, sólo el 20% de las frecuencias esperadas es inferior a 5 y en estas condiciones los resultados del contraste Chi-cuadrado son fiables. El valor del estadístico Chi-cuadrado* es 19,490 y la razón de verosimilitud 20,043. Estos valores difieren significativamente de 0 para niveles de significación superiores a 0,001, lo que significa que se rechaza la hipótesis de independencia para los niveles de significación habituales.

Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,413	,001
	V de Cramer	,413	,001
	Coeficiente de contingencia	,382	,001
N de casos válidos		114	

a. No asumiendo la hipótesis nula.

b. Empleando el error típico asintótico basado en la hipótesis nula.

***** NOTA A PIE ** Obsérvese que no se ha realizado la corrección de continuidad de Yates, porque la tabla no es de dimensiones 2x2.***

-Las medidas basadas en el estadístico Chi-cuadrado, que son los coeficientes Phi y V de Cramer, toman ***ambas**** el valor 0,413 con un nivel de significación crítico 0,001; lo que implica la existencia de asociación moderadamente fuerte entre las variables.

** ***NOTA A PIE ** En este caso, ambos coeficientes coinciden por tratarse de una tabla de dimensiones Ix2.***

-El coeficiente de contingencia de Pearson toma el valor 0,382, también con el nivel de significación 0,001. Esto es consistente con el valor de los coeficientes anteriores.

Ejemplo 4:

Analizar si existe asociación entre las variables Tipo e Internet del archivo ***Encinf.sav*** *e indicar si es posible establecer una relación de dependencia entre ambas.*

Para analizar si existe asociación e indicar el tipo de relación de dependencia se activan las opciones Chi-cuadrado, *Coeficiente de contingencia, Phi y V de Cramer y Lambda* del cuadro de diálogo *Tablas de contingencia: Estadísticos* y se seleccionan las nuevas variables, Tipo e Internet.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asint. (bilateral)
Chi-cuadrado de Pearson	40,523[a]	2	,000
Razón de verosimilitud	55,544	2	,000
Asociación lineal por lineal	39,676	1	,000
N de casos válidos	114		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 5,79.

- El estadístico Chi-cuadrado toma el valor 40,523 que permite rechazar la hipótesis de independencia para cualquier nivel de significación.

Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,596	,000
	V de Cramer	,596	,000
	Coeficiente de contingencia	,512	,000
N de casos válidos		114	

a. No asumiendo la hipótesis nula.

b. Empleando el error típico asintótico basado en la hipótesis nula.

- Los valores de los coeficientes Phi, V de Cramer y coeficiente de contingencia de Pearson son superiores a 0,5 y significativos para cualquier nivel. Teniendo en cuenta

que el coeficiente de contingencia para una tabla de dimensiones 3x2 tiene una cota superior inferior a 1, el valor 0,512 indica un grado de asociación moderadamente alto.

Medidas direccionales

			Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Nominal por nominal	Lambda	Simétrica	,245	,083	2,534	,011
		TIPO dependiente	,000	,000	,[c]	,[c]
		INTERNET dependiente	,418	,129	2,534	,011
	Tau de Goodman y Kruskal	TIPO dependiente	,157	,038		,000[d]
		INTERNET dependiente	,355	,025		,000[d]

a. No asumiendo la hipótesis nula.

b. Empleando el error típico asintótico basado en la hipótesis nula.

c. No se puede efectuar el cálculo porque el error típico asintótico es igual a cero.

d. Basado en la aproximación chi-cuadrado.

- Del análisis direccional se deduce que:

- Tomando la variable Tipo como dependiente, λ Tipo = 0,000, lo cual indica que el conocimiento de los valores de la variable Internet no permite predecir el comportamiento de la variable Tipo.

- Tomando la variable Internet como dependiente, λ Internet=0,418 con un error típico asintótico igual a 0,129. Así pues, el conocimiento de los valores de la variable Tipo permite reducir la incertidumbre en la predicción del comportamiento de la variable Internet en un 41,8%. Esta estimación de λ Internet es significativa para niveles de significación superiores a 0,011.

