Universidad Autónoma de Madrid

Instrumentos Estadísticos Avanzados
Facultad Ciencias Económicas y Empresariales
Departamento de Economía Aplicada
Profesor: Santiago de la Fuente Fernández

ANÁLISIS DE CONTINGENCIA

VARIABLES CUALITATIVAS

Las variables cualitativas son aquellas cuyos valores son un conjunto de cualidades no numéricas a las que se le suele llamar *categorías, modalidades* o *niveles* – ejemplos: sexo (mujer, hombre), filosofía política (liberal, moderada, conservadora), estado civil (soltero, casado, divorciado, viudo), nivel de estudios (ninguno, primario, medio, universitario), etc. -

Una propiedad deseable de las categorías es que sean *exhaustivas* (proporcionen suficientes valores para clasificar a toda la población) y *mutuamente excluyentes* (cada individuo se clasifica en una y solo una categoría).

A primera vista, la exhaustividad puede parecer muy restrictiva: puede que se desee saber que opinan los liberales y conservadores frente a la legalización del aborto. En este caso, la cuestión se resuelve redefiniendo la población mediante eliminación de los moderados.

CLASIFICACIÓN DE VARIABLES CUALITATIVAS

Hay varias formas de clasificar las variables cualitativas:

1. **Variables dicotómicas y politómicas** (*según el número de categorías*)

- **Dicotómicas:** Solo hay dos modalidades. Ejemplo, padecer una enfermedad (Sí, No), Sexo (Hombre, Mujer), Resultado de una oposición (Aprobar, Suspender), en general los fenómenos de respuesta binaria.

- **Politómicas:** Cuando hay mas de dos categorías. Ejemplo, fenómenos de respuesta múltiple, lugar de nacimiento, clase social, etc.

2. **Escalas nominal, ordinal y por intervalos** *(según la escala de medida de las categorías)*

- **Nominal:** No se puede definir un orden natural entre sus categorías, por ejemplo, la raza (blanca, negra, otra), la religión (católica, judía, protestante, otra), etc.

- **Ordinal:** Es posible establecer relaciones de orden entre las categorías lo conduce a establecer relaciones de tipo mayor, menor, igual o preferencia entre los individuos. Por ejemplo, el rango militar (soldado, sargento, teniente, otro), la clase social (alta, media, baja), etc.
  Sin embargo, no se pueden evaluar distancias absolutas entre categorías. Así, se puede decir que una persona de clase alta tiene mayor poder adquisitivo que una persona de clase media, pero no se puede decir exactamente cuál es la diferencia en poder adquisitivo entre ambas.

- **Por Intervalo:** Proceden de variables cuantitativas agrupadas en intervalos. Estas variables pueden ser tratadas como ordinales pero para ellas se pueden calcular, además, distancias numéricas entre dos niveles de la escala ordinal, ejemplos de este tipo son el sueldo, la edad, los días del mes o el nivel de presión sanguínea.

Existen variables que pueden ser medidas en escala nominal, ordinal o cuantitativa. Por ejemplo, el tipo de educación (privado, público) es nominal, el nivel de educación (primaria, secundaria, universitaria, postgraduado) es ordinal, y el número de años de educación (0, 1, 2, . . .) es cuantitativa.

Los métodos estadísticos propios para analizar variables nominales pueden ser usados para variables ordinales pero no al revés. Lo mejor es usar métodos apropiados para cada tipo de escala.

TABLAS DE CONTINGENCIA: VARIABLES NOMINALES

Variable nominal es aquella que conlleva información sobre un conjunto de valores no ordenado.

La tabla de contingencia recoge n_{ij} incidencias entre dos variables nominales (x_i, y_j)

Y / **X**	y_1	y_2	.....	y_j	.....	y_m	
x_1	n_{11} (e_{11})	n_{12} (e_{12})	.....	n_{1j} (e_{1j})	.....	n_{1m} (e_{1m})	$N_{1\bullet}$
x_2	n_{21} (e_{21})	n_{22} (e_{22})	.....	n_{2j} (e_{2j})	.....	n_{2m} (e_{2m})	$N_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		.....	$\vdots$	$\vdots$
x_i	n_{i1} (e_{i1})	n_{i2} (e_{i2})	.....	n_{ij} (e_{ij})	.....	n_{im} (e_{im})	$N_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	.....		.....	$\vdots$	$\vdots$
x_k	n_{k1} (e_{k1})	n_{k2} (e_{k2})	.....	n_{kj} (e_{kj})	.....	n_{km} (e_{km})	$N_{k\bullet}$
	$N_{\bullet 1}$	$N_{\bullet 2}$		$N_{\bullet j}$		$N_{\bullet m}$	N

$$N_{i\bullet} = \sum_{j=1}^{m} n_{ij}$$

$$N_{\bullet j} = \sum_{i=1}^{k} n_{ij}$$

$$N = \sum_i N_{i\bullet} = \sum_j N_{\bullet j}$$

- Se analizan dos variables (que admiten distintas modalidades) mediante una tabla de contingencia, en donde una ocupa las filas y otra las columnas.
  La intersección entre una fila y una columna da lugar a una celda o casilla, cuya frecuencia observada es n_{ij}

- Se contrasta la hipótesis nula que presupone la independencia entre ambas variables, mediante el estadístico χ^2 de Pearson.

$$\begin{cases} H_0: \text{Ambas variables son independientes} \\ H_1: \text{Existe una relación de dependencia} \end{cases}$$

Se define el estadístico observado:

$$\sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \chi^2_{(k-1).(m-1)}$$

que sigue asintóticamente una distribución χ^2 con $(k-1).(m-1)$ grados de libertad si es cierta la hipótesis nula H_0, con $e_{ij} > 5$, $1 \le i \le k$, $1 \le j \le m$ (en caso contrario, es necesario agrupar filas o columnas contiguas).

La región crítica para el contraste de independencia se determina: $P\left[\chi^2_{(k-1).(m-1)} \ge k / H_0\right] = \alpha$

Así, pues, para un nivel de significación α:

$$\begin{cases} \chi^2_{(k-1).(m-1)} < \chi^2_{\alpha;(k-1).(m-1)} & \Rightarrow \text{ Se acepta } H_0 \text{ (no existe diferencia significativa al nivel } \alpha) \\ \chi^2_{(k-1).(m-1)} \ge \chi^2_{\alpha;(k-1).(m-1)} & \Rightarrow \text{ Se rechaza } H_0 \text{ (existe diferencia significativa al nivel } \alpha) \end{cases}$$

Muy frecuentemente, se utiliza para ver si existe o no relación entre los caracteres (X, Y), es decir, si son o no independientes. Entonces recibe el nombre de *contraste de independencia de caracteres*:

$$\begin{cases} \chi^2_{(k-1).(m-1)} < \chi^2_{\alpha;(k-1).(m-1)} & \Rightarrow \quad \text{X e Y son independientes al nivel } \alpha \\ \chi^2_{(k-1).(m-1)} \geq \chi^2_{\alpha;(k-1).(m-1)} & \Rightarrow \quad \text{X e Y no son independientes al nivel } \alpha \end{cases}$$

- Es muy útil la igualdad: $\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{n_{ij}^2}{e_{ij}} - N$

- **SPSS** $\begin{cases} \text{Sig.asintótica}(p_\text{value}) \leq 0{,}05 & \Rightarrow \quad \text{Se rechaza } H_0 \\ \text{Sig.asintótica}(p_\text{value}) > 0{,}05 & \Rightarrow \quad \text{Se acepta } H_0 \end{cases}$

- En tablas 2x2 las decisiones concernientes al uso de la prueba χ^2 debe guiarse por las recomendaciones de Cochran:

Y / **X**	y_1	y_2	
x_1	n_{11}	n_{12}	$n_{11}+n_{12}$
x_2	n_{21}	n_{22}	$n_{21}+n_{22}$
	$n_{11}+n_{21}$	$n_{12}+n_{22}$	N

☞ Cuando $N > 40$ se debe utilizar corregida por la continuidad:

$$\chi_1^2 = \frac{N\left(\left|n_{11}.n_{22} - n_{12}.n_{21}\right| - \frac{N}{2}\right)^2}{(n_{11}+n_{12})(n_{21}+n_{22})(n_{11}+n_{21})(n_{12}+n_{22})}$$

☞ Cuando $20 \leq N \leq 40$, se debe utilizar siempre que las frecuencias esperadas $(e_{ij} > 5)$ sean superiores a 5. Si fuese más pequeña se utilizaría la *Prueba exacta de Fisher*.

☞ Cuando $N < 20$ se utiliza siempre la *Prueba exacta de Fisher*.

PRUEBA EXACTA DE FISHER.- Es una técnica válida tanto para datos nominales u ordinales, siempre que la muestra sea pequeña.

La prueba determina si los grupos difieren en la proporción correspondiente a las clasificaciones.

Se caracteriza porque no utiliza una aproximación de probabilidad sino la distribución de probabilidad exacta de la configuración de las frecuencias observadas.

Como para totales marginales fijos, la distribución de probabilidad de las frecuencias observadas sigue una ley hipergeométrica, en el caso de que las dos variables observadas sean independientes la probabilidad **p** de obtener cualquier disposición de las n_{ij} viene dada por:

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

	Test -	Test +	Marginal
Hombres	10	0	10
Mujeres	4	5	9
Marginal	14	5	19

$$p = \frac{(10+0)!\ (4+5)!\ (10+4)!\ (0+5)!}{10!\ 0!\ 4!\ 5!\ 19!} = 0{,}0108$$

La probabilidad de la distribución de frecuencias es $p = 0{,}0108$.

Ahora bien, en este ejercicio ha sido fácil de calcular porque en una de las celdillas existe una frecuencia cero.

En otro ejemplo, donde no existe un cero en ninguna celdilla:

(b)	Test -	Test +	Marginal
Hombres	1	6	7
Mujeres	4	1	5
Marginal	5	7	12

$$p_b = \frac{(1+6)!\ (4+1)!\ (1+4)!\ (6+1)!}{1!\ 6!\ 4!\ 1!\ 12!} = 0{,}04399$$

Sin alterar los totales marginales, una posibilidad más extrema sería la que aparece en la tabla:

(c)	Test -	Test +	Marginal
Hombres	0	7	7
Mujeres	5	0	5
Marginal	5	7	12

$$p_c = \frac{7!\ 5!\ 5!\ 7!}{0!\ 7!\ 5!\ 0!\ 12!} = 0{,}00126$$

La posibilidad de ocurrencia de la tabla es:

$$p = p_b + p_c = 0{,}04399 + 0{,}00126 = 0{,}04525 < 0{,}05 = \alpha$$

$p = 0{,}04525$ es el valor que se utiliza para saber si los datos de la tabla permiten rechazar la hipótesis nula H_0, como $p = 0{,}04525 < 0{,}05 = \alpha$, se rechaza la hipótesis nula concluyendo que los test afectan en mayor medida a los hombres.

Se observa que si el valor más pequeño de la tabla de contingencia es muy grande, la prueba de Fisher puede ser complicada de calcular (sí el valor más pequeño fuera 2 habría que determinar tres probabilidades exactas y sumarlas, sí fuera 3 habría que determinar cuatro probabilidades exactas y sumarlas, y así sucesivamente).

MODIFICACIONES DE TOCHER: Con una pequeña modificación de la prueba de Fisher, Tocher probó (1950) que se consigue una prueba más poderosa para datos de una tabla 2x2.

Para ilustrar la modificación, se parte de la tabla:

(a)	Test -	Test +	Marginal
Hombres	2	5	7
Mujeres	3	2	5
Marginal	5	7	12

$$p_a = \frac{(2+5)!\ (3+2)!\ (2+3)!\ (5+2)!}{2!\ 5!\ 3!\ 2!\ 12!} = 0{,}26515$$

La probabilidad asociada con la ocurrencia de valores tan extremos como las puntuaciones observadas (a) conforme a la hipótesis nula será:

$$p = p_a + p_b + p_c = 0{,}26515 + 0{,}04399 + 0{,}00126 = 0{,}31040 > 0{,}05 = \alpha$$

$p = 0{,}31040$ es *la probabilidad encontrada con la prueba Fisher,* que siendo mayor que el nivel de significación $\alpha = 0{,}05$, conduce a rechazar la hipótesis nula.

La *probabilidad de Tocher* determina antes todos los casos extremos (b) y (c) sin incluir aún el observado (a), con lo cual: $p_b + p_c = 0{,}04399 + 0{,}00126 = 0{,}04525$

Y recomienda el cálculo de la proporción: $p_T = \dfrac{\alpha - p(\text{casos más raros extremos})}{p(\text{casos observados aisladamente})} = \dfrac{\alpha - (P_b + P_c)}{P_a}$

$$p_T = \frac{0{,}05 - 0{,}04525}{0{,}26515} = 0{,}0179$$

MEDIDAS DE ASOCIACIÓN

En caso de rechazar la independencia entre los dos factores de una tabla de contingencia, se plantea la necesidad de definir índices que describan no solo la intensidad de la asociación, sino también su dirección. El estudio de estos índices, que se conocen con el nombre genérico de *medidas de asociación*.

Para detectar las *fuentes de asociación* existen diferentes métodos, unos directos, y otros de conversión en tablas 2x2. Entre los directos, el *análisis de residuos*, y entre los segundos, la *partición de la tabla original* en tablas 2x2.

ANÁLISIS DE LOS RESIDUOS

Los residuos son las diferencias entre la frecuencia observada y la frecuencia esperada en cada casilla: $r_{ij} = n_{ij} - e_{ij}$. En el caso de que el contraste de χ^2 haya resultado significativo, estos residuos indicarán qué casillas contribuyen en mayor grado al valor del estadístico.

Cuanto mayor sea el valor de los residuos mayor es la probabilidad de que una determinada combinación de valores de las variables, esto es, una casilla, sea significativa.

Para que el análisis de los residuos resulte adecuado es necesario que previamente éstos hayan sido ajustados y estandarizados, para lo cual se suele aplicar la fórmula propuesta por Haberman (1978), que consiste en dividir el valor del residuo en cada casilla por su error típico.

Residuos tipificados $r_{ij} = \dfrac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}}$

Residuos tipificados corregidos $d_{ij} = \dfrac{r_{ij}}{\sqrt{V(r_{ij})}} = \dfrac{(n_{ij} - e_{ij}) / \sqrt{e_{ij}}}{\sqrt{\left(1 - \dfrac{N_{i\bullet}}{N}\right)\left(1 - \dfrac{N_{\bullet j}}{N}\right)}} \approx N(0,1)$

Ejemplo.- Sea la tabla adjunta, $e_{ij} = \frac{N_{i\bullet} \cdot N_{\bullet j}}{N}$, (*) valor SPSS

Opinión Sistema Público / Nivel renta	Bueno (1)	Regular (2)	Malo (3)	Total
Bajo (1)	75 ($e_{11} = 51$)	35 ($e_{12} = 48$)	40 ($e_{13} = 51$)	150 (150)
Medio (2)	60 ($e_{21} = 61{,}2$)	70 ($e_{22} = 57{,}6$)	50 ($e_{23} = 61{,}2$)	180 (180)
Alto (3)	20 ($e_{31} = 30{,}6$)	30 ($e_{32} = 28{,}8$)	40 ($e_{33} = 30{,}6$)	90 (90)
Muy Alto (4)	15 ($e_{41} = 27{,}2$)	25 ($e_{42} = 25{,}6$)	40 ($e_{43} = 27{,}2$)	80 (80)
Total	170 (170)	160 (160)	170 (170)	500

$$e_{11} = \frac{150.170}{500} = 51 \qquad e_{21} = \frac{180.170}{500} = 61{,}2 \qquad e_{31} = \frac{90.170}{500} = 30{,}6 \qquad e_{41} = \frac{80.170}{500} = 27{,}2$$

$r_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}}$ *residuos tipificados*		
3,3607	- 1,5403	- 1,8764
- 0,1534	- 1,4316	1,6338
- 1,9162	1,6930	0,2236
- 2,3392	2,4542	- 0,1186

$V(r_{ij}) = \left(1 - \frac{N_{i\bullet}}{N}\right)\left(1 - \frac{N_{\bullet j}}{N}\right)$		
0,4620	0,4620	0,7176
0,4224	0,4224	0,4352
0,5412	0,5412	0,5576
0,5544	0,5544	0,5712

$$r_{11} = \frac{75-51}{\sqrt{51}} = 3{,}3607 \qquad r_{21} = \frac{60-61{,}2}{\sqrt{61{,}2}} = -0{,}1534 \qquad r_{31} = \frac{20-30{,}6}{\sqrt{30{,}6}} = -1{,}9162 \qquad e_{41} = \frac{15-27{,}2}{\sqrt{27{,}2}} = -2{,}3392$$

$$V(r_{11}) = \left(1 - \frac{N_{1\bullet}}{N}\right)\left(1 - \frac{N_{\bullet 1}}{N}\right) = \left(1 - \frac{150}{500}\right)\left(1 - \frac{170}{500}\right) = 0{,}4620 \qquad V(r_{21}) = \left(1 - \frac{N_{2\bullet}}{N}\right)\left(1 - \frac{N_{\bullet 1}}{N}\right) = \left(1 - \frac{180}{500}\right)\left(1 - \frac{170}{500}\right) = 0{,}4224$$

$d_{ij} = \frac{r_{ij}}{\sqrt{V(r_{ij})}}$ *residuos tipificados corregidos*		
4,9444	- 2,2661	- 2,7198
- 0,2360	- 2,2028	2,4766
- 2,6046	2,3098	0,2995
- 3,1416	3,2960	- 0,1569

Comparando los valores absolutos de los residuos tipificados corregidos con el correspondiente valor tabular de la normal, para un nivel de significación del 5% (> 1,96), se observa que muchos residuos son significativos.

Analizando estos valores, tanto en sus magnitudes como en sus rangos, resulta el patrón: *<< Hay un mayor número, considerablemente alto y superior al de otras clases sociales, de encuestados que pertenecen a una clase baja (valor 4,9444) y poseen una opinión favorable sobre la opinión pública. Por el contrario, la opinión de las clases altas y muy altas tienen una percepción claramente negativa. De este modo, se evidencia que existe una relación y del tipo que es ésta>>.*

Subrayar que este método supone un análisis celda a celda. Esta diferencia se encuentra que mientras el contraste usual trabaja con $(k-1)(m-1)$ elementos independientes, el contraste por cada celda implica que la totalidad de los residuos tipificados d_{ij} son independientes y cada uno de ellos se ajusta a una distribución teórica $N(0,1)$.

