TABLAS CONTINGENCIA

ESTUDIO CASOS - CONTROLES

Estudio de Casos - Controles **4**
Estudio de Cohortes **19**
Análisis Estratificado **28**
Tablas Generales M x N **34**
Análisis Regresión Logística **58**

ESTUDIO CASOS - CONTROLES

Es un estudio observacional, analítico y longitudinal, donde los sujetos se seleccionan en función que tengan (*casos*) o no tengan (*control*) una determinada característica.

Seleccionados los sujetos en cada grupo, se analiza si estuvieron expuestos o no expuestos a una característica de interés y se compara la proporción de sujetos expuestos en el *grupo de casos* frente a los sujetos del *grupo de controles*.

El estudio de la influencia de una variable (independiente) sobre la forma en que se modifica otra variable (dependiente) se conoce como análisis bivariado.
El análisis es multivariado cuando el estudio evalúa de forma simultánea el efecto de dos o más variables independientes sobre una variable dependiente.

Las tablas de contingencia (tablas de dobble entrada) constituyen una herramienta fundamental para este tipo de análisis.

Entre filas y columnas delimitan *celdas* donde se recogen las frecuencias de cada combinación de las variables analizadas.

La expresión más elemental de las tablas de contingencia tiene 2 filas y 2 columnas (tablas de 2 x 2).

Las tablas de contingencia (tablas de doble entrada) son una herramienta imprescindible para este tipo de análisis, se componen de filas para la información de una variable y columnas para la información de la otra variable. Las filas y columnas delimitan celdas donde se recoge la información de las variables analizadas.
En general, las tablas pueden abarcar varias filas (M) y columnas (N), dando lugar a una tabla de contingencia M x N.

Las tablas de contingencia 2 x 2 simples (de único estrato) permiten el análisis de dos variables dicotómicas, una variable independiente y una variable dependiente.

📖 **Las frecuencias de una tabla de contingencia pueden obtenerse utilizando dos estrategias básicas de recogidas de datos. En la estrategia habitual, los datos representan un corte temporal *transversal*: se recogen en el mismo o aproximdamente el mismo punto temporal.**

Si, en lugar de esto, se miden una o más variables en una muestra de sujetos y se hace un seguimiento a estos sujetos para volver a tomar una medida de esas mismas variables o de otras diferentes, es una situación *longitudinal*: las medidas se toman en diferentes puntos temporales.
Los *índices de riesgo* que se estudian resultan especialmente útiles para diseños longitudinales en los que se miden dos variables *dicotómicas*.

El seguimiento de los estudios longitudinales puede hacerse *hacia adelante* o *hacia atrás*.

En los diseños *longitudinales hacia adelante*, llamados diseños *prospectivos o de cohortes*, los sujetos son clasificados en dos grupos con arreglo a la presencia o ausencia de algún factor desencadenante (por ejemplo, el hábito de fumar, fumadores y no fumadores) y se les hace un seguimiento durante un espacio de tiempo hasta determinar la proporción de sujetos de cada grupo en los que se da un determinado desenlace o incidencia objeto de estudio (por ejemplo, problemas vasculares).

En los diseños *longitudinales hacia atrás*, también denominados *retrospectivos o de caso-control*, se forman dos grupos a partir de la presencia o ausencia de una determinada condición objeto del estudio (por ejemplo, sujetos sanos y pacientes con problemas vasculares) y se hace un seguimiento hacia atrás intentando encontrar información sobre la proporción en la que se encuentra presente en cada muestra un determinado factor desencadente (por ejemplo, el hábito de fumar).

Lógicamente, cada diseño de recogida de datos permite dar respuesta a diferentes preguntas y requiere la utilización de unos estadísticos particulares.

ESTADÍSTICOS SEGÚN DISEÑO:

En el diseño de *cohortes (longitudinal hacia adelante)*, en los que se establecen dos grupos de sujetos a partir de la presencia o ausencia de una condición que se considera desencadente y se hace un seguimiento hacia adelante para determinar qué proporción de sujetos de cada grupo alcanza un determinado desenlace o incidencia, la medida de interés suele ser el *riego relativo* (RR): grado en que la proporción de desenlaces o incidencias es más alta en un grupo que en el otro.