- El estadístico Tau de Goodman con Internet dependiente toma el valor τ Internet = 0,355 con un error típico de 0,025; este resultado confirma las conclusiones del apartado anterior.

Ejemplo 5:

*Del archivo **Encinf.sav** analice si existe concordancia entre las valoraciones dadas por los alumnos al equipamiento informático del centro, con referencia a las variables*

Dotación y Software, e indique si es posible establecer una relación de dependencia entre ambas.

Para analizar la concordancia entre variables ordinales se activan las opciones *Correlaciones, Gamma, d de Somers, Tau-b de Kendall y Tau-c de Kendall* del cuadro de diálogo *Tablas de contingencia: Estadísticos*.

Medidas direccionales

			Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Ordinal por ordinal	d de Somer	Simétrica	,333	,074	4,428	,000
		dotación dependiente	,333	,074	4,428	,000
		SOFTWARE dependiente	,333	,074	4,428	,000

a. No asumiendo la hipótesis nula.

b. Empleando el error típico asintótico basado en la hipótesis nula.

Medidas simétricas

		Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Ordinal por ordinal	Tau-b de Kendall	,333	,074	4,428	,000
	Tau-c de Kendall	,311	,070	4,428	,000
	Gamma	,391	,084	4,428	,000
	Correlación de Spearman	,424	,088	4,769	,000[c]
Intervalo por	R de Pearson	,524	,086	6,281	,000[c]
N de casos válidos		106			

a. No asumiendo la hipótesis nula.

b. Empleando el error típico asintótico basado en la hipótesis nula.

c. Basada en la aproximación normal.

- Las medidas simétricas de concordancia entre Dotación y Software indican que existe una asociación moderada y positiva. El coeficiente de correlación de Spearman, que es la medida de concordancia más utilizada con datos ordinales, es 0,424 con un error típico asintótico de 0,088; por tanto, difiere significativamente de 0 para cualquier nivel de significación.

- Las medidas direccionales indican que la concordancia es simétrica ya que coinciden todos los valores del coeficiente d de Somers. Esto quiere decir, que si bien las variables presentan un moderado grado de concordancia no existe entre ellas una relación de dependencia.

Ejemplo 6:

Analice la concordancia entre las variables Mantenimiento y Aulas del archivo ***Encinf.sav****, e indique si puede admitirse una relación de dependencia entre ambas.*

En el cuadro de diálogo *Tablas de contingencia* se seleccionan las variables Mantenimiento y Aulas en las *Filas* y *Columnas*. Con el botón *Estadísticos* se accede al menú donde se activan las medidas de concordancia entre variables ordinales. Los resultados de este análisis son:

Medidas direccionales

			Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Ordinal por ordinal	d de Somer	Simétrica	,018	,083	,212	,832
		Mantenimiento dependiente	,017	,081	,212	,832
		Acceso aulas dependiente	,018	,084	,212	,832

a. No asumiendo la hipótesis nula.

b. Empleando el error típico asintótico basado en la hipótesis nula.

Medidas simétricas

		Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Ordinal por ordinal	Tau-b de Kendall	,018	,083	,212	,832
	Tau-c de Kendall	,017	,079	,212	,832
	Gamma	,020	,096	,212	,832
	Correlación de Spearman	,028	,106	,282	,778[c]
Intervalo por	R de Pearson	,111	,115	1,127	,262[c]
N de casos válidos		104			

a. No asumiendo la hipótesis nula.

b. Empleando el error típico asintótico basado en la hipótesis nula.

c. Basada en la aproximación normal.

Tanto el coeficiente de correlación de Spearman como las restantes medidas simétricas toman valores próximos a 0 y en todos los casos se acepta la hipótesis de que no existe concordancia. A la vista de este resultado no puede suponerse que exista una relación de dependencia entre estas variables; los valores de las medidas direccionales lo confirman.