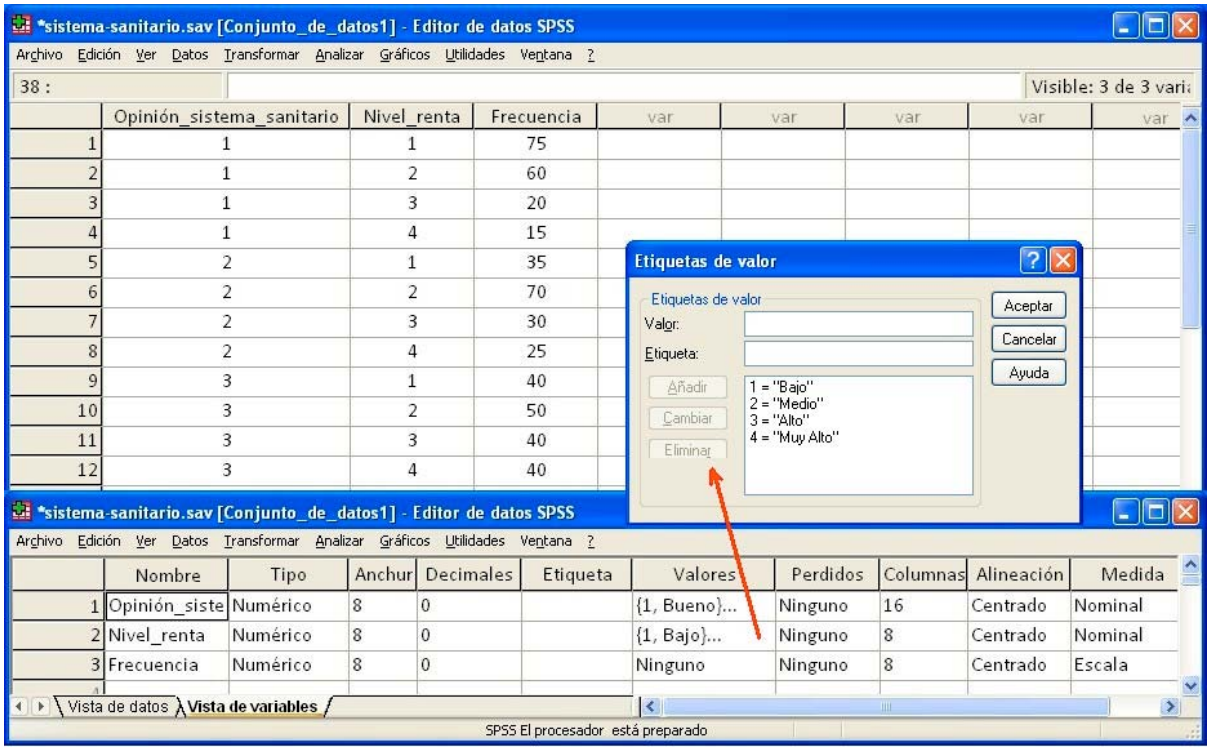
Análisis de los residuos

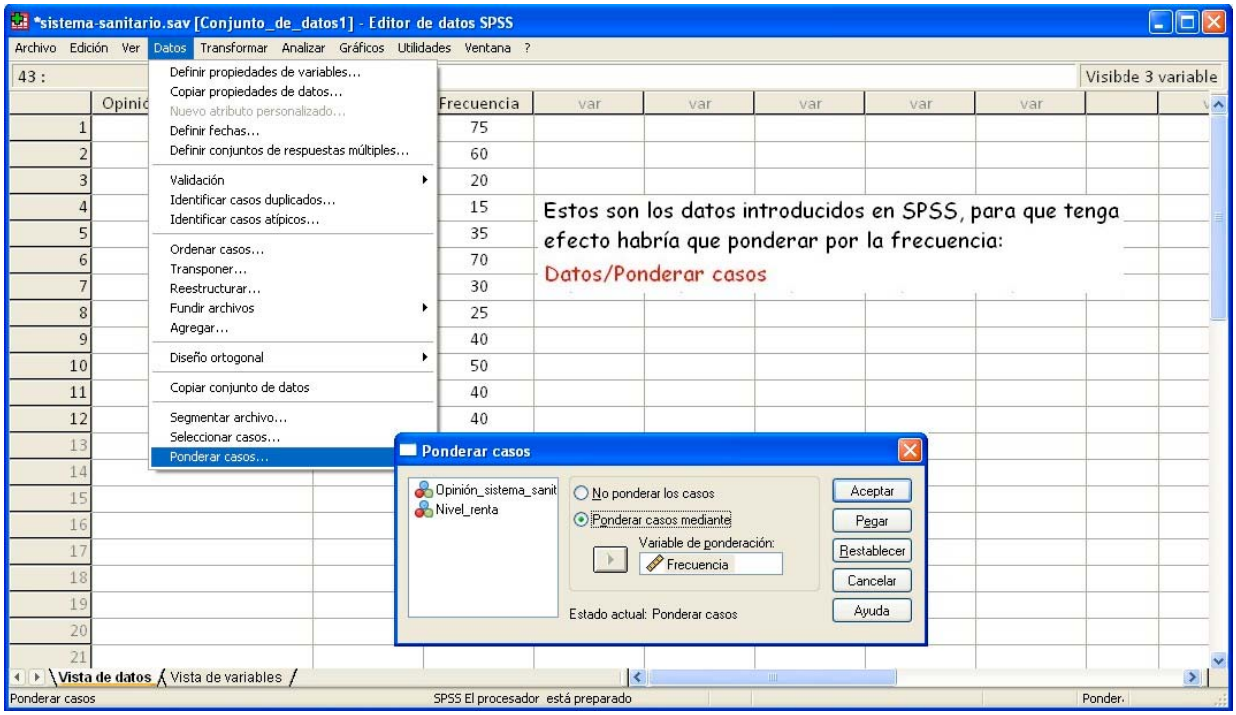
Hay diferencias significativas:

1. Prescindiendo del signo, los valores > 1,96
2. Con los valores >1,96, se analiza el signo
   - ☞ **Signo negativo:** frecuencia inferior a la teórica, se infiere una relación negativa entre los niveles de las variables.
   - ☞ **Signo positivo:** relación positiva.

En SPSS: En *Editor de datos* se introduce la tabla de valores. En *Vista de variables* se observa como en la variable (*Opinión_sistema_sanitario*) se han introducido los valores (1="Bueno", 2="Regular", 3="Malo"). Análogamente, en la variable (*Nivel_renta*) se han introducido los valores (1="Bajo", 2="Medio", 3="Alto", 4="Muy Alto"), ambas variables nominales; mientras que la variable (*Frecuencia*) la medida es escala.

En el menú [Analizar/Estadísticos descriptivos/Tablas de contingencia] se introduce en Filas (*Nivel_renta*) y en Columnas (*Opinión_sistema_sanitario*).





En el botón *[Casillas]* se seleccionan Frecuencias y Residuos.

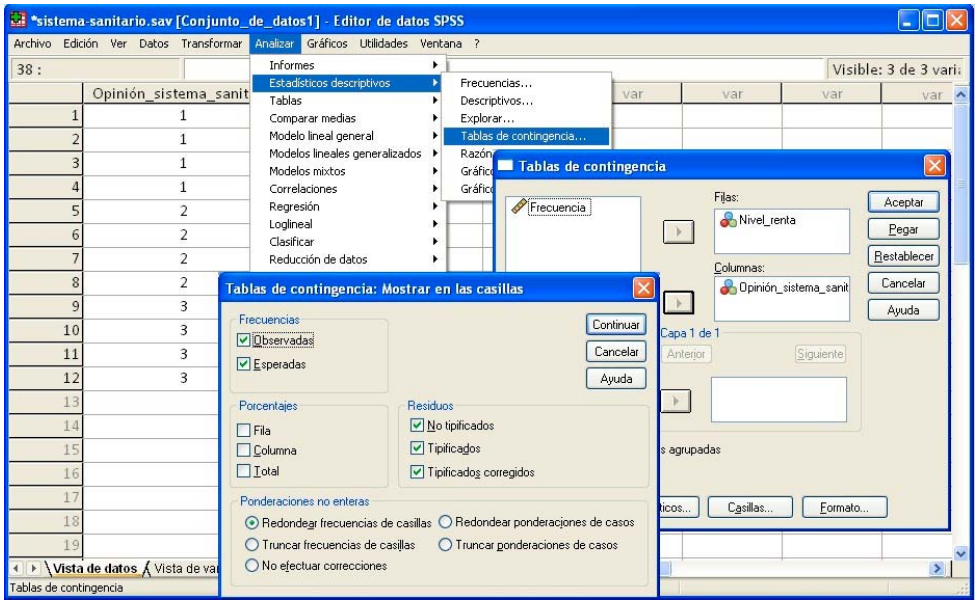


Tabla de contingencia Nivel_renta * Opinión_sistema_sanitario

			Opinión_sistema_sanitario			
			Bueno	Regular	Malo	Total
Nivel_renta	Bajo	Recuento	75	35	40	150
		Frecuencia esperada	51,0	48,0	51,0	150,0
		Residuo	24,0	-13,0	-11,0	
		Residuos tipificados	3,4	-1,9	-1,5	
		Residuos corregidos	4,9	-2,7	-2,3	
	Medio	Recuento	60	70	50	180
		Frecuencia esperada	61,2	57,6	61,2	180,0
		Residuo	-1,2	12,4	-11,2	
		Residuos tipificados	-,2	1,6	-1,4	
		Residuos corregidos	-,2	2,5	-2,2	
	Alto	Recuento	20	30	40	90
		Frecuencia esperada	30,6	28,8	30,6	90,0
		Residuo	-10,6	1,2	9,4	
		Residuos tipificados	-1,9	,2	1,7	
		Residuos corregidos	-2,6	,3	2,3	
	Muy Alto	Recuento	15	25	40	80
		Frecuencia esperada	27,2	25,6	27,2	80,0
		Residuo	-12,2	-,6	12,8	
		Residuos tipificados	-2,3	-,1	2,5	
		Residuos corregidos	-3,1	-,2	3,3	
Total		Recuento	170	160	170	500
		Frecuencia esperada	170,0	160,0	170,0	500,0

Adviértase que con la tabla de las frecuencias observadas (n_{ij}) y esperadas (e_{ij}) se calcula el estadístico χ^2 de Pearson: $e_{ij} = \dfrac{N_{i\bullet} \cdot N_{\bullet j}}{N}$

Opinión Sistema Público / Nivel renta	Bueno (1)	Regular (2)	Malo (3)	Total
Bajo (1)	75 ($e_{11} = 51$)	35 ($e_{12} = 48$)	40 ($e_{13} = 51$)	150 (150)
Medio (2)	60 ($e_{21} = 61{,}2$)	70 ($e_{22} = 57{,}6$)	50 ($e_{23} = 61{,}2$)	180 (180)
Alto (3)	20 ($e_{31} = 30{,}6$)	30 ($e_{32} = 28{,}8$)	40 ($e_{33} = 30{,}6$)	90 (90)
Muy Alto (4)	15 ($e_{41} = 27{,}2$)	25 ($e_{42} = 25{,}6$)	40 ($e_{43} = 27{,}2$)	80 (80)
Total	170 (170)	160 (160)	170 (170)	500

$$\chi^2_{(4-1)(3-1)} = \chi^2_6 = \sum_{i=1}^{4}\sum_{j=1}^{3} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \overbrace{\sum_{i=1}^{4}\sum_{j=1}^{3} \frac{n_{ij}^2}{e_{ij}} - N}^{\text{método práctico}} = 540{,}0492 - 500 = 40{,}0492$$

Pulsando el botón *[Estadísticos]* se selecciona la opción *Chi-cuadrado*.

En el Visor de resultados de SPSS:

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	40,049[a]	6	,000
Razón de verosimilitudes	39,693	6	,000
Asociación lineal por lineal	29,964	1	,000
N de casos válidos	500		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 25,60.

- **Chi-cuadrado**

El *estadístico de contraste* (observado) es 40,049, el cual, en la distribución χ^2 tiene 6 grados de libertad (*gl = 6*), tiene asociada una probabilidad (*Significación asintótica*) de 0.

Puesto que esta probabilidad (denominada *nivel crítico* o *nivel de significación observada*) es pequeña (menor que 0,05) , se decide rechazar la hipótesis nula, concluyendo que existe una relación de dependencia entre el nivel de renta y la opinión sobre la aceptación del sistema público.

Señalar que la *razón de verosimilitudes* (RV) es 39,693, tiene asociada una probabilidad (*Sig. asintótica*) de 0, que como es menor que 0,05, conduce a rechazar la hipótesis nula, concluyendo que existe dependencia entre las variables analizadas.

Los estadísticos (χ^2, RV) llevan a la misma conclusión, en caso contrario, se elige el estadístico con *menor Sig. asintótica*.

- **<u>Razón de verosimilitud Chi-cuadrado</u>** (Fisher, 1924; Neyman y Pearson, 1928)**:** Se obtiene mediante la relación: $RV = 2\sum_i \sum_j n_{ij} \log\left(\frac{n_{ij}}{e_{ij}}\right)$

Se trata de un estadístico asintóticamente equivalente a χ^2 (se distribuye y se interpreta igual) y es muy utilizado para estudiar la relación entre variables categóricas, particularmente en el contexto de los modelos *log-lineales*.

$$\begin{cases} RV < \chi^2_{\alpha;(k-1).(m-1)} & \Rightarrow \quad \text{X e Y son independientes al nivel } \alpha \\ RV \geq \chi^2_{\alpha;(k-1).(m-1)} & \Rightarrow \quad \text{X e Y no son independientes al nivel } \alpha \end{cases}$$

Se acepta la hipótesis nula cuando la significación de RV (Sig. asintótica) es mayor que 0,05.

La aplicación de los dos estadísticos (χ^2, RV) puede conducir a la misma conclusión. En los casos en que no se produzca esta coincidencia, se elige el estadístico con una significación (*probabilidad asociada*) menor.

$$RV = 2\sum_{i=}^{4} \sum_{j=1}^{3} n_{ij} \log\left(\frac{n_{ij}}{e_{ij}}\right) = 39{,}693$$

- **Corrección por continuidad de Yates** (1934)**:** Consiste en restar (0,5) puntos a $|n_{ij} - e_{ij}|$ del estadístico χ^2 (antes de elevarlo al cuadrado).

$$\chi^2_c = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(|n - e_{ij}| - 0{,}5\right)^2}{e_{ij}}$$

Algunos autores sugieren, que con muestras pequeñas, esta corrección permite que el estadístico χ^2 se ajuste mejor a las probabilidades de la distribución χ^2, pero no existe un consenso generalizado sobre la utilización de esta corrección

- **Correlaciones:** Permite obtener los coeficientes de correlación de Pearson y de Spearman.

Medidas simétricas

		Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Intervalo por intervalo	R de Pearson	,245	,043	5,640	,000[c]
Ordinal por ordinal	Correlación de Spearman	,245	,044	5,647	,000[c]
N de casos válidos		500			

a. Asumiendo la hipótesis alternativa.

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

c. Basada en la aproximación normal.

El coeficiente de correlación de Pearson es una medida de *asociación lineal* especialmente apropiada para estudiar la relación entre *variables de intervalo* o *razón*.

El coeficiente de correlación de Spearman también es una medida de *asociación lineal*, pero para *variables ordinales*.

Ambos coeficientes poseen escasa utilidad para estudiar las pautas de relación presentes en una tabla de contingencia típica, dado que lo habitual es utilizar las tablas de contingencia para cruzar variables de tipo nominal, o a lo sumo, de tipo ordinal con solo unos pocos niveles.

COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON

Suposición: Las variables ordinales (los valores discretos se pueden ordenar) o continuas (valores continuos se pueden ordenar) están distribuidas de forma *gaussiana*. *Es un test paramétrico*.

Método: Mide la desviación de las variables respecto a una *línea recta*.

Dados dos puntos $(x_i, y_i)_{i=1,2,\dots,N}$ se define el coeficiente de correlación:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \quad \text{tal que} \quad -1 \le r \le 1 \begin{cases} r = -1 & \text{no correlación} \\ r = 1 & \text{correlación perfecta} \end{cases}$$

La significancia de que no existe una correlación viene dada por la distribución t-Student con (N –2) grados de libertad, donde r está relacionado con la matriz de covarianza, que ofrece también un test paramétrico si se utiliza para buscar correlaciones

$$t_{\alpha,(N-2)} = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

COEFICIENTE DE CORRELACIÓN DE SPEARMAN (1904)**:**

Versión no paramétrica del coeficiente de correlación de Pearson, que se basa en los rangos de los datos. Resulta apropiado para datos ordinales, o los de intervalo que no satisfagan el supuesto de normalidad.

El signo del coeficiente indica la dirección de la relación y el valor absoluto del coeficiente de correlación indica la fuerza de la relación entre las variables.

$$r_S = 1 - \frac{6\sum_i d_i^2}{N(N^2-1)} \qquad -1 \le r_s \le 1$$

donde d_i es la diferencia entre el orden obtenido en el caso i-ésimo en ambas series.

MEDIDAS DE ASOCIACIÓN DE TIPO NOMINAL

Después de analizar si existe relación o no entre las variables objeto de estudio, cabe preguntarse ¿cuál es la intensidad de esa relación?. Entre las medidas utilizadas se encuentran: Coeficiente de contingencia, Q de Yule, Phi y V de Cramer, Lambda y el Coeficiente de incertidumbre.

Método: Probar que es errónea la suposición que las variables no están asociadas. Si es así, el número de incidencias esperado en el casillero (i,j) será $e_{ij} = \frac{N_{i\bullet} \cdot N_{\bullet j}}{N}$

Se define la función $\sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \chi^2_{(k-1).(m-1)}$

La significancia de que ambas distribuciones estén asociadas viene dada por la función de probabilidad χ^2 con ν grados de libertad.

$$Q(\chi^2/\nu) = \frac{1}{\Gamma(\nu)} \int_{\chi^2}^{\infty} e^{-t}\, t^{\nu-1}\, dt \quad \text{donde } \nu = (k-1)(m-1)$$

📖 **COEFICIENTE DE CONTINGENCIA.-** El coeficiente de contingencia C es una medida del grado de asociación o relación entre dos conjuntos de atributos. Es especialmente útil cuando hay una información clasificatoria (escala nominal) acerca de uno o ambos conjuntos de atributos.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \qquad 0 \le C \le 1 \quad \begin{cases} C = 0 & \mapsto \text{ Independencia} \\ C = 1 & \mapsto \text{ Asociación perfecta} \end{cases}$$

Sólo se utiliza cuando las tablas de contingencia tienen la misma dimensión.

En una tabla de contingencia (k x k) el valor máximo será: $C = \sqrt{\frac{k-1}{k}}$.

📖 **Q de YULE:** El coeficiente de contingencia Q es una medida del grado de asociación o relación entre dos conjuntos de atributos.

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} \qquad 0 \le Q \le 1$$

X \ Y	y_1	y_2	
x_1	n_{11}	n_{12}	$n_{11}+n_{12}$
x_2	n_{21}	n_{22}	$n_{21}+n_{22}$
	$n_{11}+n_{21}$	$n_{12}+n_{22}$	N

PHI: El coeficiente *phi* se obtiene: $\phi = \sqrt{\frac{\chi^2}{N}}$.

En las tablas de contingencia (2 x 2), *phi* adopta valores entre 0 y 1, y su valor es idéntico al del coeficiente de correlación de Pearson.

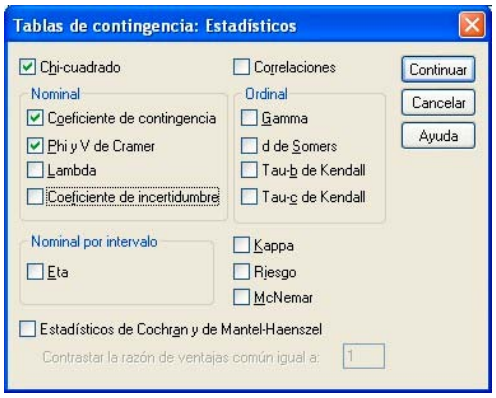
En las tablas en las que una de las variables tiene más de dos niveles, *phi* puede tomar valores mayores que 1 (pues el valor de χ^2 puede ser mayor que el tamaño muestral).

✎ El *coeficiente de contingencia* y *Phi y V de Cramer*, son medidas basadas en Chi-cuadrado, y que intentan corregir el valor del estadístico χ^2 para hacerle tomar un valor entre 0 y 1, y para minimizar el efecto del tamaño de la muestra sobre la cuantificación del grado de asociación (Pearson, 1913; Cramer, 1946).

📖 **V de CRAMER:** El coeficiente V incluye una ligera modificación de *phi*:

$$V_{Cramer} = \sqrt{\frac{\chi^2}{N.\min(k-1, m-1)}} \qquad 0 \le V_{Cramer} \le 1 \begin{cases} V = 0 & \mapsto \text{ Independencia} \\ V = 1 & \mapsto \text{ Asociación perfecta} \end{cases}$$

En las tablas de contingencia (2 x 2), V_{Cramer} y ϕ son idénticos.
El problema de este estadístico es que tiende a subestimar el grado de asociación entre las variables.



Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,283	,000
	V de Cramer	,200	,000
	Coeficiente de contingencia	,272	,000
N de casos válidos		500	

a. Asumiendo la hipótesis alternativa.

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

Cada medida viene acompañada de su correspondiente nivel crítico (*Sig. aproximada*), el cual permite decidir sobre la hipótesis de independencia, puesto que el nivel crítico de todas las medidas listadas es muy pequeño (menor que 0,05), se puede rechazar la hipótesis nula de independencia y concluir que el nivel de renta y la opinión sobre el sistema sanitario están relacionados.

$$\phi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{40{,}0492}{500}} = 0{,}283 \qquad V_{Cramer} = \sqrt{\frac{\chi^2}{N.\min(k-1, m-1)}} = \sqrt{\frac{40{,}0492}{500.2}} = 0{,}200$$

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{40{,}0492}{40{,}0492 + 500}} = 0{,}2723$$

MEDIDAS BASADAS EN LA REDUCCIÓN PROPORCIONAL DEL ERROR (RPE).- Son medidas de asociación que expresan la proporción en la que se consigue reducir la probabilidad de cometer un error de predicción cuando, al intentar clasificar un caso o grupo de casos como pertenecientes a una u otra categoría de una variable, en lugar de utilizar únicamente las probabilidades asociadas a cada categoría de esa variable, se efectúa la clasificación teniendo en cuenta las probabilidades de las categorías de esa variable en cada categoría de una segunda variable.

🕮 **COEFICIENTE LAMBDA DE GOODMAN Y KRUSKALL:** Es un coeficiente que no depende de la χ^2.

Suponiendo que Y es el factor explicado y X el explicativo, se evalúa la capacidad de X para predecir Y mediante:

$$\lambda_Y = \frac{\sum_i \max_j n_{ij} - \max_j N_{\bullet j}}{N - \max_j n_{ij}}$$

De forma análoga, cuando X es el factor explicado e Y el explicativo, se evalúa la capacidad de Y para predecir X mediante:

$$\lambda_X = \frac{\sum_j \max_i n_{ij} - \max_i N_{i\bullet}}{N - \max_i n_{ij}}$$

$0 \le (\lambda_X, \lambda_X) \le 1$ y son medidas asimétricas.