En el diseño de *caso-control (longitudinal hacia atrás)*, tras formar dos grupos de sujetos a parir de alguna condición de interés, se va hacia atrás buscando la presencia de algún factor desencadenante - Por ejemplo, en el estudio sobre el tabaquismo y problemas vasculares se podría diseñar seleccionando dos grupos de sujetos diferenciados por la presencia de problemas vasculares y buscando en la historia clínica la presencia o no de fumar -.

Puesto que el tamaño de los grupos se fija a partir de la presencia o ausencia de un determinado desenlace, no tiene sentido calcular un índice de riesgo basado en las proporciones de desenlaces o incidencias, pues el número de fumadores y no fumadores no ha sido previamente establecido sino que es producto del muestreo. Se puede calcular la *ratio* fumadores/no-fumadores tanto en el grupo de sujetos con problemas vasculares como en el grupo de sujetos sin problemas, y utilizar el *cociente de ambas ratios* como una estimación del *riesgo relativo*.

Estudios Transversales o de Prevalencia (potenciales factores de riesgo)

Estudian simultáneamente la exposición y la enfermedad en un momento determinado. La obtención de datos puede ser prolongada (semanas o meses). El estudio transversal facilita información de gran utilidad para valorar el estado de salud de una comunidad y determinar sus necesidades.
Utiliza un formato de tabla para análisis bivariado de variables dicotómicas donde la variable independiente (exposición) se presenta en filas y la variable dependiente (daño o enfermedad) en las columnas.
La prevalencia corresponde a la probabilidad de padecer una enfermedad antes de realizar la prueba.

	Enfermedad		
Factor de riesgo	Sí	No	Total
Expuestos	a	b	a + b
No expuestos	c	d	c + d
Total	a + c	b + d	a + b + c + d

PREVALENCIAS DE ENFERMEDAD:

- **Riesgo en expuestos:** $p_1 = \frac{a}{a+b}$
- **Riesgo en no expuestos:** $p_2 = \frac{c}{c+d}$
- **Razón prevalencia ≡ Riesgo relativo (RR)** $= \frac{\text{Incidencia en expuestos}}{\text{Incidencia en no expuestos}} = \frac{p_1}{p_2}$

El Riesgo Relativo (RR) expresa cuántas veces más aparece una enfermedad en los expuestos que en los no-expuestos, o bien cuántas veces más riesgo tienen los expuestos en relación con los no-expuestos.

$RR = 1 \rightarrow$ **No hay asociación**

$RR > 1 \rightarrow$ **Asociación al factor de riesgo**

$RR < 1 \rightarrow$ **Asociación al factor de protección**

- **Riesgo atribuible o diferencia de riesgos (RA)** $= \frac{a}{a+b} - \frac{c}{c+d}$

El Riesgo Atribuible o diferencia de riesgos (RA) indica la cantidad adicional de incidencia de tener una enfermedad (o exceso de riesgo) que tienen los expuestos a los no-expuestos.

- **Intervalo de confianza para la razón de prevalencia RR de enfermedad con distribución asintóticamente normal:**

$$IC_{1-\alpha}(RR) = RR \cdot e^{\pm z_{\alpha/2} \cdot \sqrt{V(\ln RR)}} = \left(RR \cdot e^{-z_{\alpha/2} \cdot \sqrt{V(\ln RR)}} \ , \ RR \cdot e^{z_{\alpha/2} \cdot \sqrt{V(\ln RR)}} \right)$$

donde $V\left[\ln RR\right] = \frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}$

PREVALENCIAS DE EXPOSICIÓN:

- **Riesgo en expuestos:** $p_1 = \frac{a}{a+c}$

- **Riesgo en no expuestos:** $p_2 = \frac{b}{b+d}$

- **Razón prevalencia** $\equiv$ **Riesgo relativo esperado (RR)** $= \frac{\text{Incidencia en enfermos}}{\text{Incidencia en sanos}} = \frac{p_1}{p_2}$

- **Intervalo de confianza para la razón de prevalencia RR de exposición con distribución asintóticamente normal:**

$$IC_{1-\alpha}(RR) = RR \cdot e^{\pm z_{\alpha/2} \cdot \sqrt{V(\ln RR)}} = \left(RR \cdot e^{-z_{\alpha/2} \cdot \sqrt{V(\ln RR)}} \ , \ RR \cdot e^{z_{\alpha/2} \cdot \sqrt{V(\ln RR)}} \right)$$

donde $V\left[\ln RR\right] = \frac{1}{a} - \frac{1}{a+c} + \frac{1}{b} - \frac{1}{b+d}$

MEDIDA DE ASOCIACIÓN: Odds Ratio (OR)

La medida de asociación que se utiliza en los estudios de casos y controles para evaluar la fuerza de la asociación ente el factor en estudio y el evento se conoce como Odds Ratio, medida que indica la frecuencia relativa de la exposición entre los casos y los controles.