MEDIDAS DE ASOCIACIÓN PARA VARIABLES CUANTITATIVAS

Para variables cuantitativas, es decir, aquellas que se miden en una escala de intervalo o de razón, las medidas de asociación más utilizadas son la covarianza y el coeficiente de correlación de Pearson. Ambas medidas hacen referencia a un tipo particular de asociación: la asociación lineal.

El análisis conjunto de dos variables cuantitativas puede ir acompañado del análisis unidimensional de cada una de ellas por separado, así como de gráficos que pongan de manifiesto el patrón de comportamiento conjunto de ambas variables.

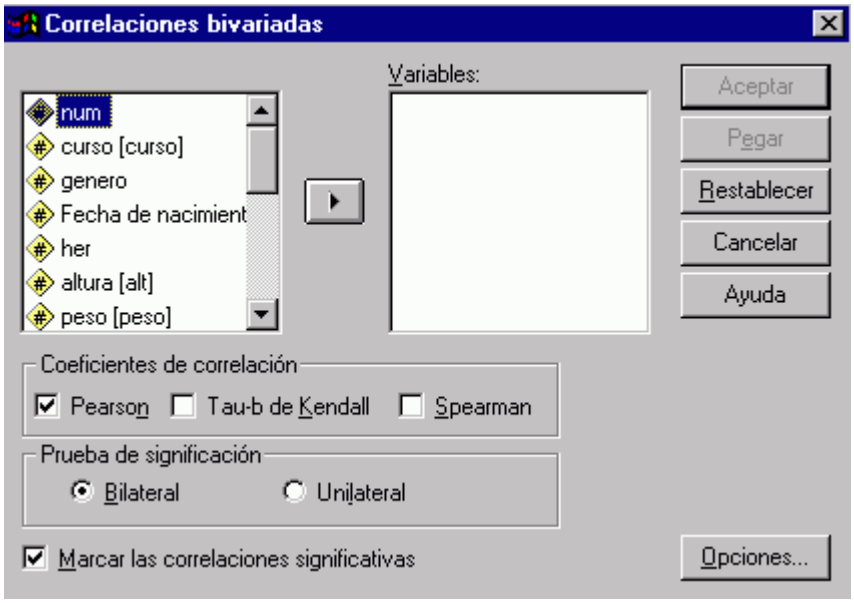
Para realizar el análisis conjunto de dos variables X_1 y X_2, la secuencia es:

Analizar

Correlaciones

Bivariadas

Aparece el siguiente cuadro de diálogo:



Por defecto, están activadas las opciones *Coeficiente de correlación de Pearson*, *Prueba de significación Bilateral y Marcar las correlaciones significativas*. Otras medidas de asociación son los coeficientes *Tau-b de Kendall y Spearman* para variables ordinales, a los que ya se ha hecho referencia en el epígrafe anterior.

COEFICIENTE DE CORRELACIÓN DE PEARSON

El coeficiente de correlación de Pearson es la principal medida de asociación lineal entre dos variables cuantitativas y se define como:

$$r = \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)}{\sqrt{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2}\sqrt{\sum_{i=1}^{n}(X_{2i} - \overline{X}_2)^2}}$$

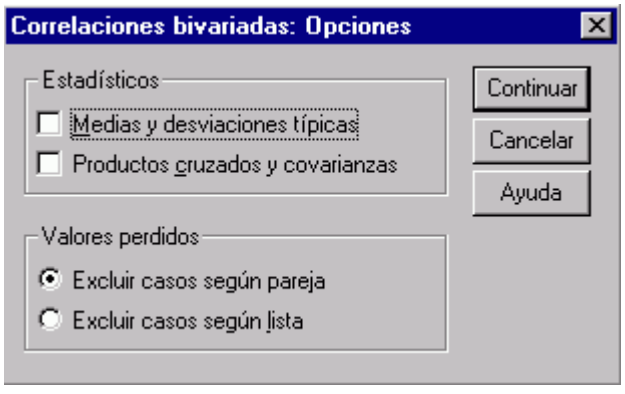
Este coeficiente, cuyo valor no depende de las unidades de medida de las variables, está acotado entre -1 y +1; su signo indica la dirección, positiva o negativa, de la asociación lineal y su valor absoluto la intensidad de la misma. En caso de asociación lineal perfecta toma el valor |1| ; si no hay asociación lineal toma el valor 0, lo cual no implica que no pueda haber asociación de otro tipo.