El coeficiente Lambda presenta tres versiones: dos asimétricas (cuando una de las dos variables se considera independiente) y una simétrica (cuando no existe argumento para distinguir), $0 \le \lambda \le 1$, el valor 0 indica que la variable independiente no contribuye en absoluto en reducir el error de predicción; el valor 1 indica que se ha conseguido por completo reducir el error de predicción, es decir, que la variable independiente predice con toda precisión a qué categoría de la variable dependiente pertenecen los casos clasificados.

Dos variables son independientes cuando $\lambda = 0$, pero $\lambda = 0$ no implica independencia estadística.

Cuando dos variables son estadísticamente independientes $\Rightarrow \lambda = 0$

$\lambda = 0 \xrightarrow{\text{No implica}}$ independencia estadística, pues lambda únicamente es sensible a un tipo particular de asociación: A la derivada de la reducción en el error que se consigue al predecir las categorías de una variable utilizando las de otra.

Cuando no es posible determinar objetivamente cuál de los dos factores es el explicado o el explicativo, se opta por la versión simétrica, cuyo valor es:

$$\lambda = \frac{\sum_i \max_j n_{ij} + \sum_j \max_i n_{ij} - \max_i N_{i\bullet} - \max_j N_{\bullet j}}{2N - \max_i N_{i\bullet} - \max_j N_{\bullet j}}$$

El valor de λ está comprendido entre λ_X y λ_Y presenta el inconveniente de ser muy sensible a la presencia de totales marginales desequilibrados.

☞ TAU DE GOODMAN Y KRUSKALL: $\tau = \dfrac{\frac{1}{N}\sum_i (N - N_{i\bullet})N_{i\bullet} - \sum_j \frac{(N_{\bullet j} - n_{ij})n_{ij}}{N_{\bullet j}}}{\frac{1}{N}\sum_i (N - N_{i\bullet})N_{i\bullet}}$

Opinión Sistema Público / Nivel renta	Bueno	Regular	Malo	Total marginal	
Bajo	75	35	40	$N_{1\bullet} = 150$	$\text{máx } n_{1j} = 75$
Medio	60	70	50	$N_{2\bullet} = 180$	$\text{máx } n_{2j} = 70$
Alto	20	30	40	$N_{3\bullet} = 90$	$\text{máx } n_{3j} = 40$
Muy Alto	15	25	40	$N_{4\bullet} = 80$	$\text{máx } n_{4j} = 40$
Total marginal	$N_{\bullet 1} = 170$	$N_{\bullet 2} = 160$	$N_{\bullet 3} = 170$	$N = 500$	$\sum_{i=1}^{4} \text{máx}_j\, n_{ij} = 225$
$\sum_{j=1}^{3} \text{máx}_i\, n_{ij} = 195$	$\text{máx } n_{i1} = 75$	$\text{máx } n_{i2} = 70$	$\text{máx } n_{i3} = 50$		

$$\lambda = \frac{\sum_i \text{máx}_j\, n_{ij} + \sum_j \text{máx}_i\, n_{ij} - \text{máx}_i\, N_{i\bullet} - \text{máx}_j\, N_{\bullet j}}{2N - \text{máx}_i\, N_{i\bullet} - \text{máx}_j\, N_{\bullet j}} = \frac{225 + 195 - 180 - 170}{2.500 - 180 - 170} = 0{,}1076 \quad \textit{coeficiente lambda}$$

$$\tau = \frac{\frac{1}{N}\sum_i (N - N_{i\bullet})N_{i\bullet} - \sum_j \frac{(N_{\bullet j} - n_{ij})n_{ij}}{N_{\bullet j}}}{\frac{1}{N}\sum_i (N - N_{i\bullet})N_{i\bullet}} = \frac{361{,}2 - 351{,}3051}{361{,}2} = 0{,}02739 \quad \textit{tau de Goodman y Kruskall}$$

$$\frac{1}{N}\sum_{i=1}^{4}(N - N_{i\bullet})N = \frac{1}{500}\left[(500-150)150 + (500-180)180 + (500-90)90 + (500-80)80\right] = 361{,}2$$

$$\sum_j \frac{(N_{\bullet j} - n_{ij})n_{ij}}{N_{\bullet j}} = \left[\frac{(170-75)75 + (170-60)60 + (170-20)20 + (170-15)15}{170}\right] +$$
$$+ \left[\frac{(160-35)35 + (160-70)70 + (160-30)30 + (160-25)25}{160}\right] +$$
$$+ \left[\frac{(170-40)40 + (170-50)50 + (170-40)40 + (170-40)40}{170}\right] = 351{,}3051$$

🕮 **COEFICIENTE DE INCERTIDUMBRE** (THEIL, 1970)**:** Es una medida semejante a *lambda* y *tau* en Cuanto a su concepción de la asociación de las variables, en relación a la capacidad predictiva y la disminución del error de dicha predicción.

La diferencia estriba en su cálculo ya que en este caso la expresión de estos coeficientes depende de toda la distribución y no sólo de los valores modales, por lo que sólo toma el *valor 0* en casos de *total independencia*. Ésta es su ventaja respecto a *lambda*, pero es más difícil de interpretar.
Oscila entre 0 y 1.

Posee dos versiones *asimétricas* (dependiendo de cuál de las dos variables se considera independiente) y una *simétrica* (cuando no se hace distinción entre variable dependiente e independiente).

Se obtiene de la siguiente forma: $I_{Y/X} = \dfrac{I(X)+I(Y)-I(XY)}{I(Y)}$

$$I(X) = -\sum_i \left(\frac{N_{i\bullet}}{N}\right) Ln\left(\frac{N_{i\bullet}}{N}\right) \qquad I(Y) = -\sum_j \left(\frac{N_{\bullet j}}{N}\right) Ln\left(\frac{N_{\bullet j}}{N}\right) \qquad I(XY) = -\sum_i\sum_j \left(\frac{n_{ij}}{N}\right) Ln\left(\frac{n_{ij}}{N}\right)$$

Para obtener $I_{X/Y}$ basta intercambiar los papeles de I(X), I(Y).

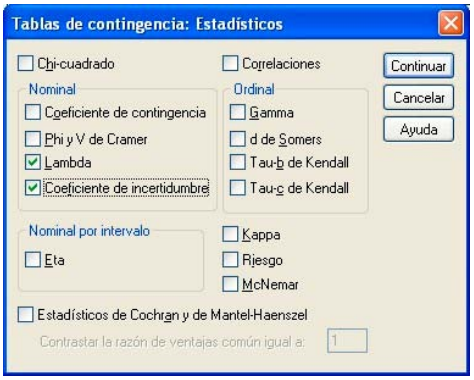
La versión *simétrica* se obtiene: $I_{Y/X} = 2\,\dfrac{I(X)+I(Y)-I(XY)}{I(X)+I(Y)}$

En el ejemplo:

$$I(X) = -\sum_i \left(\frac{N_{i\bullet}}{N}\right) Ln\left(\frac{N_{i\bullet}}{N}\right) = 1{,}3309 \qquad I(Y) = -\sum_j \left(\frac{N_{\bullet j}}{N}\right) Ln\left(\frac{N_{\bullet j}}{N}\right) = 1{,}0982$$

$$I(XY) = -\sum_i\sum_j \left(\frac{n_{ij}}{N}\right) Ln\left(\frac{n_{ij}}{N}\right) = 2{,}3894$$

$$I_{Y/X} = 2\,\frac{I(X)+I(Y)-I(XY)}{I(X)+I(Y)} = 2\,\frac{1{,}3309+1{,}0982-2{,}3894}{1{,}3309+1{,}0982} = 0{,}03268$$



Medidas direccionales

			Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Nominal por nominal	Lambda	Simétrica	,108	,033	3,162	,002
		Nivel_renta dependiente	,047	,035	1,293	,196
		Opinión_sistema_ sanitario dependiente	,167	,043	3,558	,000
	Tau de Goodman y Kruskal	Nivel_renta dependiente	,027	,009		,000[c]
		Opinión_sistema_ sanitario dependiente	,040	,013		,000[c]
	Coeficiente de incertidumbre	Simétrica	,033	,010	3,188	,000[d]
		Nivel_renta dependiente	,030	,009	3,188	,000[d]
		Opinión_sistema_ sanitario dependiente	,036	,011	3,188	,000[d]

a. Asumiendo la hipótesis alternativa.
b. Empleando el error típico asintótico basado en la hipótesis nula.
c. Basado en la aproximación chi-cuadrado.
d. Probabilidad del chi-cuadrado de la razón de verosimilitudes.

Cada medida (Lambda, Tau de Goodman-Kruskall, Coeficiente de incertidumbre) figura acompañada de su nivel crítico (*Sig*. *aproximada*), que como es pequeño, menor que 0,05, conduce a rechazar la hipótesis nula de independencia, concluyendo que las variables en estudio (Nivel_renta, Opinión_sistema_sanitario) están relacionadas.

Junto al valor concreto adoptado por cada medida de asociación aparece su valor estandarizado (*T aproximada*), que se obtiene dividiendo el valor de la medida entre su error típico (calculado éste suponiendo independencia entre las variables. La tabla también muestra el error típico de cada medida calculado sin suponer independencia (*Error típico asintótico*).

DATOS ORDINALES.- Recoge una serie de medidas de asociación que permite aprovechar la información ordinal que las medidas diseñadas para datos nominales pasan sin analizar.

Con datos ordinales tiene sentido hablar de *dirección* de la relación: una *dirección positiva* indica que los valores altos de una variable se asocian con los valores altos de la otra variable, y los valores bajos con valores bajos; una *dirección negativa* indica que los valores altos de una variable se asocian con los valores bajos de la otra, y los valores bajos con los valores altos.

Muchas de las medidas de asociación diseñadas para estudiar la relación entre variables ordinales se basan en el concepto de *concordancia* (*inversión)* y *discordancia* (*no-inversión)*.

- ☞ Si los dos valores de un caso en ambas variables son mayores (o menores) que los dos valores de otro caso, se da una *concordancia o no-inversión* (C).
- ☞ Si el valor de un caso en una de las variables es mayor que el del otro caso, y en la segunda variable el valor del segundo caso es mayor que el del primero, se da una *discordancia* o *inversión* (D).
- ☞ Si los dos casos tienen valores idénticos en una o en las dos variables, se da un *empate* (E). Hay tres tipos de empates: empate en la variable X y no en la variable Y (E_X), empate en la variable Y y no en la variable X (E_Y), y empate en ambas variables (E_{XY}).

Cuando predominan las *concordancias*, la relación es *positiva*, a medida que aumentan (o disminuyen) los valores de una de las variables, aumentan (o disminuyen) los de la otra.

Cuando predominan las *discordancias*, la relación es *negativa*, a medida que aumentan (o disminuyen) los valores de una de las variables, disminuyen (o aumentan) los de la otra.

Todas las medidas de asociación mencionadas en este punto, utilizan en el numerador la diferencia la *diferencia* entre el número de *discordancias* y *concordancias* resultantes de comparar cada caso con cada caso, diferenciándose en el tratamiento dado a los *empates*.

En adelante, $C \equiv$ número de concordancias, $D \equiv$ número de discordancias, $E_X \equiv$ número empates en la variable X (tomando a Y como independiente), $E_Y \equiv$ número empates en la variable Y (tomando a X como independiente), $E_{XY} \equiv$ número empates en ambas variables.

El total de pares de valores que es posible encontrar (T), sin repeticiones, siendo N el total de casos, viene dado por la expresión:

$$T = \frac{N(N-1)}{2} \qquad T = C + D + E_X + E_Y + E_{XY}$$

🕮 **GAMMA:** Uno de los coeficientes más conocidos es el coeficiente *gamma* (γ) de Goodman y Kruskall, para la gamma los empates son irrelevantes, se basa en la relación que siguen los rangos de los dos atributos:

$$\gamma = \frac{C-D}{C+D} \begin{cases} \gamma = -1 & \text{perfecta asociación negativa} \\ \gamma = 0 & \text{independencia} \\ \gamma = 1 & \text{perfecta asociación positiva} \end{cases}$$

📖 **d DE SOMMER:** A diferencia de los anteriores, este estadístico considera que las variables pueden ser simétricas o dependientes. En el caso de ser simétricas, el estadístico *d de Somers* coincide con la *Tau-b* de Kendall. En caso de ser dependientes, se diferencia del estadístico *Gamma* en que incluye los empates a la que considera dependiente, lo que da lugar a tres índices (dos asimétricos y uno simétrico):

☞ Tomando a la variable Y como independiente: $d_X = \dfrac{(C-D)}{C+D+E_X}$ (E_X son empates en la variable X)

☞ Tomando a la variable X como independiente: $d_Y = \dfrac{(C-D)}{C+D+E_Y}$ (E_Y son empates en la variable Y)

☞ Versión simétrica para X e Y: $d = \dfrac{(C-D)}{C+D+\left[\dfrac{E_X+E_Y}{2}\right]}$

$$-1 \le d \le 1 \quad \begin{cases} d=-1 & \rightarrow \text{ perfecta asociación negativa} \\ d=0 & \rightarrow \text{ independencia} \\ |d| \mapsto 1 & \rightarrow \text{ asociación a medida} \\ d=1 & \rightarrow \text{ perfecta asociación positiva} \end{cases}$$

📖 **COEFICIENTE DE CONCORDANCIA DE KENDALL:** En un proceso de evaluación, se emplea en situaciones, en la que interesa obtener una medida de la coincidencia de un conjunto de evaluadores.

$$W = \frac{12\sum_{i=1}^{N}(R_i - \overline{R})}{k^2 N(N^2-1)} \quad \text{donde} \begin{cases} \overline{R} \equiv \text{media de los rangos} \\ k \equiv \text{número variables incluidas} \end{cases}$$

📖 **Tau-a de Kendall:** Se define como la razón entre la diferencia del número pares concordantes y discordantes con la totalidad del número de pares posibles (T).

$$\tau_a = \frac{(C-D)}{T}$$

El coeficiente τ_a toma valores entre [-1, 1], el signo indica el sentido de la asociación, el 0 la independencia estadística. Una limitación importante es que en ciertos casos no alcanza el 1 (–1) aún existiendo asociación perfecta, pues para esto requiere que todas las celdas centrales tengan frecuencia 0.

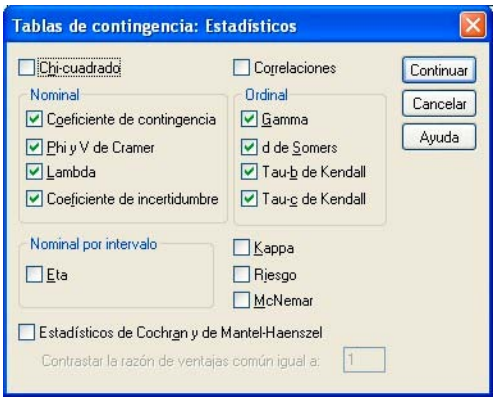
📖 **Tau-b de Kendall:** Utiliza el mismo criterio que la *d* de Sommer simétrica, si bien utiliza la media geométrica en lugar de la media aritmética:

$$\tau_b = \frac{(C-D)}{\sqrt{(C+D+E_X)(C+D+E_Y)}}$$

El coeficiente τ_b toma valores entre [-1, 1], solo en las tablas de contingencia cuadradas y si ninguna frecuencia marginal vale cero.

🕮 **Tau-c de Kendall:** Utiliza el mínimo de filas y de columnas:

$$\tau_c = \frac{2m(C-D)}{N^2(m-1)} \qquad m = \text{mín}\{\text{nº filas, nº columnas}\}, \qquad -1 \le \tau_c \le 1$$



Medidas simétricas

		Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Nominal por nominal	Phi	,283			,000
	V de Cramer	,200			,000
	Coeficiente de contingencia	,272			,000
Ordinal por ordinal	Tau-b de Kendall	,216	,039	5,591	,000
	Tau-c de Kendall	,225	,040	5,591	,000
	Gamma	,307	,054	5,591	,000
N de casos válidos		500			

a. Asumiendo la hipótesis alternativa.

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

Cada coeficiente aparece con su correspondiente nivel crítico (*Sig. aproximada*), puesto que estos niveles críticos son menores que 0,05, se rechaza la hipótesis nula de independencia, afirmando que las variables (Nivel_renta, Opinión_sistema_sanitario) están relacionadas.

Y como el valor de las medias es positivo (relación positiva), se puede interpretar que a un mayor nivel de renta corresponde una mejor opinión del sistema sanitario.

🕮 **NOMINAL POR INTERVALO (ETA):** El coeficiente de correlación *eta* (η) sirve para cuantificar Grado de asociación existente ente una variable cuantitativa (medida en escala de intervalo o razón) y una variable categórica (medida en escala nominal u ordinal).

Se trata de un coeficiente de correlación que no supone linealidad y cuyo cuadrado puede interpretarse, si el diseño lo permite, como la proporción de varianza de la variable cuantitativa qué esta explicada por (que depende de) la variable categórica.
Su mayor utilidad no está asociada a las tablas de contingencia, pues éstas se construyen utilizando variables categóricas. A pesar de ello, como es una opción de SPSS, se puede marcar la opción *eta* y obtener el valor de la relación entre dos variables cuando una de ellas es cuantitativa y la otra categórica.

☞ *Cuando las filas son el atributo dependiente*: $$\eta = \sqrt{1 - \frac{\sum_j \left(\sum_i x_i^2 n_{ij} \right) - \frac{\left(\sum_i x_i^2 n_{ij} \right)^2}{\sum_i n_{ij}}}{\sum_i \sum_j x_i^2 n_{ij} - \frac{\left(\sum_i \sum_j x_i n_{ij} \right)^2}{N}}}$$

☞ *Cuando las columnas son el atributo dependiente*: $\eta = \sqrt{1 - \dfrac{\sum_i \left(\sum_j y_j^2 n_{ij} \right) - \dfrac{\left(\sum_j y_j^2 n_{ij} \right)^2}{\sum_j n_{ij}}}{\sum_j \sum_i y_j^2 n_{ij} - \dfrac{\left(\sum_i \sum_j y_j n_{ij} \right)^2}{N}}}$

🕮 **ÍNDICE DE ACUERDO (kappa):** Proporciona una medida del grado de acuerdo existente entre dos observadores o jueces al evaluar una serie de sujetos u objetos (Cohen, 1960):

$$\kappa = \frac{N \sum_i n_{ii} - \sum_i N_{i\bullet} N_{\bullet i}}{N^2 - \sum_i N_{i\bullet} N_{\bullet i}}$$

El valor kappa κ, $0 \le \kappa \le 1$, toma el valor 0 (acuerdo nulo) y 1 (acuerdo máximo).

Si el acuerdo alcanzado es menor que el esperado por azar, kappa toma un valor negativo.

Fleiss, Cohen y Everitt (1969) demostraron que el error típico del coeficiente de Kappa puede estimarse mediante la expresión:

$$\hat{\sigma}_\kappa^2 = \frac{1}{N \left(N^2 - \sum_i N_{i\bullet} N_{\bullet i} \right)^2} \left[N^2 \sum_i \sum_i N_{i\bullet} N_{\bullet i} + \left(\sum_i N_{i\bullet} N_{\bullet i} \right)^2 - N \sum_i N_{i\bullet} N_{\bullet i} \left(N_{i\bullet} + N_{\bullet i} \right) \right]$$

La hipótesis de que los dos observadores(jueces) son independientes $(\kappa = 0)$ se puede contrastar tipificando el valor de kappa: $z_\kappa = \dfrac{\kappa}{\sigma_\kappa} \approx N(0,1)$

Al margen de la significación estadística del coeficiente de Kappa, Landis y Roch (1977) argumentaron que:

$$\begin{cases} \kappa < 0 & \rightarrow \text{ sin acuerdo} \\ 0 \le \kappa < 0{,}2 & \rightarrow \text{ insuficiente} \\ 0{,}2 \le \kappa < 0{,}4 & \rightarrow \text{ bajo} \end{cases} \qquad \begin{cases} 0{,}4 \le \kappa < 0{,}6 & \rightarrow \text{ moderado} \\ 0{,}6 \le \kappa < 0{,}8 & \rightarrow \text{ bueno} \\ 0{,}8 \le \kappa < 1 & \rightarrow \text{ muy bueno} \end{cases}$$

🕮 **ÍNDICES DE RIESGO:** Son útiles para diseños longitudinales en los que se miden dos variables dicotómicas. El seguimiento en estudios longitudinales puede realizarse: *hacia delante* o *hacia atrás*.

☞ En los estudios longitudinales *hacia delante*, denominados *diseños prospectivos* o *de cohortes*, los sujetos son clasificados en dos grupos dependiendo de la presencia o ausencia de algún factor desencadenante y, se les hace un seguimiento durante un período de tiempo, hasta determinar la proporción de sujetos de cada grupo en los que se da un determinado desenlace objeto de estudio.

La medida de interés es el *riesgo relativo* (R_r): $R_r = \dfrac{n_{11}/N_{1\bullet}}{n_{21}/N_{2\bullet}}$

La interpretación: << *La proporción de desenlaces entre los sujetos expuestos al factor desencadenante es* R_r *más alta que entre los sujetos no expuestos* >>.

Un riesgo relativo de 1 indica que la probabilidad de encontrar el desenlace es la misma tanto en el grupo de sujetos expuestos como en el grupo de sujetos no expuestos.

Para valorar si el riesgo obtenido es significativamente distinto de 1, se puede calcular el intervalo de confianza:

$$\left[R_r\, e^{z_{\alpha/2}\sqrt{\frac{n_{12}}{n_{11}N_{1\bullet}}+\frac{n_{22}}{n_{21}N_{2\bullet}}}} \;;\; R_r\, e^{z_{1-\alpha/2}\sqrt{\frac{n_{12}}{n_{11}N_{1\bullet}}+\frac{n_{22}}{n_{21}N_{2\bullet}}}} \right]$$

Si el intervalo de confianza no cubre el valor 1, el riesgo de experimentar el desenlace no es lo mismo en los grupos que se comparan.

☞ En los diseños longitudinales *hacia atrás*, también llamados *retrospectivos* o *caso-control*, se forman grupos de sujetos a partir de la presencia o ausencia de una determinada condición objeto de estudio, y se hace el seguimiento hacia atrás, intentando encontrar información sobre la proporción en la que se encuentra presente en cada muestra un determinado factor desencadenante.

Como el tamaño de los grupos se fija a partir de la presencia o ausencia de un determinado desenlace, se calcula *odds ratio* (*razón de las ventajas* o *razón de productos cruzados*):

$$O_r = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}\,n_{22}}{n_{12}\,n_{21}}$$

El valor de la *odds ratio* (O_r) es tanto mejor estimador del riesgo relativo cuanto más pequeñas sean las proporciones de desenlace en cada grupo.

Un índice de 1 indica que la probabilidad de encontrarse con el factor desencadenante en los grupos estudiados es la misma. Para determinar si este riesgo es significativamente distinto de 1, se calcula el intervalo de confianza mediante:

$$\left[O_r\, e^{z_{\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}} \;;\; O_r\, e^{z_{1-\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}} \right]$$

🕮 **ESTADÍSTICOS DE COCHRAN Y DE MANTEL-HAENSZEL** (Combinación de Tablas 2x2)**:** En ocasiones, puede interesar analizar los *diseños de cohortes* y de *caso-control* (tratados en los *índices de riesgo*) controlando el efecto de terceras variables.

Esta situación es propia, por ejemplo, cuando se desea evaluar el efecto de un tratamiento sobre una determinada respuesta utilizando distintos grupos de pacientes.

Se trata de estudiar si existe o no asociación entre una variable *factor* y otra variable *respuesta*, ambas dicotómicas, cuando se dispone de información referida a varios *estratos* (distintos grupos de edad o sexo, pacientes con diferente sintomatología, distintos grupos étnicos, distintas dosis de fármaco, etc.).

En estos casos, utilizar el estadístico χ^2 sobre el conjunto de datos agrupados puede arrojar resultados inadecuados. Por otra parte, analizar separadamente cada *estrato* no proporciona una idea global del efecto de la variable *factor*.

Utilizando los estadísticos de Cochran y Mantel-Haenszel (según un modelo χ_1^2) se obtiene una información más ajustada para contrastar la hipótesis de independencia condicional, esto es, la hipótesis de independencia entre las variables *factor* y *respuesta* una vez que se ha controlado el efecto de los estratos.

El estadístico de Cochran (1954) se expresa mediante: $\chi^2_{Cochran} = \dfrac{\left(\sum_k n_k - \sum_k e_k\right)^2}{\sum_k \sigma^2_{n_k}}$

$$\begin{cases} k \equiv \text{ cada uno de los estratos} \\ n_k \equiv \text{frecuencia observada en casilla del estrato } k \text{ (sólo una y siempre la misma en todos los estratos)} \\ e_k \equiv \text{frecuencia esperada en cada correspondiente } n_k \\ \sigma^2_{n_k} \equiv (n_{1+k} + n_{2+k} + n_{+1k} + n_{+2k})/N^3 \text{ donde } (n_{1+k}, n_{2+k}, n_{+1k}, n_{+2k}) \text{ frecuencias marginales asociadas tabla 2x2} \end{cases}$$

El estadístico de Mantel-Haenszel (1959) es idéntico al de Cohran, excepto en dos detalles:

- Utiliza la corrección por continuidad (resta 0,5 al numerador antes de elevarlo al cuadrado)
- Cambia el denominador de la varianza, utilizando $N^2(N-1)$ en lugar de N^3

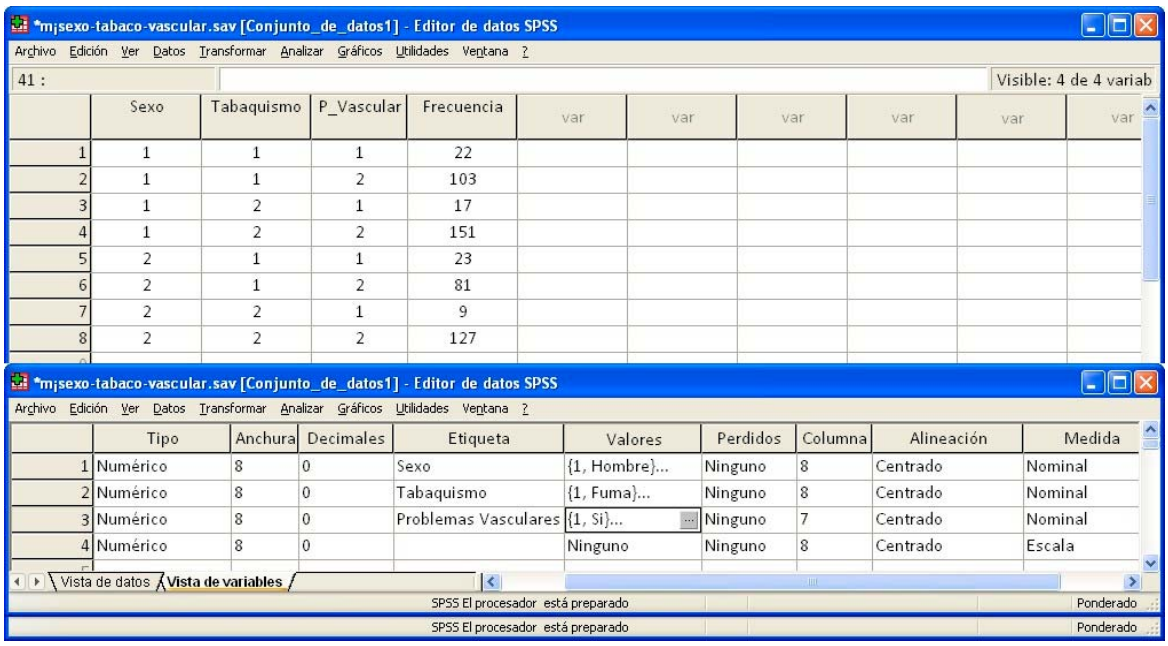
Los dos estadísticos se distribuyen según el modelo de probabilidad χ_1^2.

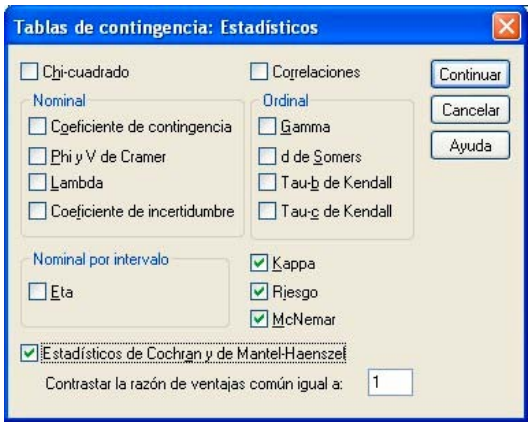
Cuando el nivel crítico asociado a ellos es menor que 0,05, se rechaza la hipótesis nula de independencia condicional y se concluye que, una vez controlado el efecto de los estratos, las variables *factor* y *respuesta* están asociadas.

$$RV_{comun} = \frac{\sum_k \left(n_{11k}\, n_{22k} / n_{++k}\right)}{\sum_k \left(n_{12k}\, n_{21k} / n_{++k}\right)}$$ *Riesgo homogéneo en todos los estratos de Mantel-Haenszel*

Ejemplo.- Tabla de contingencia de tabaquismo por problemas vasculares de hombres y mujeres.

Sexo			Problemas Vasculares Si	Problemas Vasculares No	Total
Hombre	Tabaquismo	Fuma	22	103	125
		No Fuma	17	151	168
	Total		39	254	293
Mujer	Tabaquismo	Fuma	23	81	104
		No Fuma	9	127	136
	Total		32	208	240





Pruebas de independencia condicional

	Chi-cuadrado	gl	Sig. asintótica (bilateral)
De Cochran	13,933	1	,000
Mantel-Haenszel	12,939	1	,000

Bajo el supuesto de independencia condicional, el estadístico de Cochran se distribuye asintóticamente según una distribución de chi-cuadrado con 1 gl, sólo si el número de estratos es fijo, mientras que el estadístico de Mantel-Haenszel se distribuye siempre asintóticamente según una distribución de chi-cuadrado con 1 gl. Tenga presente que se suprime la corrección por continuidad del estadístico de Mantel-Haenszel cuando la suma de las diferencias entre lo observado y lo esperado es igual a 0.

Pruebas de homogeneidad de la razón de las ventajas

	Chi-cuadrado	gl	Sig. asintótica (bilateral)
Breslow-Day	1,911	1	,167
De Tarone	1,910	1	,167

El estadístico de Cochran tiene un valor de 13,933 con un nivel crítico asociado (*Sig. asintótica bilateral*) de 0,000 (menor que 0,05) con lo que *se rechaza* la hipótesis nula de independencia

condicional y se concluye, controlado el efecto de la variable sexo, que las variables tabaquismo y problemas vasculares están relacionadas.
Análoga conclusión se llega con el estadístico de Mantel-Haenszel.

Rechazada la hipótesis de independencia condicional, el interés se centra hacia el *grado de dependencia* existente entre las variables *factor* y *respuesta*. En esta línea, SPSS ofrece una estimación del riesgo *odds-ratio* común para todos los estratos. Esta estimación común solo tiene sentido si existe interacción triple, esto es, si la interacción detectada es homogénea en todos los estratos.

La hipótesis de homogeneidad de las odds-ratio puede contrastarse utilizando los estadísticos Breslow-Day (1980, 1987) y de Tarone (1983). En la tabla de las pruebas de homogeneidad de la razón de las ventajas, se observa que el nivel crítico asociado a ambos estadísticos es de 0,167 (mayor que 0,05) por lo que se puede mantener la hipótesis nula de homogeneidad.

Partiendo de que el riesgo es homogéneo en todos los estratos, tiene sentido calcular una estimación común del riesgo.

El Visor de resultados de SPSS ofrece el dato basado en el estadístico de Mantel-Haenszel (1959).

Estimación de la razón de las ventajas común de Mantel-Haenszel

Estimación			2,608
ln(estimación)			,959
Error típ. de ln(estimación)			,264
Sig. asintótica (bilateral)			,000
Intervalo de confianza asintótico al 95%	Razón de ventajas común	Límite inferior	1,555
		Límite superior	4,373
	ln(Razón de ventajas común)	Límite inferior	,442
		Límite superior	1,475

La estimación de la razón de las ventajas común de Mantel-Haenszel se distribuye de manera asintóticamente normal bajo el supuesto de razón de las ventajas común igual a 1,000. Lo mismo ocurre con el log natural de la estimación.

Pruebas de chi-cuadrado

Sexo		Valor	Sig. exacta (bilateral)
Hombre	Prueba de McNemar		,000[a]
	N de casos válidos	293	
Mujer	Prueba de McNemar		,000[a]
	N de casos válidos	240	

a. Utilizada la distribución binomial

En la prueba de *NcNemar*, se presenta el nivel crítico asociado al número de cambios observados (*Sig. exacta bilateral*) y el número de casos válidos.

La tabla no muestra el valor del estadístico de *NcNemar*, lo que significa que el nivel crítico se ha calculado utilizando la *distribución binomial* (como se detalla), permitiendo de este modo obtener la probabilidad exacta y no aproximada (que se obtendría con el estadístico *chi-cuadrado*).

Cualquiera que fuera la forma de obtenerlo, el nivel crítico indica el grado de compatibilidad existente entre los datos muestrales y la hipótesis nula de *igualdad de proporciones antes-después*.

Siendo el nivel crítico 0,000 (< 0,005), se rechaza la hipótesis nula, concluyendo que la proporción de hombres y mujeres con problemas vasculares cambia significativamente entre fumadores y no fumadores.

Medidas simétricas

Sexo			Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Hombre	Medida de acuerdo	Kappa	,082	,045	1,864	,062
	N de casos válidos		293			
Mujer	Medida de acuerdo	Kappa	,169	,050	3,500	,000
	N de casos válidos		240			

a. Asumiendo la hipótesis alternativa.

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

El valor del estadístico Kappa (0,082 y 0,169, respectivamente) y su nivel crítico (*Sig. aproximada*) permiten decidir sobre la hipótesis de *acuerdo nulo.*

En el caso de los hombres, al ser el nivel crítico en hombres (0,062 > 0,05) se *acepta* la hipótesis nula de *acuerdo nulo*. Según Landis y Roch (1977), siendo $\kappa = 0{,}082 < 0{,}20$ habría un *acuerdo insuficiente*. En el caso de las mujeres, el nivel crítico (0 < 0,05), se puede *rechazar* la hipótesis nula de acuerdo nulo, concluyendo que existe un acuerdo significativamente más alto que el esperado por azar. Según Landis y Roch, siendo $\kappa = 0{,}169 < 0{,}20$ habría un *acuerdo insuficiente*.

Junto al valor del índice de Kappa aparece su valor estandarizado (*T aproximada*), obtenido al dividir el valor de kappa entre su error típico (calculado bajo el supuesto de acuerdo nulo). Por otra parte, el error típico de Kappa (*Error típico asintótico*) se calcula sin suponer acuerdo nulo.

Estimación de riesgo

Sexo		Valor	Intervalo de confianza al 95% Inferior	Superior
Mujer	Razón de las ventajas para Tabaquismo (Fuma / No Fuma)	4,007	1,766	9,093
	Para la cohorte Problemas Vasculares = Si	3,342	1,615	6,915
	Para la cohorte Problemas Vasculares = No	,834	,746	,933
	N de casos válidos	240		

La primera fila indica el *riesgo* estimado entre mujeres fumadoras y no fumadoras, en un diseño de *caso-control* (razón de las ventajas), el valor 4,007 se interpreta como que, entre las mujeres con problemas vasculares, la *probabilidad o riesgo* de encontrar mujeres fumadoras es 4 veces mayor que la de encontrar mujeres no fumadoras.

La *razón de ventajas* también puede interpretarse como una estimación del riesgo relativo (en particular, sí la proporción de desenlaces es pequeña), es decir, el problema de padecer problemas vasculares es 4 veces mayor entre mujeres fumadoras que en mujeres no fumadoras.

El intervalo de confianza $[1{,}766\ ;\ 9{,}093]$, con un nivel de confianza del 95%, indica que el riesgo obtenido es mayor que 1.

La segunda y tercera fila se encargan de índices de riesgo para un *diseño de cohortes.*

Si se desea analizar la *presencia de problemas vasculares* (2ª fila), *el riesgo o la probabilidad* de encontrar tal problema entre las fumadoras es 3,342 veces mayor que entre las no fumadoras, en otras palabras: << por cada mujer con problemas vasculares entre las no fumadoras, se pueden encontrar 3,34 mujeres con problemas vasculares entre las fumadoras>>.

Para analizar la *ausencia de problemas vasculares* (3ª fila), *la probabilidad o riesgo* de encontrar tal desenlace es menor entre las mujeres fumadoras, esto es: <<por cada mujer sin problemas vasculares entre las no fumadoras se pueden encontrar 0,834 mujeres sin problemas vasculares entre las fumadoras>>.

Las casillas o celdas de una tabla de contingencia contienen información muy variada: frecuencias (observadas, esperadas), porcentajes y residuos.

☞ **Frecuencias**:

- *Frecuencias observadas:* valor real de cada par de valores.
- *Frecuencias esperadas:* valor teórico que tuviera que tener cada casilla para que fueran independientes las variables.

☞ **Porcentajes:**

- *Porcentaje por fila:* valor de cada casilla entre el total marginal de su fila.
- *Porcentaje por columna:* valor de cada casilla entre el total marginal de su columna.
- *Porcentaje total:* valor que la frecuencia observada de una casilla representa respecto al número total de casos.

☞ **Residuos:** Diferencias entre las frecuencias observadas y esperadas de cada casilla. Son útiles para interpretar las pautas de asociación en una tabla:

- *No tipificados:* Diferencia ente la frecuencia observada y esperada.
- *Tipificados:* Es el residuo no tipificado dividido por la raíz cuadrada de su correspondiente frecuencia esperada. El valor esperado es 0 y su desviación típica es menor que 1, lo que hace que no se pueden interpretar como puntuaciones normales z. No obstante, sirven como indicadores del grado en que cada casilla contribuye al valor del estadístico χ^2, sumando el cuadrado de todos los residuos tipificados se obtiene el valor de chi-cuadrado.
- *Tipificados corregidos de Haberman (1973):* Residuos que se distribuyen con puntuaciones normales $N(0,1)$, se obtienen dividiendo el residuo de cada casilla por su error típico. Al distribuirse normalmente, son interpretables con mucha facilidad, de este modo, utilizando un nivel de confianza de 0,95, se puede afirmar que los residuos mayores de 1,96 avisan de casillas con más casos de los que debería haber en esa casilla si las variables analizadas fueran independientes, mientras que los residuos menores de –1,96 delatan casillas con menos casos de los que debería haber en esa casilla bajo la hipótesis de independencia.
  En tablas de contingencia con valores nominales, una vez establecido que entre dos variables existe una asociación significativa (con el estadístico chi-cuadrado), y habiendo cuantificado esta asociación (coeficiente de contingencia), los residuos tipificados corregidos son una herramienta muy útil para interpretar el significado de la asociación.

Tabla de contingencia por categoría vascular: Análisis de los Residuos.