La prueba de significación del coeficiente de correlación de Pearson puede plantearse:

- *Bilateral* (a doble cola) si se contrasta la hipótesis nula de ausencia de asociación lineal sin especificar de antemano en la hipótesis alternativa la dirección o sentido de la asociación, en caso de que ésta exista.
- *Unilateral* (a una sola cola) si se contrasta la hipótesis nula especificando de antemano en la hipótesis alternativa la dirección de la asociación. Si se desea un contraste unilateral es necesario activar la opción correspondiente.

OPCIONES

Al activar el botón *Opciones* se abre el cuadro de diálogo siguiente.



Este cuadro permite incluir en los resultados los siguientes *Estadísticos*:

- *Medias y desviaciones típicas* de cada una de las variables analizadas.

- Productos cruzados y covarianzas. La suma de los productos cruzados es el numerador del coeficiente de correlación lineal que dividido por n-1 recibe el nombre de covarianza cuya expresión es:

$$S_{X_1X_2} = \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X_1})(X_{2i} - \overline{X_2})}{n-1}$$

La covarianza es una medida de asociación lineal cuyo signo indica la dirección o sentido de la asociación, pero cuyo valor numérico es de difícil interpretación porque depende de las unidades de medida de las variables.

El cuadro de diálogo *Opciones* permite modificar la forma en que se gestionan los valores missing. Por defecto, está activada la opción *Excluir casos según pareja* con la que se calculan los coeficientes de correlación utilizando todos los casos para los que existe información sobre las dos variables. Como alternativa puede activarse la opción *Excluir casos según lista* con la que se calculan los coeficientes de correlación utilizando únicamente los casos para los que se dispone de información sobre todas las variables. Si únicamente se han seleccionado dos variables en el cuadro de diálogo *Correlaciones bivariadas* ambas opciones proporcionan los mismos resultados.

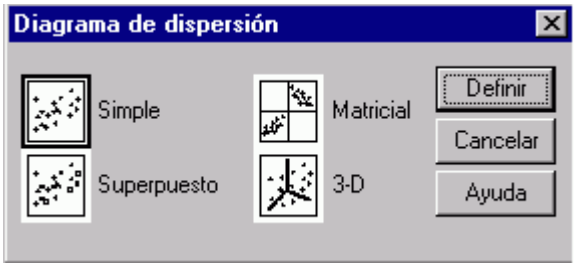
DIAGRAMA DE DISPERSIÓN

La representación gráfica más útil para describir el comportamiento conjunto de dos variables es el diagrama de dispersión o nube de puntos, donde cada caso aparece representado como un punto en el plano definido por las variables X_1 y X_2. Para obtener un diagrama de dispersión la secuencia es:

Gráficos

Dispersión

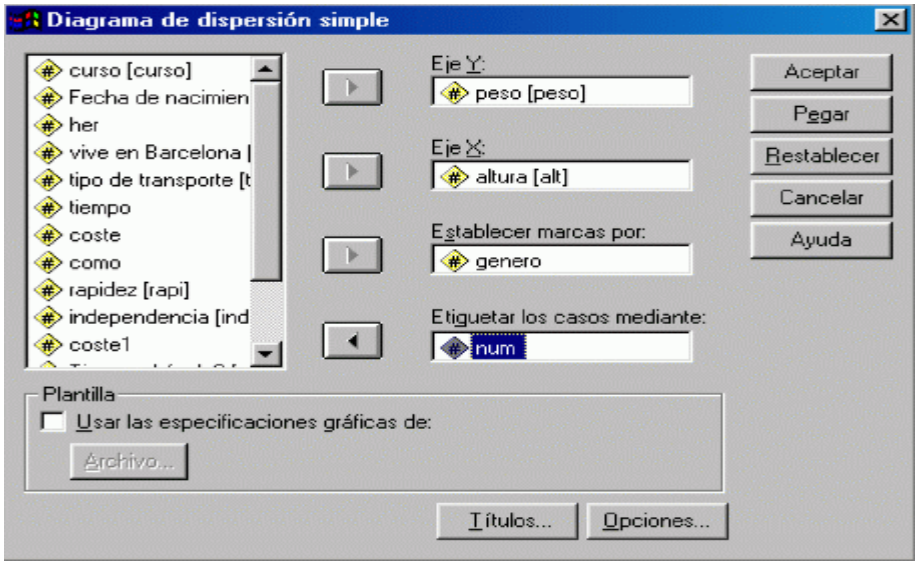
El cuadro de diálogo siguiente:



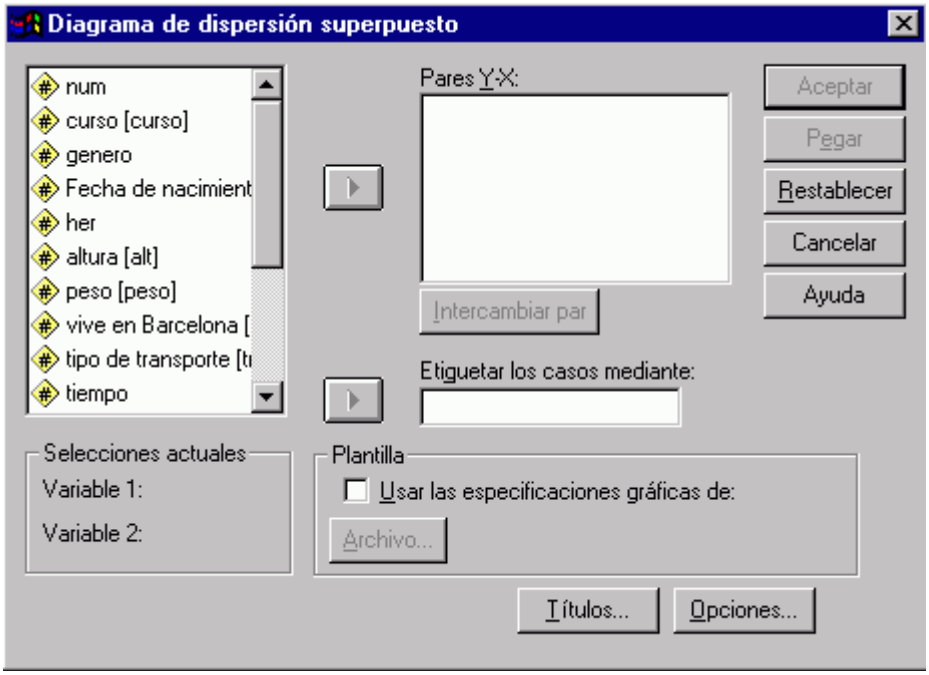
recoge diferentes tipos de diagramas de dispersión. Éstos pueden ser:

1. Simple: si el diagrama sólo recoge el comportamiento simultáneo de dos variables, una definida en el eje X (abscisas) y la otra en el eje Y (ordenadas). Con el botón *Definir* se abre el siguiente cuadro:

- En *Eje X* se selecciona la variable que se considera independiente y en *Eje Y* la dependiente.
- En *Establecer marcas por* puede indicarse alguna variable de control cuyas categorías o valores se representan con un símbolo o color distintivo. Esto permite identificar los puntos pertenecientes a cada categoría y poner de manifiesto si existen comportamientos diferenciados.
- En *Etiquetar los casos mediante* se puede indicar alguna variable cuyos valores se tomarán como etiquetas de los casos. Para visualizar las etiquetas es preciso activar la opción *Mostrar el gráfico* con las etiquetas de caso del cuadro de diálogo *Opciones*.
- El botón *Títulos* ofrece la posibilidad de definir dos líneas de título y un subtítulo, y dos líneas de nota al pie del gráfico.