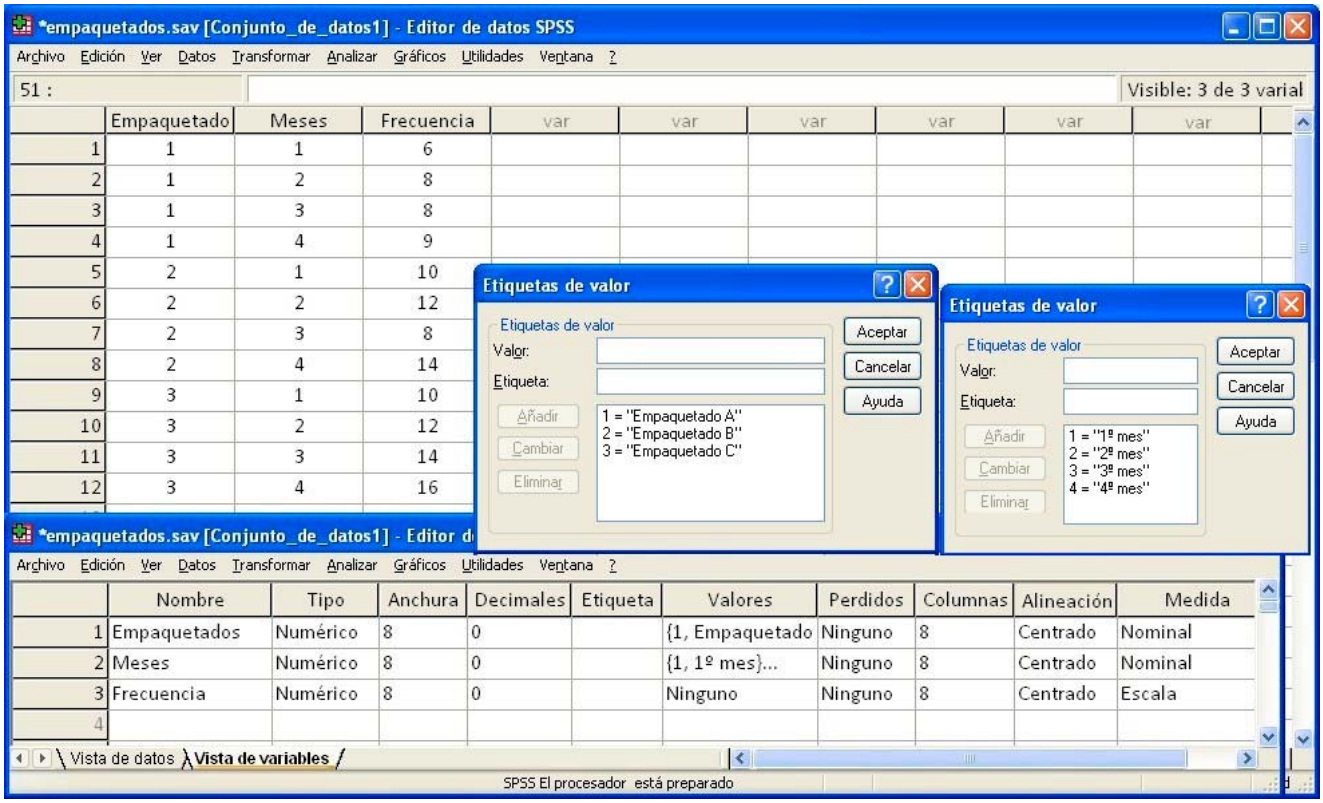
Tabla de contingencia Tabaquismo * Problemas Vasculares * Sexo

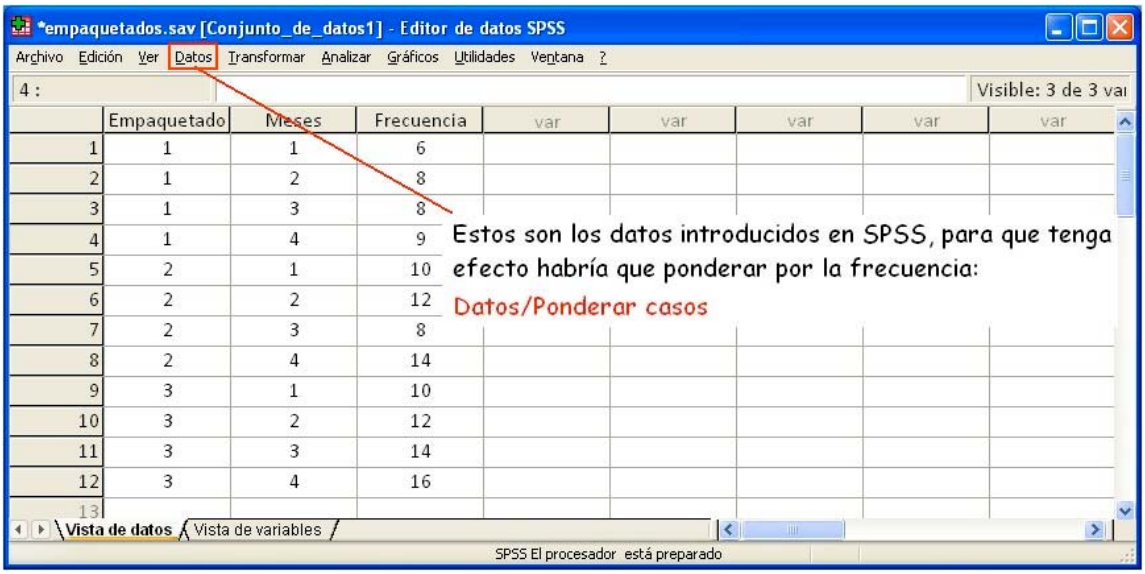
Sexo				Problemas Vasculares Si	Problemas Vasculares No	Total
Hombre	Tabaquismo	Fuma	Recuento	22	103	125
			Frecuencia esperada	16,6	108,4	125,0
			Residuos tipificados	1,3	-,5	
		No Fuma	Recuento	17	151	168
			Frecuencia esperada	22,4	145,6	168,0
			Residuos tipificados	-1,1	,4	
	Total		Recuento	39	254	293
			Frecuencia esperada	39,0	254,0	293,0
Mujer	Tabaquismo	Fuma	Recuento	23	81	104
			Frecuencia esperada	13,9	90,1	104,0
			Residuos tipificados	2,5	-1,0	
		No Fuma	Recuento	9	127	136
			Frecuencia esperada	18,1	117,9	136,0
			Residuos tipificados	-2,1	,8	
	Total		Recuento	32	208	240
			Frecuencia esperada	32,0	208,0	240,0

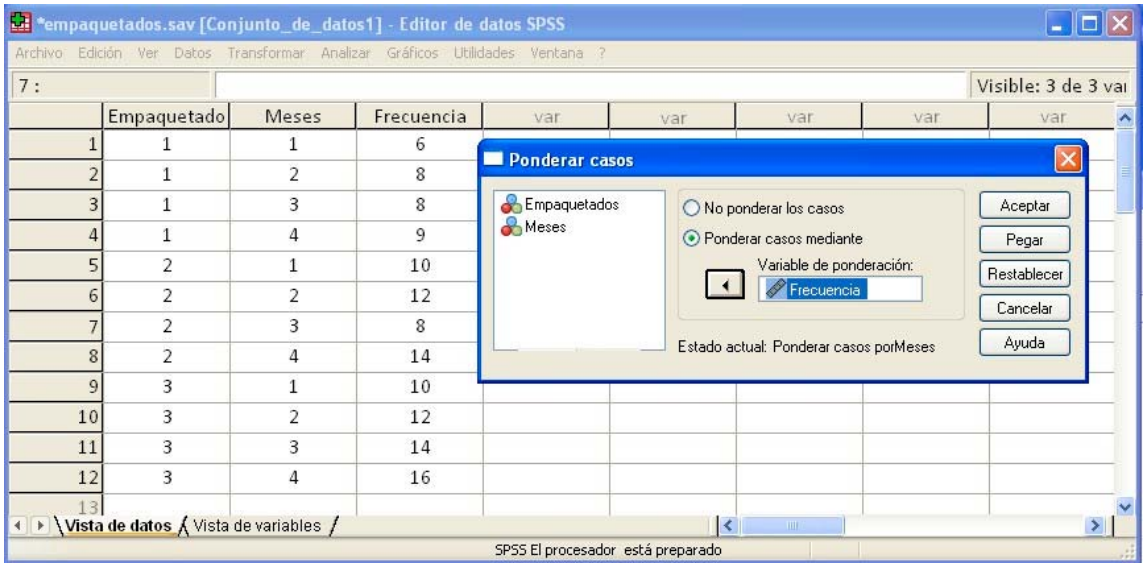
Los residuos tipificados fuera del intervalo $[-1{,}96\ ;\ 1{,}96]$, por ejemplo, en hombres y mujeres que fuman con problemas vasculares, existe una proporción significativa más alta en las mujeres (2,5 frente 1,3); mientras que entre los no fumadores existen una proporción más alta en hombres (-1,1 frente a –2,1).

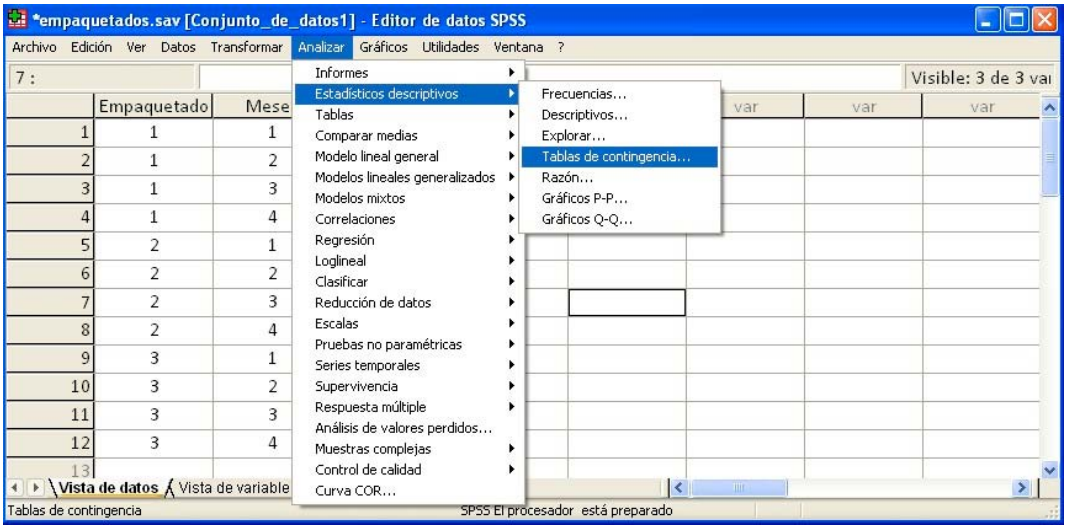
Ejercicio 1.- Tres métodos de empaquetado de tomates fueron probados durante un período de cuatro meses; se hizo un recuento del número de kilos por 1000 que llegaron estropeados, obteniéndose la tabla adjunta. Con un nivel de significación de 0,05, ¿tienen los tres métodos la misma eficacia?.

Meses	A	B	C	Total
1	6	10	10	26
2	8	12	12	32
3	8	8	14	30
4	9	14	16	39
Total	31	44	52	127









Para tomar una decisión sobre si hay diferencia entre los diferentes métodos de empaquetado, se contrasta la hipótesis nula, H_0: No hay diferencia entre los diferentes método de empaquetado, mediante una χ^2 de Pearson.

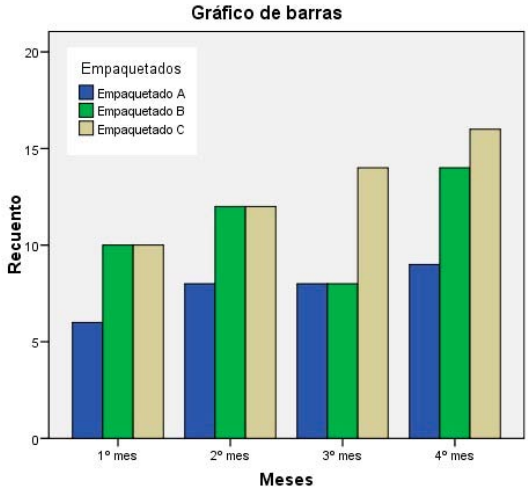
Se introducen los datos por filas y columnas.

- **Mostrar los gráficos de barras agrupadas:**

Activando la opción, el *Visor de resultados* muestra un gráfico de barras con las categorías de la variable fila (eje de abscisas) y las categorías de la variable columna anidadas dentro de las categorías de la variable fila.

En consecuencia, cada barra representa una casilla, y su altura viene dada por la frecuencia de la casilla.

- **Suprimir tablas:** Esta opción puede activarse si no se desea ninguna *tabla de contingencia*, esta decisión tendría sentido si solo se estuviera interesado en obtener un gráfico de barras o alguno de los estadísticos o medidas de asociación disponibles en el procedimiento *Tablas de contingencia*.



Para visualizar *frecuencias observadas* (n_{ij}) y *esperadas* (e_{ij}) en SPSS:

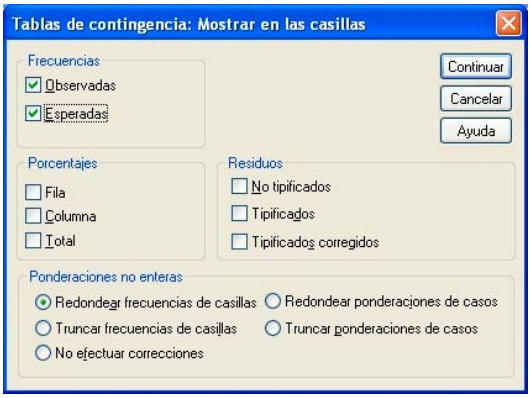


Tabla de contingencia Meses * Empaquetados

			Empaquetados			
			Empaquetado A	Empaquetado B	Empaquetado C	Total
Meses	1º mes	Recuento	6	10	10	26
		Frecuencia esperada	6,3	9,0	10,6	26,0
	2º mes	Recuento	8	12	12	32
		Frecuencia esperada	7,8	11,1	13,1	32,0
	3º mes	Recuento	8	8	14	30
		Frecuencia esperada	7,3	10,4	12,3	30,0
	4º mes	Recuento	9	14	16	39
		Frecuencia esperada	9,5	13,5	16,0	39,0
Total		Recuento	31	44	52	127
		Frecuencia esperada	31,0	44,0	52,0	127,0

Empaquetado / Meses	A	B	C	$N_{i\bullet}$
1	6 ($e_{11}=6,35$)	10 ($e_{12}=9,01$)	10 ($e_{13}=10,62$)	26 (26)
2	8 ($e_{21}=7,81$)	12 ($e_{22}=11,09$)	12 ($e_{23}=13,10$)	32 (32)
3	8 ($e_{31}=7,32$)	8 ($e_{32}=10,39$)	14 ($e_{33}=12,28$)	30 (30)
4	9 ($e_{41}=9,52$)	14 ($e_{42}=13,51$)	16 ($e_{43}=15,97$)	39 (39)
$N_{\bullet j}$	31	44	52	$N=127$

$$e_{11}=\frac{26.31}{127}=6,35 \qquad e_{21}=\frac{32.31}{127}=7,81 \qquad e_{31}=\frac{30.31}{127}=7,32 \qquad e_{41}=\frac{39.31}{127}=9,52$$

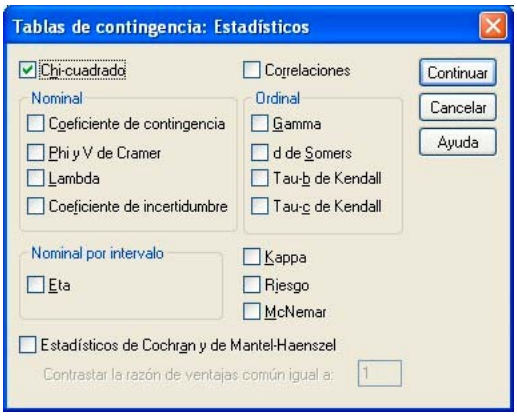
$$e_{12}=\frac{26.44}{127}=9,01 \qquad e_{22}=\frac{32.44}{127}=11,09 \qquad e_{32}=\frac{30.44}{127}=10,39 \qquad e_{42}=\frac{39.44}{127}=13,51$$

$$e_{13}=\frac{26.52}{127}=10,65 \qquad e_{23}=\frac{32.52}{127}=13,10 \qquad e_{33}=\frac{30.52}{127}=12,28 \qquad e_{43}=\frac{39.52}{127}=15,97$$

Estadístico de contraste: $\chi^2_{(4-1)(3-1)}=\chi^2_6=\sum_{i=1}^{4}\sum_{j=1}^{3}\frac{n_{ij}^2}{e_{ij}}-n=128,24-127=1,24$ (*estadístico observado*)

El estadístico teórico o esperado $\chi^2_{0,05\,;\,6} = 12,592$

Como $\chi^2_6 = 1,24 < \chi^2_{0,05\,;\,6} = 12,592$, el estadístico observado es menor que el estadístico teórico o esperado, se *acepta* la hipótesis nula, concluyendo que los tres métodos de empaquetado tienen la misma eficiencia.



Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	1,240[a]	6	,975
Razón de verosimilitudes	1,274	6	,973
Asociación lineal por lineal	,059	1	,808
N de casos válidos	127		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 6,35.

- *Estadístico de contraste (observado)* es 1,24, el cual, en la distribución χ^2 de Pearson tiene 6 grados de libertad $(gl = 6)$, tiene asociada una probabilidad Sig. asintótica (*Significación asintótica*) de 0,975.

Puesto que esta probabilidad (denominada *nivel crítico* o *nivel de significación observado*) es grande (0,975 > 0,05), se decide *aceptar* la hipótesis nula, y se concluye que los tres métodos de empaquetado tienen la misma eficiencia.

- *Razón de verosimilitud Chi-cuadrado:* Siendo, $RV = 2\sum_{i=1}^{4}\sum_{j=1}^{3} n_{ij} \log\left(\frac{n_{ij}}{e_{ij}}\right) = 1,274 < 12,592 = \chi^2_{0,05\,;\,6}$,

se acepta la hipótesis nula, y se concluye que los tres métodos de empaquetado tienen la misma eficiencia.

En la tabla, se observa como RV tiene asociada una probabilidad (*Sig. asintótica*) de 0,973, que como es mayor que 0,05, conduce a aceptar la hipótesis nula, llegando a la misma conclusión.

Señalar, que en caso contrario, se elige el estadístico con menor *Sig. asintótica*.

- *La corrección por continuidad de Yates:* $\chi^2_c = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left(\left|n - e_{ij}\right| - 0,5\right)^2}{e_{ij}} = 0,59$

Algunos autores sugieren, que con muestras pequeñas, esta corrección permite que el estadístico χ^2 se ajuste mejor a las probabilidades de la distribución χ^2, pero no existe un consenso generalizado sobre la utilización de esta corrección.

En el análisis de **MEDIDAS SIMÉTRICAS** se encuentran las *medidas nominales, medidas ordinales, coeficiente de correlación de Spearman y el coeficiente de correlación de Pearson.*

Las medidas nominales permiten contrastar la independencia sin decir nada sobre la fuerza de asociación entre las variables, informan *únicamente del grado de asociación existente*, no de la dirección o de la naturaleza de tal asociación. Son medidas basadas en el estadístico chi-cuadrado: Phi, V de Cramer y el Coeficiente de Contingencia.

Las *medidas ordinales* que recogen la *dirección de la asociación* de las variables: una relación *positiva* indica que los *valores altos* de una variable se asocian con los *valores altos* de la otra variable, y los *valores bajos* con los *valores bajos*; una relación *negativa* indica que los *valores altos* de una variable se asocian con los *valores bajos* de la otra variable, y los *valores bajos* con los *valore altos*.

Estas medidas se basan en el concepto de *concordancias* (o *inversión)* y *discordancias* (o *no-inversión*). Las medidas de asociación (Gamma, Tau-b, Tau-c) utilizan en el numerador la *diferencia* entre el número de *concordancias* o *inversiones* y *discordancias* o *no-inversiones* resultantes de comparar cada caso con otro, diferenciándose en el tratamiento dado a los *empates*.

Medidas simétricas

		Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Nominal por nominal	Phi	,099			,975
	V de Cramer	,070			,975
	Coeficiente de contingencia	,098			,975
Ordinal por ordinal	Tau-b de Kendall	,020	,075	,271	,786
	Tau-c de Kendall	,021	,078	,271	,786
	Gamma	,029	,107	,271	,786
	Correlación de Spearman	,024	,087	,266	,791[c]
Intervalo por intervalo	R de Pearson	,022	,087	,243	,809[c]
N de casos válidos		127			

a. Asumiendo la hipótesis alternativa.

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

c. Basada en la aproximación normal.

Cada medida de asociación aparece acompañada de su correspondiente nivel crítico (Sig. aproximada), permitiendo decidir sobre la hipótesis de igualdad de eficiencia, puesto que el nivel crítico de todas las medidas listadas es grande (mayor que 0,05 en todos los casos) se acepta la hipótesis nula de igualdad de eficiencia.

Al lado del valor de cada coeficiente se encuentra su valor estandarizado (T aproximada: valor del coeficiente dividido por su error típico), así como el error típico del valor de cada coeficiente obtenido sin suponer independencia (Error típico asintótico).

- Phi: $\phi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{1{,}240}{127}} = 0{,}099$

- V de CRAMER: $V_{Cramer} = \sqrt{\frac{\chi^2}{N.\min(k-1, m-1)}} = \sqrt{\frac{1{,}240}{127.\min(4-1, 3-1)}} = \sqrt{\frac{1{,}240}{127.2}} = 0{,}07$

- Coeficiente de Contingencia (grado de relación o dependencia):

$$C=\sqrt{\frac{\chi^2}{\chi^2+N}}=\sqrt{\frac{1{,}240}{1{,}240+127}}=0{,}098$$

$$\underset{\text{independen cia}}{0} \leq C \leq \underset{\text{asociación perfecta}}{1}$$

- Para calcular los coeficientes ordinales (Tau-b, Tau-c y Gamma) se requiere saber el número de pares concordantes (C) , discordantes (D) y empates (E). Partiendo de la información obtenida:

Meses	A	B	C
1	6	10	10
2	8	12	12
3	8	8	14
4	9	14	16

Número de pares concordantes: surgen del producto de las celdas externas por el sumando de las frecuencias de las celdas internas.

6	10	10
8	12	12
8	8	14
9	14	16

$6(12+12+8+14+14+16)=456$

6	10	10
8	12	12
8	8	14
9	14	16

$10(12+14+16)=420$

6	10	10
8	12	12
8	8	14
9	14	16

$8(8+14+14+16)=416$

6	10	10
8	12	12
8	8	14
9	14	16

$12(14+16)=360$

6	10	10
8	12	12
8	8	14
9	14	16

$8(14+16)=240$

6	10	10
8	12	12
8	8	14
9	14	16

$8(16)=128$

$C=456+420+416+360+240+128=2020$ *número de pares concordantes*

Número de pares discordantes: razonamiento análogo, partiendo de la celda opuesta.

6	10	10
8	12	12
8	8	14
9	14	16

$10(8+12+8+8+9+14)=590$

6	10	10
8	12	12
8	8	14
9	14	16

$10(8+8+9)=250$

6	10	10
8	12	12
8	8	14
9	14	16

$12(8+8+9+14)=468$

6	10	10
8	12	12
8	8	14
9	14	16

$12(8+9)=204$

6	10	10
8	12	12
8	8	14
9	14	16

$14(9+14)=322$

6	10	10
8	12	12
8	8	14
9	14	16

$8(9)=72$

$D=590+250+468+204+322+72=1906$ *número de pares discordantes*

Como predominan las *concordancias (2020)*, la relación es *positiva*, a medida que aumentan (o disminuyen) los valores de una de las variables, aumentan (o disminuyen) los de la otra.

Cálculo de pares empatados (E_X) *en la variable* X:

6		
8		
8		
9		

$6(8+8+9)=150$

8		
8		
9		

$8(8+9)=136$

8		
9		

$8(9)=72$

	10	
	12	
	8	
	14	

$10(12+8+14)=340$

	12	
	8	
	14	

$12(8+14)=264$

	8	
	14	

$8(14)=112$

		10
		12
		14
		16

$10(12+14+16)=420$

		12
		14
		16

$12(14+16)=360$

		14
		16

$14(16)=224$

El número de pares empatados en la variable X será:

$$E_X=150+136+72+340+264+112+420+360+224=2078$$

Cálculo de pares empatados (E_Y) *en la variable* Y:

6	10	10

$6(10+10)=120$

8	12	12

$8(12+12)=192$

8	8	14

$8(8+14)=176$

	10	10

$10(10)=100$

	12	12

$12(12)=144$

	8	14

$8(14)=112$

9	14	16

$9(14+16)=270$

	14	16

$14(16)=224$

El número de pares empatados en la variable Y será:

$$E_Y = 120+192+176+100+144+112+270+224 = 1338$$

El cálculo de pares empatados en ambas variables viene expresado: $E_{XY} = \sum_{i,j} \frac{n_{ij}(n_{ij}-1)}{2}$

Meses	A	B	C
1	6 (15)	10 (45)	10 (45)
2	8 (28)	12 (66)	12 (66)
3	8 (28)	8 (28)	14 (91)
4	9 (36)	14 (91)	16 (120)

$$E_{XY} = \sum_{i=1}^{4}\sum_{j=1}^{3} \frac{n_{ij}(n_{ij}-1)}{2} = 659$$

Calculados el número de pares de valores concordantes, discordantes, y empates, se puede determinar los distintos coeficientes para determinar el grado de asociación entre las variables ordinales.

El total de pares de valores que es posible encontrar (T), sin repeticiones, siendo N el total de casos, viene dado por la expresión:

$$T = \frac{N(N-1)}{2} = \frac{127.126}{2} = 8001$$

Adviértase que, $T = C+D+E_X+E_Y+E_{XY} = 2020+1906+2078+1338+659 = 8001$

- Gamma (los empates son irrelevantes): $\gamma = \frac{C-D}{C+D} = \frac{2020-1906}{2020+1906} = 0{,}029$

- Tau-a de Kendall:

$$\tau_a = \frac{(C-D)}{T} = \frac{(2020-1906)}{8001} = 0{,}0142$$

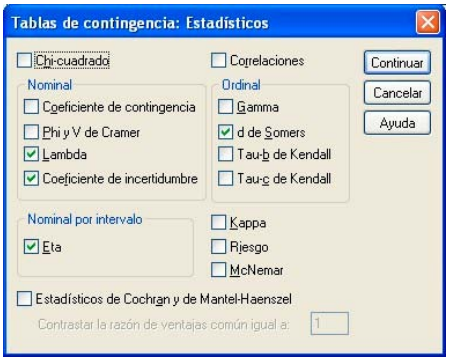
- Tau-b de Kendall:

$$\tau_b = \frac{(C-D)}{\sqrt{(C+D+E_X)(C+D+E_Y)}} = \frac{(2020-1906)}{\sqrt{(2020+1906+2078)(2020+1906+1338)}} = 0{,}0203$$

- Tau-c de Kendall:

$$\tau_c = \frac{2m(C-D)}{N^2(m-1)} = \frac{2.3(2020-1906)}{127^2.\,2} = 0{,}021 \quad \text{donde } m = \text{mín}\{\text{nº filas, nº columnas}\}$$

En el análisis de **MEDIDAS DIRECCIONALES** se encuentran las *medidas nominales (lambda, coeficiente de incertidumbre), medidas ordinales (d de Somers), y el nominal por intervalo (eta).*



Medidas direccionales

			Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Nominal por nominal	Lambda	Simétrica	,000	,041	,000	1,000
		Meses dependiente	,000	,000	.[c]	.[c]
		Empaquetados dependiente	,000	,088	,000	1,000
	Tau de Goodman y Kruskal	Meses dependiente	,003	,005		,978[d]
		Empaquetados dependiente	,005	,009		,969[d]
	Coeficiente de incertidumbre	Simétrica	,004	,007	,574	,973[e]
		Meses dependiente	,004	,006	,574	,973[e]
		Empaquetados dependiente	,005	,008	,574	,973[e]
Ordinal por ordinal	d de Somers	Simétrica	,020	,075	,271	,786
		Meses dependiente	,022	,080	,271	,786
		Empaquetados dependiente	,019	,070	,271	,786
Nominal por intervalo	Eta	Meses dependiente	,039			
		Empaquetados dependiente	,035			

a. Asumiendo la hipótesis alternativa.
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.
e. Probabilidad del chi-cuadrado de la razón de verosimilitudes.

El valor de los coeficientes aparece acompañado de su correspondiente nivel crítico (Sig. aproximada), puesto que el nivel crítico de todas las medidas listadas es grande $(>0{,}05)$ *se acepta la hipótesis nula de independencia, concluyendo que los meses y el método de empaquetado no están relacionados.*

Meses	A	B	C	Total marginal	
1	6	10	10	$N_{1\bullet}=26$	$\text{máx } n_{1j}=10$
2	8	12	12	$N_{2\bullet}=32$	$\text{máx } n_{2j}=12$
3	8	8	14	$N_{3\bullet}=30$	$\text{máx } n_{3j}=14$
4	9	14	16	$\mathbf{N_{4\bullet}=39}$	$\text{máx } n_{4j}=16$
Total marginal	$N_{\bullet 1}=31$	$N_{\bullet 2}=44$	$\mathbf{N_{\bullet 3}=52}$	$N=127$	$\mathbf{\sum_{i=1}^{4}\text{máx}_j\, n_{ij}=52}$
$\mathbf{\sum_{j=1}^{3}\text{máx}_i\, n_{ij}=39}$	$\text{máx } n_{i1}=9$	$\text{máx } n_{i2}=14$	$\text{máx } n_{i3}=16$		

- Coeficiente Lambda:

$$\lambda=\frac{\sum_i \text{máx}_j\, n_{ij}+\sum_j \text{máx}_i\, n_{ij}-\text{máx}_i\, N_{i\bullet}-\text{máx}_j\, N_{\bullet j}}{2N-\text{máx}_i\, N_{i\bullet}-\text{máx}_j\, N_{\bullet j}}=\frac{52+39-39-52}{2.127-39-52}=0 \quad \textit{coeficiente lambda}$$

En consecuencia, las variables analizadas son independientes al ser $\lambda=0$

- Coeficiente de Goodman y Kruskall (*variable X dependiente*):

$$\tau = \frac{\frac{1}{N}\sum_i (N - N_{i\bullet})N_{i\bullet} - \sum_j \frac{(N_{\bullet j} - n_{ij})n_{ij}}{N_{\bullet j}}}{\frac{1}{N}\sum_i (N - N_{i\bullet})N_{i\bullet}} = \frac{94{,}551 - 94{,}2578}{94{,}2578} = 0{,}003 \quad \textit{tau de Goodman y Kruskall}$$

$$\frac{1}{N}\sum_{i=1}^{4}(N - N_{i\bullet})N = \frac{1}{127}\left[(127-26)26 + (127-32)32 + (127-30)30 + (127-39)39\right] = 94{,}551$$

$$\sum_j \frac{(N_{\bullet j} - n_{ij})n_{ij}}{N_{\bullet j}} = \left[\frac{(31-6)6 + (31-8)8 + (31-8)8 + (31-9)9}{31}\right] +$$
$$+ \left[\frac{(44-10)10 + (44-12)12 + (44-8)8 + (44-14)14}{44}\right] +$$
$$+ \left[\frac{(52-10)10 + (52-12)12 + (52-14)14 + (52-16)16}{52}\right] = 88{,}322$$

- Coeficiente de Goodman y Kruskall (variable Y dependiente):

$$\tau = \frac{\frac{1}{N}\sum_j (N - N_{\bullet j})N_{\bullet j} - \sum_i \frac{(N_{i\bullet} - n_{ij})n_{ij}}{N_{i\bullet}}}{\frac{1}{N}\sum_j (N - N_{\bullet j})N_{\bullet j}} = \frac{82{,}898 - 82{,}456}{82{,}898} = 0{,}005$$

$$\frac{1}{N}\sum_{j=1}^{3}(N - N_{\bullet j})N = \frac{1}{127}\left[(127-31)31 + (127-44)44 + (127-52)52\right] = 82{,}898$$

$$\sum_i \frac{(N_{i\bullet} - n_{ij})n_{ij}}{N_{i\bullet}} = \left[\frac{(26-6)6 + (26-10)10 + (26-10)10}{26}\right] + \left[\frac{(32-8)8 + (32-12)12 + (32-12)12}{32}\right] +$$
$$+ \left[\frac{(30-8)8 + (30-8)8 + (30-14)14}{30}\right] + \left[\frac{(39-9)9 + (39-14)14 + (39-16)16}{39}\right] = 82{,}456$$

- Coeficiente de Incertidumbre: $I_{Y/X} = \dfrac{I(X) + I(Y) - I(XY)}{I(Y)}$

$N_{i\bullet}$	$N_{i\bullet}/N$	$Ln(N_{i\bullet}/N)$	$(N_{i\bullet}/N)\,Ln(N_{i\bullet}/N)$
26	0,2047	-1,5861	-0,3247
32	0,2520	-1,3785	-0,3473
30	0,2362	-1,4430	-0,3409
39	0,3071	-1,1806	-0,3626
$N = 127$			-1,3755

$N_{\bullet j}$	$N_{\bullet j}/N$	$Ln(N_{\bullet j}/N)$	$(N_{\bullet j}/N)\,Ln(N_{\bullet j}/N)$
31	31	0,2441	-1,4102
44	44	0,3465	-1,0600
52	52	0,4094	-0,8929
$N = 127$			-1,0771

$$I(X) = -\sum_i \left(\frac{N_{i\bullet}}{N}\right) Ln\left(\frac{N_{i\bullet}}{N}\right) = 1{,}3755 \qquad I(Y) = -\sum_j \left(\frac{N_{\bullet j}}{N}\right) Ln\left(\frac{N_{\bullet j}}{N}\right) = 1{,}0771$$

(n_{ij}/N)			$Ln\,(n_{ij}/N)$			$(n_{ij}/N)\ Ln\,(n_{ij}/N)$		
0,0472	0,079	0,079	-3,0524	-2,5416	-2,5416	-0,1442	-0,2001	-0,2001
0,0630	0,094	0,094	-2,7647	-2,3593	-2,3593	-0,1742	-0,2229	-0,2229
0,0630	0,063	0,110	-2,7647	-2,7647	-2,2051	-0,1742	-0,1742	-0,2431
0,0709	0,110	0,126	-2,6470	-2,2051	-2,0716	-0,1876	-0,2431	-0,2610
						$\sum_i \sum_j (n_{ij}/N)\,Ln(n_{ij}/N) = -2{,}4475$		

$$I(XY) = -\sum_i \sum_j \left(\frac{n_{ij}}{N}\right) Ln\left(\frac{n_{ij}}{N}\right) = 2{,}4475$$

$$I_{Y/X} = 2\,\frac{I(X)+I(Y)-I(XY)}{I(X)+I(Y)} = 2\,\frac{1{,}3755+1{,}0771-2{,}4475}{1{,}3755+1{,}0771} = 0{,}00416 \quad \textit{(simétrica)}$$

Para obtener $I_{X/Y}$ basta intercambiar los papeles de I(X), I(Y).

- d de Sommer (simétrica):

$$d = \frac{(C-D)}{C+D+\left[\frac{E_X+E_Y}{2}\right]} = \frac{(2020-1906)}{2020+1906+\left[\frac{2078+1338}{2}\right]} = 0{,}020$$

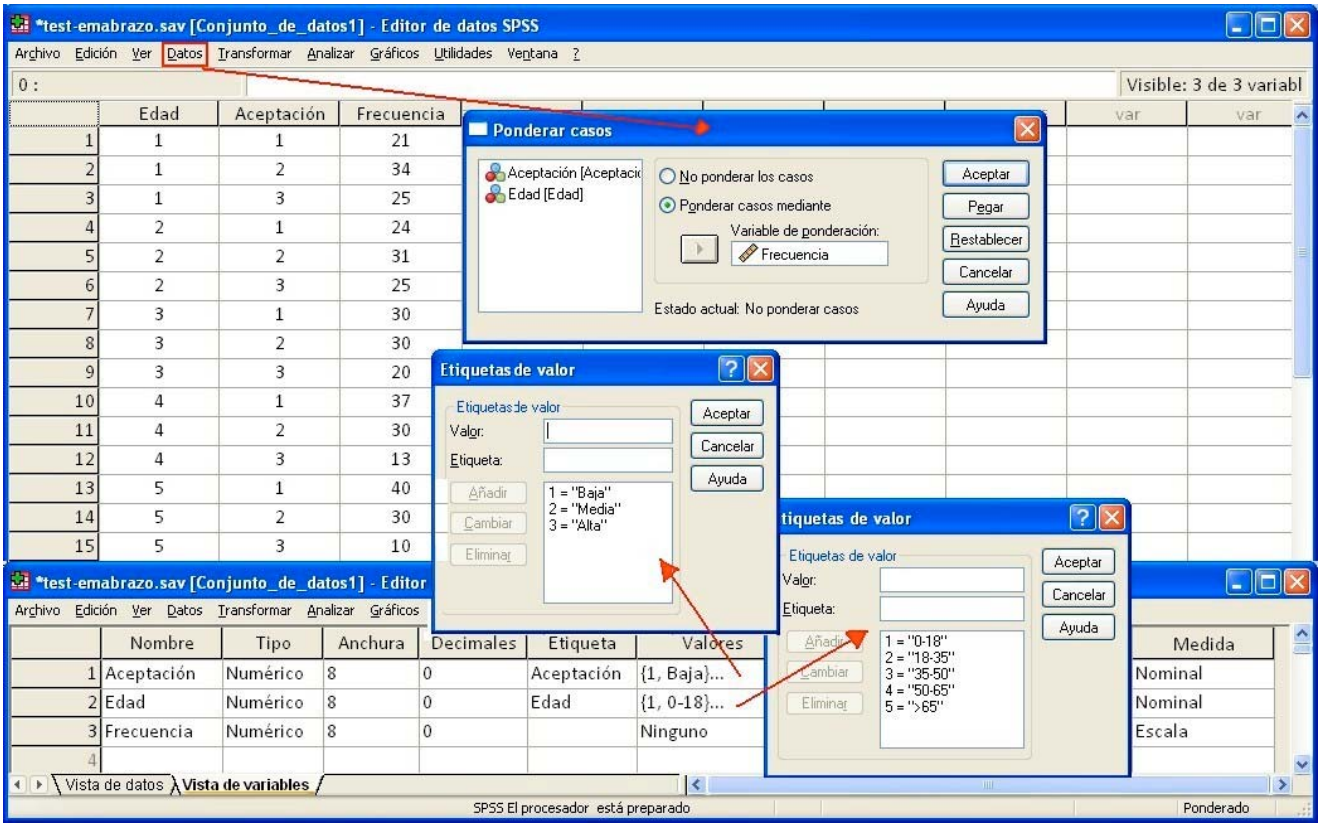
Variable Y como independiente: $d_X = \frac{(C-D)}{C+D+E_X} = \frac{(2020-1906)}{2020+1906+2078} = 0{,}019$

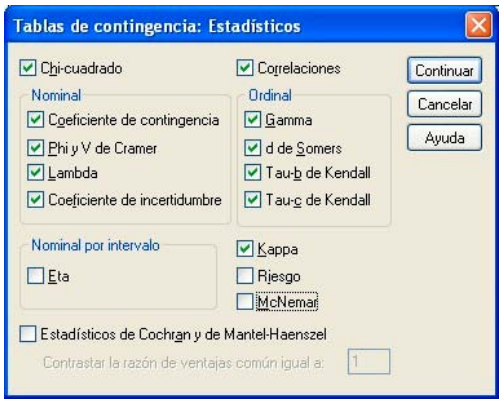
Variable X como independiente: $d_Y = \frac{(C-D)}{C+D+E_Y} = \frac{(2020-1906)}{2020+1906+1338} = 0{,}022$

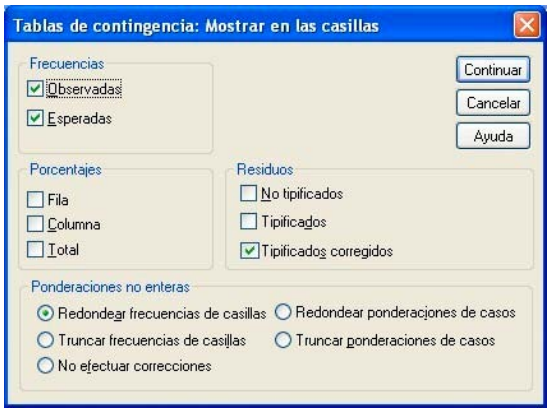
Ejercicio 2.- Se quiere estudiar la relación entre la edad de las mujeres y su aceptación de una ley sobre interrupción del embarazo. Para ello se ha llevado a cabo una encuesta sobre 400 mujeres cuyos resultados se adjuntan en la tabla:

	Aceptación		
Edad	**Baja**	**Media**	**Alta**
0-18	21	34	25
18-35	24	31	25
35-50	30	30	20
50-65	37	30	13
> 65	40	30	10

Con el *Editor de datos* de SPSS:







En el *Visor de resultados* de SPSS:

Tabla de contingencia Edad * Aceptación

			Aceptación			Total
			Baja	Media	Alta	
Edad	0-18	Recuento	21	34	25	80
		Frecuencia esperada	30,4	31,0	18,6	80,0
		Residuos corregidos	-2,4	,8	1,9	
	18-35	Recuento	24	31	25	80
		Frecuencia esperada	30,4	31,0	18,6	80,0
		Residuos corregidos	-1,6	,0	1,9	
	35-50	Recuento	30	30	20	80
		Frecuencia esperada	30,4	31,0	18,6	80,0
		Residuos corregidos	-,1	-,3	,4	
	50-65	Recuento	37	30	13	80
		Frecuencia esperada	30,4	31,0	18,6	80,0
		Residuos corregidos	1,7	-,3	-1,7	
	>65	Recuento	40	30	10	80
		Frecuencia esperada	30,4	31,0	18,6	80,0
		Residuos corregidos	2,5	-,3	-2,5	
Total		Recuento	152	155	93	400
		Frecuencia esperada	152,0	155,0	93,0	400,0

Residuos corregidos	Aceptación		
	Baja	Media	Alta
0-18	-2,4	,8	1,9
18-35	-1,6	,0	1,9
35-50	-,1	-,3	,4
50-65	1,7	-,3	-1,7
>65	2,5	-,3	-2,5

Comparando los valores absolutos de los residuos tipificados corregidos con el correspondiente valor tabular de la normal, para un nivel de significación del 5% (> 1,96), se observa que muchos residuos no son significativos.

Analizando estos valores, tanto en sus magnitudes como en sus rangos, resulta el patrón: *<< Las jóvenes (menores de 18 años) de clase baja y las mayores de 65 años de clase alta tienen una opinión favorable sobre la interrupción del embarazo. Por el contrario, la opinión de las mayores de 65 años de clase baja tienen una percepción claramente negativa. De este modo, se evidencia que existe una relación y del tipo que es ésta>>.*

Subrayar que este método supone un análisis celda a celda. Esta diferencia se encuentra que mientras el contraste usual trabaja con $[(5-1)(3-1)=8]$ elementos independientes, el contraste por cada celda implica que la totalidad de los residuos tipificados d_{ij} son independientes y cada uno de ellos se ajusta a una distribución teórica $N(0,1)$.