2. *Superpuestos*: presenta dos o más parejas de variables en un mismo gráfico.



- En *Pares Y-X* se indican las parejas de variables a representar seleccionándolas de dos en dos en la lista de variables. Si se quiere intercambiar X por Y se utiliza el botón *Intercambiar par*.
- *Etiquetar los casos mediante* tiene la misma función que en el diagrama simple.
- Los botones *Títulos y Opciones* ofrecen las mismas posibilidades ya vistas para el diagrama de dispersión simple.

3. *Matricial:* ofrece una matriz de diagramas de dispersión simples de todos los pares y todas las ordenaciones posibles que se pueden formar con las variables seleccionadas. En el cuadro de diálogo que aparece con el botón *Definir* se deben seleccionar las variables cuyos diagramas de dispersión simples aparecerán en la matriz.

4. *3-D*: proporciona en tres dimensiones el diagrama de dispersión de tres variables.

Si el diagrama de dispersión es *Simple o Superpuesto* se puede visualizar con la recta que mejor se ajusta a la nube de puntos. Para ello se edita el gráfico en el visor de resultados haciendo doble clic sobre el mismo.

En la barra de menú del editor de gráficos se activa *Diseño* > *Opciones* y se abre el cuadro de diálogo:

Se selecciona *Ajustar línea* > *Total*. En *Opciones de ajuste* se puede elegir el método de ajuste deseado entre: *Regresión lineal* (activado por defecto), *Regresión cuadrática, Regresión cúbica y Minsce*. También es posible incluir en el diagrama de dispersión una línea paralela al eje de abscisas que pasa por la media de la variable Y con la opción *Línea de referencia* para la media en *Y* > *Total.*

Cuando el diagrama recoge un gran número de observaciones algunos puntos representan a más de un caso ya que estos se superponen. Con la opción *Girasoles* > *Mostrar girasoles* cada punto aparece con tantas rayas o 'pétalos' como casos

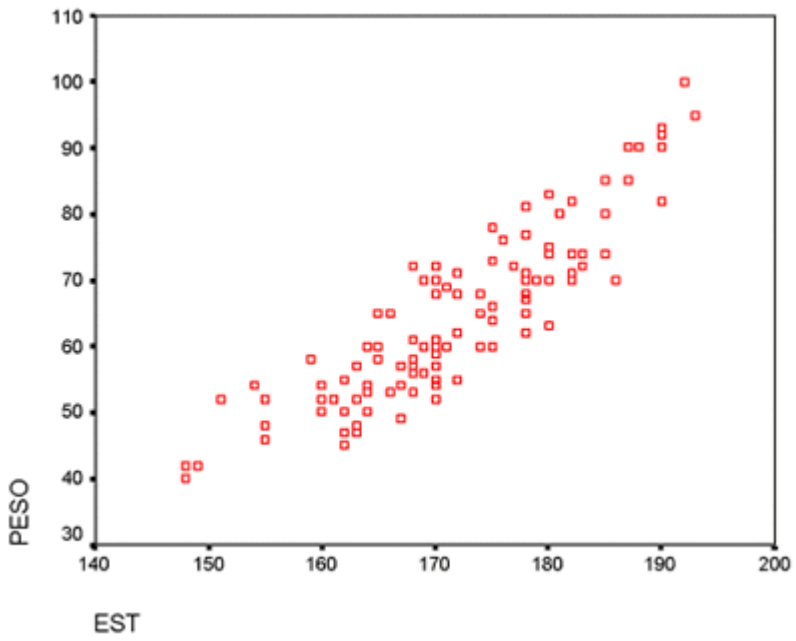
representa. Ésta es una forma gráfica de indicar cuantos casos están representados por un punto.

EJEMPLOS

Con referencia a las variables Peso y Est (estatura) del archivo ***Encinf.sav*** *comprobar gráfica y analíticamente la existencia de una relación lineal entre ellas.*

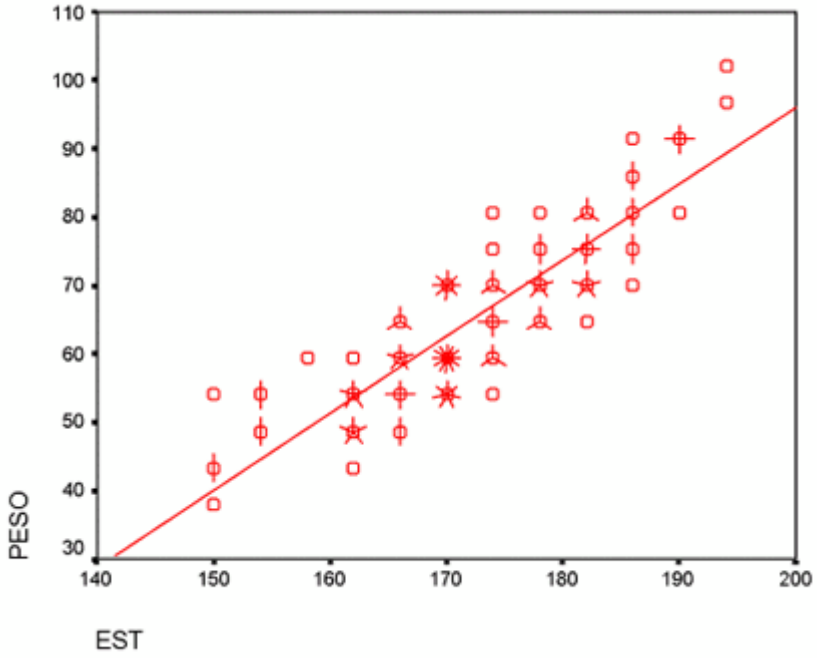
1. La representación gráfica que permite comprobar la existencia de relación lineal entre dos variables es el diagrama de dispersión y la medida analítica adecuada es el coeficiente de correlación lineal.

Con la secuencia *Gráficos* > *Dispersión* > *Simple* > *Definir* se abre el cuadro de diálogo *Diagrama de dispersion simple*. Se seleccionan en el *Eje Y* la variable Peso y en el *Eje X* la variable Est, y se obtiene el siguiente gráfico:



Como se observa en el gráfico ambas variables presentan una relación lineal positiva; es decir, a medida que aumenta el valor de la variable Est aumenta también el valor de la variable Peso.

Si se desea ver la n **No se encuentran elementos de tabla de ilustraciones.**ube de puntos con la línea de mejor ajuste superpuesta, y que los casos iguales o muy próximos entre si queden representados por un sólo punto, se edita el gráfico haciendo doble clic sobre el mismo. En el editor de gráficos se selecciona *Diseño* > *Opciones y en el cuadro Opciones del diagrama de dispersión* se activa *Mostrar girasoles y Ajustar línea* > *Total*.



Para cuantificar el grado de asociación lineal entre las variables la medida adecuada es el coeficiente de correlación lineal de Pearson. Con la secuencia *Analizar* > *Correlaciones* > *Bivariadas* se abre un cuadro de diálogo donde se seleccionan las variables Peso y Est, y con las opciones *Coeficiente de correlación* > *Pearson* se obtiene la siguiente matriz de correlaciones.

Correlaciones

		EST	PESO
EST	Correlación de Pearson		
	Sig. (bilateral)		
	N		
PESO	Correlación de Pearson	,883**	
	Sig. (bilateral)	,000	
	N	114	

**. La correlación es significativa al nivel 0,01

El valor de r=0,883 es positivo y elevado, así como significativo para cualquier nivel, con lo cual se confirma la impresión proporcionada por el gráfico acerca de la existencia de asociación lineal entre las variables.