El valor del *estadístico de contraste* (observado) es $19{,}283$, el cual, en la distribución χ^2 tiene 8 grados de libertad (*gl = 8*), tiene asociada una probabilidad (*Significación asintótica*) de $0{,}013$.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	19,283[a]	8	,013
Razón de verosimilitudes	19,945	8	,011
Asociación lineal por lineal	18,255	1	,000
N de casos válidos	400		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 18,60.

Puesto que esta probabilidad (denominada *nivel crítico* o *nivel de significación observada*) es pequeña (menor que 0,05) , se decide rechazar la hipótesis nula, indicando que hay evidencia de asociación entre el grado de aceptación del aborto y la edad de las mujeres.

Señalar que el valor de la *razón de verosimilitudes* (RV) es 19,945, tiene asociada una probabilidad (*Sig. asintótica*) de $0{,}011$, que como es menor que 0,05, indicando que hay evidencia de asociación entre el grado de aceptación del aborto y la edad de las mujeres.

Los estadísticos (χ^2, RV) llevan a la misma conclusión, en caso contrario, se elige el estadístico con *menor Sig. asintótica*.

El valor del estadístico *Asociación lineal por lineal* (corrección por continuidad de Yates) tiene un valor de $18{,}255$ con un nivel crítico de $(0<0{,}05)$, por lo que se rechaza la hipótesis nula de independencia, llegando a la misma conclusión que con los estadísticos anteriores.

Medidas simétricas

		Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Nominal por nominal	Phi	,220			,013
	V de Cramer	,155			,013
	Coeficiente de contingencia	,214			,013
Ordinal por ordinal	Tau-b de Kendall	-,180	,040	-4,485	,000
	Tau-c de Kendall	-,195	,043	-4,485	,000
	Gamma	-,248	,054	-4,485	,000
	Correlación de Spearman	-,213	,047	-4,358	,000[c]
Intervalo por intervalo	R de Pearson	-,214	,047	-4,368	,000[c]
N de casos válidos		400			

a. Asumiendo la hipótesis alternativa.

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

c. Basada en la aproximación normal.

El valor de cada coeficiente aparece acompañado de su correspondiente nivel crítico (*Sig. aproximada*), que permite tomar una decisión sobre la hipótesis nula de independencia. Puesto que estos niveles críticos son menores que 0,05, se puede afirmar que hay relación entre la aceptación del aborto y la edad de las mujeres.

Por su parte, los valores obtenidos del Coeficiente de contingencia y V de Cramer (*como medidas nominales cuantifican el grado de asociación*) indican una asociación baja entre la edad de las mujeres y la aceptación del aborto.

De otra parte, los valores obtenidos de la Tau-b de Kendall, Tau-c de Kendall, Gamma y Correlación de Spearman (*como medidas ordinales indican además el tipo de asociación*) presentan una asociación baja negativa, es decir, que el grado de aceptación del aborto disminuye al aumentar la edad.

Medidas direccionales

			Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Nominal por nominal	Lambda	Simétrica	,064	,032	1,968	,049
		Edad dependiente	,059	,024	2,451	,014
		Aceptación dependiente	,069	,055	1,213	,225
	Tau de Goodman y Kruskal	Edad dependiente	,012	,005		,014[c]
		Aceptación dependiente	,022	,010		,022[c]
	Coeficiente de incertidumbre	Simétrica	,019	,008	2,300	,011[d]
		Edad dependiente	,015	,007	2,300	,011[d]
		Aceptación dependiente	,023	,010	2,300	,011[d]
Ordinal por ordinal	d de Somers	Simétrica	-,179	,040	-4,485	,000
		Edad dependiente	-,199	,044	-4,485	,000
		Aceptación dependiente	-,162	,036	-4,485	,000

a. Asumiendo la hipótesis alternativa.

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

c. Basado en la aproximación chi-cuadrado.

d. Probabilidad del chi-cuadrado de la razón de verosimilitudes.

Los valores obtenidos de Lambda, Tau de Goodman y Kruskall, Coeficiente de incertidumbre, y d de Somers (*como medidas nominales cuantifican el grado de asociación*) indican una asociación baja entre la edad de las mujeres y la aceptación del aborto.

Cada medida acompañada de un nivel crítico (Sig. aproximada), que en los casos que es menor que 0,05, conduce a rechazar la hipótesis nula de independencia y concluir que las variables (edad de las mujeres, aceptación del aborto) están asociadas.

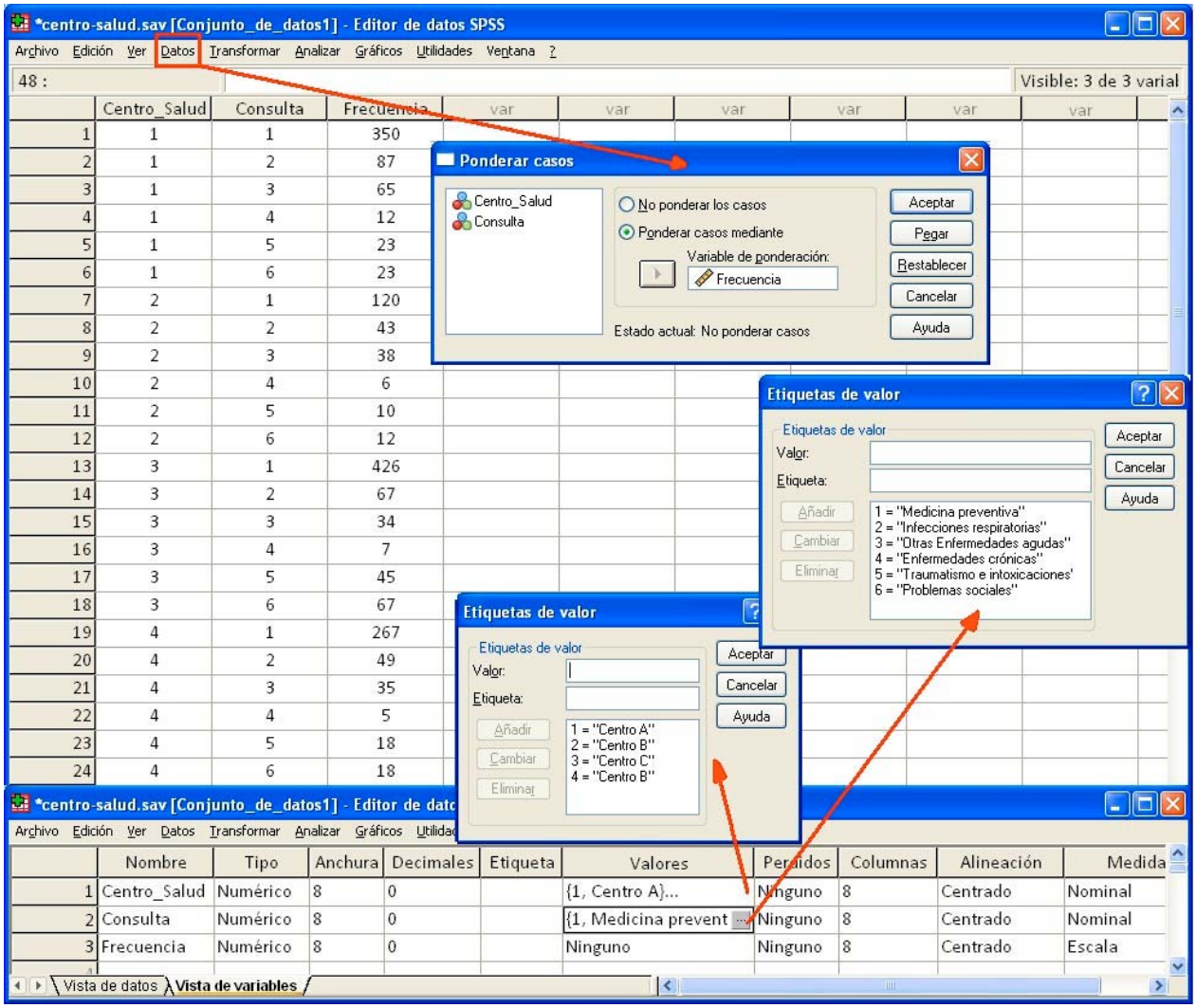
El valor 0,012 del coeficiente Tau de Goodman y Kruskall calculado considera la variable "Aceptación del aborto" como independiente, tiene la interpretación:
<<Conociendo la edad de la mujer consultada (filas), se reduce en un 1,2% la probabilidad de cometer un error al predecir su aceptación al aborto (columnas). Esto significa que la edad de la mujer no tiene capacidad predictiva sobre la aceptación del aborto>>.

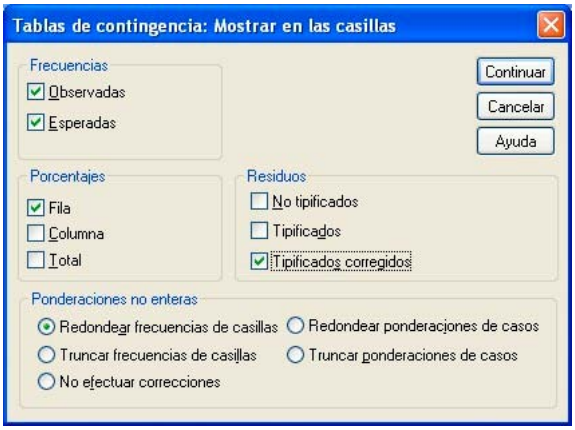
Se concluye, existe evidencia de asociación entre el grado de aceptación del aborto y la edad de las mujeres, disminuyendo el grado de aceptación al aumentar la edad.

Ejercicio 3.- Para analizar si la distribución de los motivos de consulta en 4 centros de atención ambulatoria pediátrica eran similares, se clasificaron las consultas en 6 grupos: (1) Medicina preventiva; (2) Infecciones respiratorias altas; (3) Otras enfermedades agudas; (4) Enfermedades crónicas; (5)Traumatismos e intoxicaciones; y (6) Problemas sociales.

Centro Médico	*Grupo motivo de consulta*						
	1	2	3	4	5	6	Total
A	350	87	65	12	23	23	560
B	120	43	38	6	10	12	229
C	426	67	34	7	45	67	646
D	267	49	35	5	18	18	392
Total	1163	246	172	30	96	120	1827

Con el Editor de SPSS:





Se inicia el análisis *celda* a *celda* con el botón [Casillas] para comprobar el uso adecuado de la chi-cuadrado, así como para obtener un patrón en la distribución observada.

Tabla de contingencia Centro_Salud * Consulta_médica

			Consulta_médica						Total
			Medicina preventiva	Infecciones respiratorias	Otras Enfermedades agudas	Enfermedades crónicas	Traumatismo e intoxicaciones	Problemas sociales	
Centro_Salud	Centro A	Recuento	350	87	65	12	23	23	560
		Frecuencia esperada	356,5	75,4	52,7	9,2	29,4	36,8	560,0
		% de Centro_Salud	62,5%	15,5%	11,6%	2,1%	4,1%	4,1%	100,0%
		Residuos corregidos	-,7	1,7	2,1	1,1	-1,5	-2,8	
	Centro B	Recuento	120	43	38	6	10	12	229
		Frecuencia esperada	145,8	30,8	21,6	3,8	12,0	15,0	229,0
		% de Centro_Salud	52,4%	18,8%	16,6%	2,6%	4,4%	5,2%	100,0%
		Residuos corregidos	-3,8	2,5	4,0	1,2	-,6	-,9	
	Centro C	Recuento	426	67	34	7	45	67	646
		Frecuencia esperada	411,2	87,0	60,8	10,6	33,9	42,4	646,0
		% de Centro_Salud	65,9%	10,4%	5,3%	1,1%	7,0%	10,4%	100,0%
		Residuos corregidos	1,5	-2,9	-4,5	-1,4	2,4	4,9	
	Centro B	Recuento	267	49	35	5	18	18	392
		Frecuencia esperada	249,5	52,8	36,9	6,4	20,6	25,7	392,0
		% de Centro_Salud	68,1%	12,5%	8,9%	1,3%	4,6%	4,6%	100,0%
		Residuos corregidos	2,1	-,6	-,4	-,6	-,7	-1,8	
Total		Recuento	1163	246	172	30	96	120	1827
		Frecuencia esperada	1163,0	246,0	172,0	30,0	96,0	120,0	1827,0
		% de Centro_Salud	63,7%	13,5%	9,4%	1,6%	5,3%	6,6%	100,0%

El porcentaje de celdas, con una frecuencia esperada menor que 5, está vinculado a las exigencias para la utilización de la prueba Chi-cuadrado: (a) *Menos de un 20% de celdas con frecuencia esperada menor que 5.* (b) *Ninguna celda con frecuencia esperada menor que 1.*

Comparando los valores absolutos de los residuos tipificados corregidos con el correspondiente valor tabular de la normal, para un nivel de significación del 5% (> 1,96), se observa que algunos residuos son significativos.

Analizando estos valores, tanto en sus magnitudes como en sus rangos, resulta el patrón:
<< La distribución observada entre los centros médicos no es uniforme, en el centro C los problemas sociales (4,9) fueron el mayor motivo de consulta médica; mientras que, en el centro B otras enfermedades agudas (4) fueron el mayor motivo de consulta>>.

Subrayar que este método supone un análisis celda a celda. Esta diferencia se encuentra que mientras el contraste usual trabaja con $[(4-1)(6-1)=15]$ elementos independientes, el contraste por cada celda implica que la totalidad de los residuos tipificados d_{ij} son independientes y cada uno de ellos se ajusta a una distribución teórica $N(0,1)$.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	76,944[a]	15	,000
Razón de verosimilitudes	75,422	15	,000
Asociación lineal por lineal	,021	1	,886
N de casos válidos	1827		

a. 1 casillas (4,2%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 3,76.

El valor del *estadístico de contraste* (observado) es $76{,}944$, el cual, en la distribución χ^2 tiene 15 grados de libertad (*gl = 15*), tiene asociada una probabilidad (*Significación asintótica*) de $0{,}000$.

Puesto que esta probabilidad (denominada *nivel crítico* o *nivel de significación observada*) es pequeña (menor que 0,05), se decide rechazar la hipótesis nula, indicando que hay evidencia de asociación entre el centro médico y el motivo de las consultas.

Señalar que el valor de la *razón de verosimilitudes* (RV) es 75,422, tiene asociada una probabilidad (*Sig. asintótica*) de $0{,}000$, que como es menor que 0,05, indica que hay evidencia de asociación entre el grado de aceptación entre el centro médico y el motivo de las consultas.

Los estadísticos (χ^2, RV) llevan a la misma conclusión, en caso contrario, se elige el estadístico con *menor Sig. asintótica*.

El valor del estadístico *Asociación lineal por lineal* (*corrección por continuidad de Yates, cuando se aproxima una variable discreta a una variable continua, se aplica a la prueba de la chi-cuadrado cuando la frecuencia en algunas observaciones es menor que 10*) tiene un valor de $0{,}021$ con un nivel crítico $(0{,}886 > 0{,}05)$, por lo que se acepta la hipótesis nula de independencia. No existe un consenso generalizado sobre la utilización de esta corrección.

Medidas simétricas

		Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Nominal por nominal	Phi	,205			,000
	V de Cramer	,118			,000
	Coeficiente de contingencia	,201			,000
Ordinal por ordinal	Tau-b de Kendall	-,034	,019	-1,768	,077
	Tau-c de Kendall	-,029	,016	-1,768	,077
	Gamma	-,054	,030	-1,768	,077
	Correlación de Spearman	-,041	,023	-1,759	,079[c]
Intervalo por intervalo	R de Pearson	,003	,022	,144	,886[c]
N de casos válidos		1827			

a. Asumiendo la hipótesis alternativa.

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

c. Basada en la aproximación normal.

El valor de cada coeficiente aparece acompañado de su correspondiente nivel crítico (*Sig. aproximada*), que permite tomar una decisión sobre la hipótesis nula de independencia. Puesto que estos niveles críticos son menores que 0,05, se puede afirmar que hay asociación entre el centro médico y el motivo de las consultas

Por su parte, los valores obtenidos del Coeficiente de contingencia y V de Cramer (*como medidas nominales cuantifican el grado de asociación*) indican una asociación baja entre el centro médico y el motivo de la consulta.

De otra parte, los valores obtenidos de la Tau-b de Kendall, Tau-c de Kendall, Gamma y Correlación de Spearman (*como medidas ordinales indican además el tipo de asociación*) presentan una asociación baja negativa.

Medidas direccionales

			Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Nominal por nominal	Lambda	Simétrica	,030	,009	3,406	,001
		Centro_Salud dependiente	,047	,014	3,406	,001
		Consulta_médica dependiente	,000	,000	.[c]	.[c]
	Tau de Goodman y Kruskal	Centro_Salud dependiente	,016	,004		,000[d]
		Consulta_médica dependiente	,010	,003		,000[d]
	Coeficiente de incertidumbre	Simétrica	,017	,004	4,356	,000[e]
		Centro_Salud dependiente	,016	,004	4,356	,000[e]
		Consulta_médica dependiente	,017	,004	4,356	,000[e]
Ordinal por ordinal	d de Somers	Simétrica	-,034	,019	-1,768	,077
		Centro_Salud dependiente	-,039	,022	-1,768	,077
		Consulta_médica dependiente	-,030	,017	-1,768	,077

a. Asumiendo la hipótesis alternativa.

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.

e. Probabilidad del chi-cuadrado de la razón de verosimilitudes.

Los valores obtenidos de Lambda, Tau de Goodman y Kruskall, Coeficiente de incertidumbre, y d de Somers (*como medidas nominales cuantifican el grado de asociación, la capacidad de hacer pronósticos de una variable respecto de la otra*) indican una asociación baja entre el centro médico y el motivo de la consulta, es decir, la capacidad de hacer pronósticos de una variable respecto de la otra es realmente escasa.

Cada medida acompañada de un nivel crítico (Sig. aproximada), en los casos que es menor que 0,05, conduce a rechazar la hipótesis nula de independencia y concluir que las variables (*centro médico*, *motivo de la consulta*) están asociadas.

El valor 0,016 del coeficiente Tau de Goodman y Kruskall calculado considera la variable "Centro médico" como independiente, tiene la interpretación:
<<Conociendo el centro médico (filas), se reduce en un 1,6% la probabilidad de cometer un error al predecir donde se hizo la consulta (columnas). Esto significa que el centro médico no tiene capacidad predictiva sobre el motivo de la consulta>>.

En definitiva, se puede concluir que el centro médico y el motivo de las consultas están relacionados, pero en ningún caso se podría considerar un claro factor de pronóstico sobre las consultas.

(Estudio de Casos y Controles)

Ejercicio 4.- Para analizar si la lactancia constituye un factor de protección para el cáncer de mama, un estudio incluyó a 755 mujeres menores de 35 años de todas las comunidades españolas, a las que se diagnóstico cáncer de mama durante el período 2000-2005. Los controles tenían una diferencia de edad con los casos inferior a seis meses.
Cada caso y control fueron controlados por el mismo investigador. Los resultados reflejan que en el grupo de casos, 255 mujeres realizaron una lactancia plena de al menos 3 meses, mientras que entre los controles este antecedente estaba presente en 487 mujeres (de los 255 controles de los casos que tuvieron una lactancia plena, 160 lactaron y 95 no, en tanto de los 500 controles de los casos que no lactaron, 327 si lo habían hecho y 173 no).
Los datos quedan reflejados en las dos tablas siguientes:

Casos y Controles

	Casos	Controles	Total
Expuesto	255	487	742
No expuesto	500	268	768
Total	755	755	1.510

$$OR = \frac{255.268}{500.487} = 0{,}281 \quad RR = \frac{255.768}{500.742} = 0{,}528$$

Casos y Controles emparejados

	Controles		
Casos	Expuesto	No expuesto	Total
Expuesto	160	95	255
No expuesto	327	173	500
Total	487	268	755

$$OR = \frac{160.173}{327.95} = 0{,}891 \quad RR = \frac{160.500}{327.255} = 0{,}959$$

☞ En los diseños *longitudinales hacia delante*, conocidos como *diseños prospectivos o de cohortes*, las mujeres son clasificadas en dos grupos dependiendo de la presencia o ausencia de lactancia y se les hace un seguimiento durante un período de tiempo hasta determinar la proporción de mujeres de cada grupo en los que se da un determinado desenlace (cáncer de mama).

La medida de interés es el riesgo relativo (RR): $RR = \frac{n_{11}.N_{2\bullet}}{n_{21}.N_{1\bullet}}$

La interpretación es: <<La proporción de cáncer de mama entre las mujeres expuestas es RR veces más alta que entre las mujeres no expuestas>>. En este sentido, en casos y controles, la proporción de cáncer de mama de mujeres expuestas es 0,528 más alta que entre las mujeres no expuestas. En el emparejamiento de casos y controles, la proporción aumenta hasta 0,959.

Un riesgo relativo de 1 indica que la probabilidad de encontrar cáncer de mama es la misma tanto en el grupo de mujeres expuestas como en el grupo de mujeres no expuestas. Para valorar si el riesgo obtenido es significativamente distinto de 1, se calcula el intervalo de confianza:

$$IC_{RR} = \left[RR.e^{-z_{\alpha/2}.\sqrt{\frac{n_{12}}{n_{11}.N_{1\bullet}} + \frac{n_{22}}{n_{21}.N_{2\bullet}}}} \;,\; RR.e^{z_{\alpha/2}.\sqrt{\frac{n_{12}}{n_{11}.N_{1\bullet}} + \frac{n_{22}}{n_{21}.N_{2\bullet}}}} \right]$$

$$IC_{RR} = \left[0{,}528.e^{-1{,}96.\sqrt{\frac{487}{255.742} + \frac{268}{500.768}}} \;,\; 0{,}528.e^{1{,}96.\sqrt{\frac{487}{255.742} + \frac{268}{500.768}}} \right] = [0{,}472 \;,\; 0{,}590]_{\text{casos y controles}}$$

$$IC_{RR} = \left[0{,}959 \,.\, e^{-1{,}96 \cdot \sqrt{\frac{95}{160.255} + \frac{173}{327.500}}} \;,\; 0{,}959 \,.\, e^{1{,}96 \cdot \sqrt{\frac{95}{160.255} + \frac{173}{327.500}}} \right] = [0{,}856 \;,\; 1{,}075]_{emparejados}$$

Si el intervalo de confianza *no contiene* el 1, se concluye que el riesgo de tener cáncer de mama no es lo mismo entre las mujeres expuestas y no expuestas a la lactancia.

☞ En los diseños *longitudinales hacia atrás*, llamados *diseños retrospectivos o de caso-control*, se forman grupos de mujeres (lactaron y no lactaron) a partir de la presencia o ausencia de cáncer de mama y se hace un seguimiento hacia atrás intentando encontrar información sobre la proporción en la que se encuentra presente en cada muestra el cáncer de mama.

Puesto que el tamaño de los grupos (lactaron y no lactaron) se fija a partir de la presencia o ausencia del cáncer de mama, se calcula odds-ratio (*razón de ventajas o razón de productos cruzados o en qué medida que lactaron es un riesgo de tener cáncer de mama*):

$$OR = \frac{n_{11} \,.\, n_{22}}{n_{12} \,.\, n_{21}}$$

Como se observa, el odds-ratio (OR) es tanto mejor estimador del riesgo relativo cuanto más pequeñas sean las proporciones de desenlace en cada grupo.

Un índice de 1 indica que la probabilidad de encontrarse con el cáncer de mama en los grupos estudiados es la misma. Para determinar si este riesgo es significativamente distinto de 1, se calcula el intervalo de confianza:

$$IC_{OR} = \left[OR \,.\, e^{-z_{\alpha/2} \cdot \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}} \;,\; OR \,.\, e^{z_{\alpha/2} \cdot \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}} \right]$$

$$IC_{OR} = \left[0{,}281 \,.\, e^{-1{,}96 \cdot \sqrt{\frac{1}{255} + \frac{1}{487} + \frac{1}{500} + \frac{1}{268}}} \;,\; 0{,}281 \,.\, e^{1{,}96 \cdot \sqrt{\frac{1}{255} + \frac{1}{487} + \frac{1}{500} + \frac{1}{268}}} \right] = [0{,}227 \;,\; 0{,}347]_{caso\ y\ control}$$

$$IC_{OR} = \left[0{,}891 \,.\, e^{-1{,}96 \cdot \sqrt{\frac{1}{160} + \frac{1}{95} + \frac{1}{327} + \frac{1}{173}}} \;,\; 0{,}891 \,.\, e^{1{,}96 \cdot \sqrt{\frac{1}{160} + \frac{1}{95} + \frac{1}{327} + \frac{1}{173}}} \right] = [0{,}651 \;,\; 1{,}219]_{emparejados}$$

Preparando datos de variables dicotómicas para introducir en SPSS:

		Controles	
Casos	Casos	Expuesto (1)	No expuesto (2)
Expuesto (1)	Expuesto (1)	164	40
	No expuesto (2)	32	64
No expuesto (2)	Expuesto (1)	41	10
	No expuesto (2)	48	96

NOTA: En el caso de regresión logística la codificación de (2) hubiera sido (0) para no expuesto

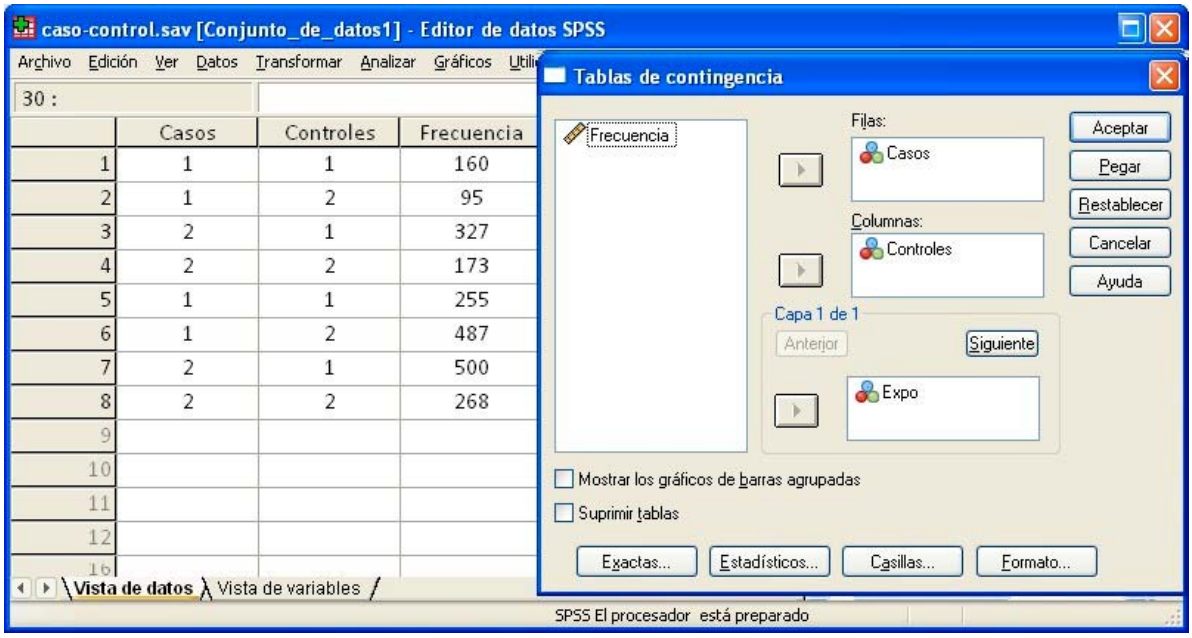


Tabla de contingencia Casos * Controles * Expo

Estadísticos | Recuento

Expo			Controles: Expuesto	Controles: No Expuesto	Total
Expo	Casos	Expo	160	95	255
		No Expo	327	173	500
	Total		487	268	755
No Expo	Casos	Expo	255	487	742
		No Expo	500	268	768
	Total		755	755	1510

Haciendo doble click sobre la tabla, en el menú *Pivotar/Paneles de Pivotado*

Tabla de contingencia Casos * Controles * Expo

Expo | No Expo

			Controles: Expuesto	Controles: No Expuesto	Total
Casos	Expo	Recuento	255	487	742
	No Expo	Recuento	500	268	768
Total		Recuento	755	755	1510

Pulsando, a la izquierda, en Casos, van apareciendo las dos tablas introducidas.

El hecho de que la tabla no muestre el valor del estadístico de McNemar significa que el nivel crítico se ha calculado utilizando la distribución binomial (obteniendo la probabilidad exacta en lugar de aproximada).

Pruebas de chi-cuadrado

Expo		Valor	Sig. exacta (bilateral)
Expo	Prueba de McNemar		,000[a]
	N de casos válidos	755	
No Expo	Prueba de McNemar		,703[a]
	N de casos válidos	1510	

a. Utilizada la distribución binomial

Cualquiera que sea la forma de obtenerlo, el nivel crítico indica el grado de compatibilidad existente entre los datos muestrales y la hipótesis nula de igualdad de proporciones antes-después.

En el caso de los expuestos, como p_value < 0,05, se rechaza la hipótesis nula y se concluye que se puede descartar el azar como explicación del cáncer de mama con una pequeña probabilidad (inferior al 1:10.000) de cometer el error tipo I (rechazo indebido).

En el caso de los no expuestos, p_value > 0,05, se acepta la hipótesis nula, no habiendo asociación alguna con el cáncer de mama.

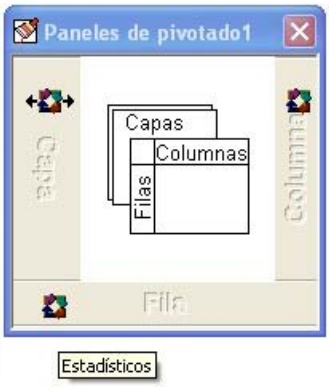
Estimación de riesgo

Expo		Valor	Intervalo de confianza al 95% Inferior	Superior
Expo	Razón de las ventajas para Casos (Expo / No Expo)	,891	,651	1,219
	Para la cohorte Controles = Expuesto	,959	,856	1,075
	Para la cohorte Controles = No Expuesto	1,077	,882	1,315
	N de casos válidos	755		
No Expo	Razón de las ventajas para Casos (Expo / No Expo)	,281	,227	,347
	Para la cohorte Controles = Expuesto	,528	,472	,590
	Para la cohorte Controles = No Expuesto	1,881	1,685	2,099
	N de casos válidos	1510		

Estimación de riesgo

Expo | No Expo

	Valor	Intervalo de confianza al 95% Inferior	Superior
Razón de las ventajas para Casos (Expo / No Expo)	,281	,227	,347
Para la cohorte Controles = Expuesto	,528	,472	,590
Para la cohorte Controles = No Expuesto	1,881	1,685	2,099
N de casos válidos	1510		



Observando la segunda fila, para la cohorte *Controles=Expuesto*, 'La proporción de cáncer de mama entre las mujeres expuestas es RR=0,528 veces más alta que entre las mujeres no expuestas' En el emparejamiento de casos y controles, la proporción aumenta hasta 0,959.

Para valorar si el riesgo obtenido es significativo, se calcula el intervalo de confianza. En Casos y Controles si lo es, mientras que en el Emparejamiento de Casos-Controles no lo es, su intervalo de confianza cubre el 1, indicando que la probabilidad de encontrar cáncer de mama es la misma en los casos expuestos y no expuestos.

Estimación de riesgo

Expo | Expo

	Valor	Intervalo de confianza al 95% Inferior	Superior
Razón de las ventajas para Casos (Expo / No Expo)	,891	,651	1,219
Para la cohorte Controles = Expuesto	,959	,856	1,075
Para la cohorte Controles = No Expuesto	1,077	,882	1,315
N de casos válidos	755		

En la primera fila aparece el odds-ratio (OR), que es tanto mejor estimador del riesgo relativo cuanto más pequeñas sean las proporciones de desenlace en cada grupo.

El OR *(razón de ventajas o qué medida que lactaron es un riesgo de tener cáncer de mama)* de Casos y Controles es 0,281 y es significativo porque su intervalo de confianza no cubre el 1; mientras que el Emparejamiento tiene un OR de 0,891 y no es significativo porque su intervalo de confianza cubre el uno, indicando que la probabilidad de encontrarse con el cáncer de mama en los grupos estudiados es la misma.

Tablas 2X2 estratificadas (efecto de confusión)

Ejercicio 5.- La tabla adjunta refleja la asociación positiva de un estudio de casos y controles entre el consumo de café y el cáncer de páncreas:

Café	Casos	Controles
Sí	196	104
No	89	106

$$OR = \frac{196.106}{89.104} = 2{,}24 \qquad \text{Probabilidad} = \frac{OR}{OR+1} = 0{,}69$$

NOTA: *En la tabla (los casos serían los enfermos al final del estudio y los controles los no enfermos). Un OR = 2,24 se lee como (2,24:1), es decir, el cáncer de páncreas es 2,24 veces más que si no estuviera presente el café.*

Considerando el tabaco como un tercer factor, y dividir los individuos del estudio en dos estratos (fumadores y no fumadores):

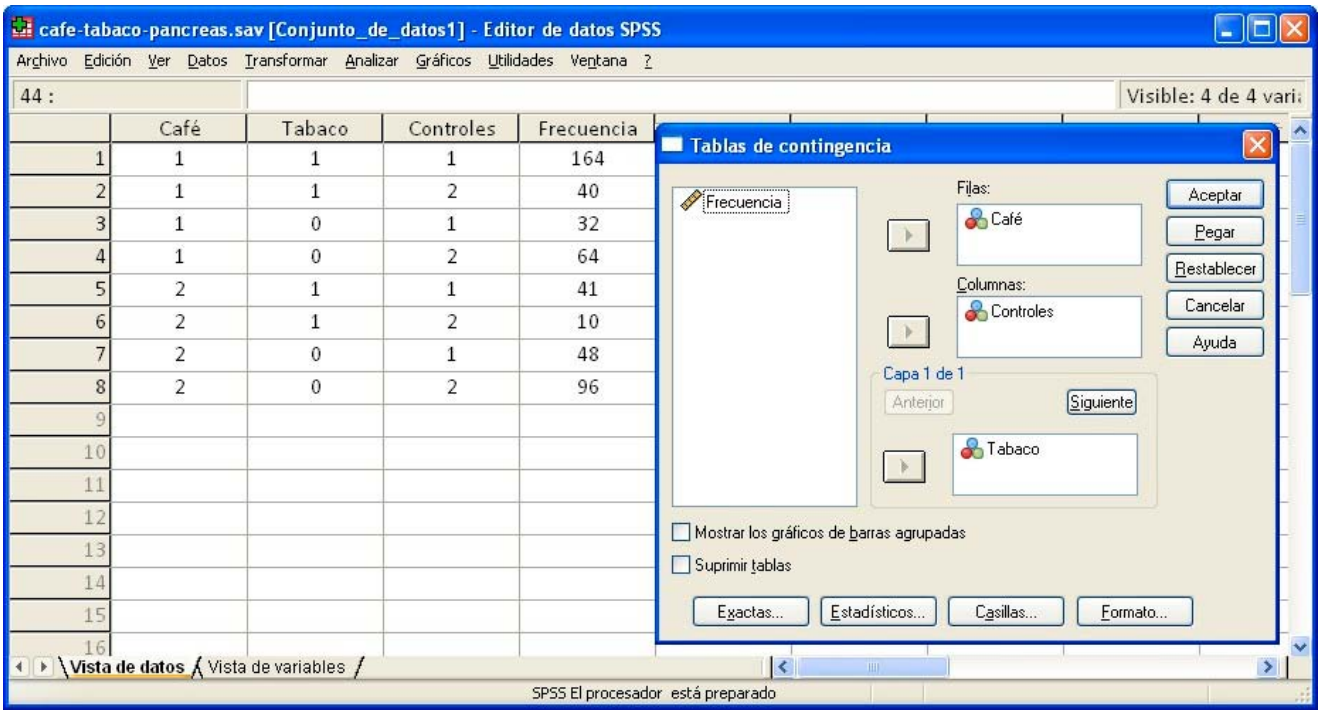
No Fumadores		
Café	Casos	Controles
Sí	32	64
No	48	96

$$OR = \frac{32.96}{48.64} = 1$$

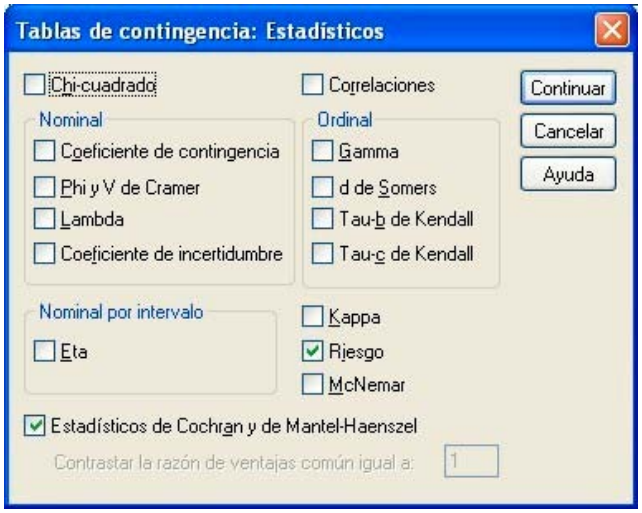
Fumadores		
Café	Casos	Controles
Sí	164	40
No	41	10

$$OR = \frac{164.10}{41.40} = 1$$

		Casos-Controles	
Café	Tabaco	Casos (1)	Controles (2)
Sí (1)	Fumadores (1)	164	40
	No Fumadores (0)	32	64
No (0)	Fumadores (1)	41	10
	No Fumadores (0)	48	96



Se conoce que las personas que toman café generalmente fuman, entonces el tabaquismo puede ser un *factor confusor*. Para estimar el efecto confusor, el análisis individual de cada estrato debe ser complementado con un análisis general considerando todos los valores de los estratos. El método de *Mantel-Haenszel* es uno de los más útiles para estimar el efecto de confusión.



Estimación de riesgo

Tabaco		Valor	Intervalo de confianza al 95% Inferior	Superior
No Fumadores	Razón de las ventajas para Café (Cafe / No Cafe)	1,000	,578	1,729
	Para la cohorte Controles = Caso	1,000	,694	1,441
	Para la cohorte Controles = Control	1,000	,833	1,200
	N de casos válidos	240		
Fumadores	Razón de las ventajas para Café (Cafe / No Cafe)	1,000	,462	2,166
	Para la cohorte Controles = Caso	1,000	,859	1,164
	Para la cohorte Controles = Control	1,000	,537	1,861
	N de casos válidos	255		

Se observa que hay dos tablas de contingencia (No Fumadores, Fumadores), ambas tienen un intervalo de confianza que contiene a la unidad, por tanto ambos OR son no significativos.

Pruebas de independencia condicional

	Chi-cuadrado	gl	Sig. asintótica (bilateral)
De Cochran	,000	1	1,000
Mantel-Haenszel	,000	1	1,000

El estadístico de Cochran es muy bajo y tiene un nivel crítico asociado de 1 > 0,05, con lo que se acepta la hipótesis nula de independencia condicional. A la misma conclusión se llega con el estadístico de Mantel-Haenszel.

La prueba de homogeneidad permite examinar las diferencias entre los *odds-ratio* de los estratos.

Pruebas de homogeneidad de la razón de las ventajas

	Chi-cuadrado	gl	Sig. asintótica (bilateral)
Breslow-Day	,000	1	1,000
De Tarone	,000	1	1,000

El estadístico es bajo y el nivel crítico asociado a ambos estadísticos es mayor que 5% (p > 0,05), lo que hace pensar que no hay diferencias apreciables entre los OR en los estratos y que, en consecuencia, los resultados ajustados pueden considerarse para el conjunto, esto es, se mantiene la hipótesis de homogeneidad.

Cómo se puede asumir que el riesgo es homogéneo en todos los estratos, tiene sentido una estimación común del riesgo. La solución viene dada por el estadístico de Mantel-Haenszel.

Estimación de la razón de las ventajas común de Mantel-Haenszel

Estimación			1,000
ln(estimación)			,000
Error típ. de ln(estimación)			,228
Sig. asintótica (bilateral)			1,000
Intervalo de confianza asintótico al 95%	Razón de ventajas común	Límite inferior	,640
		Límite superior	1,564
	ln(Razón de ventajas común)	Límite inferior	-,447
		Límite superior	,447

La estimación de la razón de las ventajas común de Mantel-Haenszel se distribuye de manera asintóticamente normal bajo el supuesto de razón de las ventajas común igual a 1,000. Lo mismo ocurre con el log natural de la estimación.

El valor del riesgo común (estimación) es 1, con un intervalo de confianza definido por los límites 0,640 y 1,564, como el intervalo de confianza cubre el 1, se puede concluir que el OR no es significativo, es decir, el tabaco es un *factor confusor*.

Hoja nº .....

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

Hoja nº .....

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

(IMPRESO EN PAPEL RECICLADO)

UNICAMENTE PARA USO ESCOLAR

Instrumentos Estadísticos Avanzados
Facultad Ciencias Económicas y Empresariales
Departamento de Economía Aplicada
Profesor: Santiago de la Fuente Fernández