Razón de Odds anticipado: $OR = \frac{p_1 \cdot (1-p_2)}{p_2 \cdot (1-p_1)} = \frac{a \times d}{b \times c}$

En el estudio de casos y controles no se puede estimar directamente la incidencia de la enfermedad en los expuestos y no expuestos, dado que los sujetos son seleccionados basándose en la presencia o ausencia del evento en estudio y no por

el estatus de exposición (a excepción de variantes del estudio como los anidados y caso-cohorte).
Sin embargo, cuando la presencia de la enfermedad es baja, el Odds ratio puede ser un estimador no sesgado de la razón de tasas de incidencia o de riesgo relativo.

El Odds ratio indica cuantas veces es mayor, o menor si la exposición actúa como un factor protector, la probabilidad de los casos que han estado expuestos al factor en estudio en comparación con los controles: $0 \leq OR < \infty$

$OR = 1 \rightarrow$ **Indica que la exposición analizada no se asocia con la enfermedad.**
$OR < 1 \rightarrow$ **La exposición disminuye la probabilidad de desarrollar el evento.**
$OR > 1 \rightarrow$ **La exposición aumenta la probabilidad de desarrollar el evento.**

El Odds ratio es una estimación puntual de la magnitud de asociación entre un determinado factor y una enfermedad, para encontrar una medida de variabilidad de esta estimación se recurre al inervalo de confianza.
Cuanto más amplio sea el intervalo de confianza menor es la precisión de la estimación. Un intervalo de confianza que incluya el valor 1 indica que la asociación no es significativa.

Inconvenientes de OR

- **Es más susceptible a sesgos que otros diseños, posibilidad que disminuye si el estudio considera el uso de datos recogidos con anterioridad a la ocurrencia de la enfermedad.**
- **El riesgo o la incidencia de la enfermedad no se puede medir directamente.**

Ventajas de OR

- **Es económico en términos de recursos y tiempo debido al menor tamaño de muestra requerido.**
- **Es más adecuado para enfermedades con largo período de investigación. La elección de un diseño prospectivo no resultaría eficiente pues para detectar los casos se tendría que seguir a la población completa durante un largo período.**
- **Puede evaluar simultáneamente la exposición a múltiples factores etiológicos**

- **Intervalo de confianza para OR (Método de Woolf)**

$$IC_{1-\alpha}(OR) = OR \,.\, e^{\pm z_{\alpha/2} \cdot \sqrt{V(\ln OR)}} = \left(OR \,.\, e^{-z_{\alpha/2} \cdot \sqrt{V(\ln OR)}} \;,\; OR \,.\, e^{z_{\alpha/2} \cdot \sqrt{V(\ln OR)}} \right)$$

$$\text{donde } V\left[\ln OR\right] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

⊕ **Cuando el tamaño de la muestra es inferior a 30 se introduce la corrección:**

$$OR = \frac{(a + 0{,}5) \times (d + 0{,}5)}{(b + 0{,}5) \times (c + 0{,}5)}$$

TAMAÑO MUESTRAL:

Siendo, $OR = \frac{p_1 . (1-p_2)}{p_2 . (1-p_1)}$ con $p_1 = \frac{a}{a+c}$, $p_2 = \frac{b}{b+d}$, se expresa:

$$p_1 = \frac{p_2}{\frac{(1-p_2)}{OR} + p_2} \quad y \quad p_2 = \frac{p_1}{OR \,.\, (1-p_1) + p_1}$$

en este caso, $n = z^2_{\alpha/2} \frac{\frac{1}{p_1(1-p_1)} + \frac{1}{p_2(1-p_2)}}{\ln^2(1-\in)}$ $\in \equiv$ precisión relativa

CONTRASTE CHI-CUADRADO DE ASOCIACIÓN:

Hipótesis nula H_0 : Las distribuciones teórica y empírica son independientes

Estadístico observado $\chi^2_{k-1} = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{O_i^2}{e_i} - n$

Se acepta H_0 si $\sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} < \chi^2_{\alpha, k-1}$, siendo $e_{ij} = \frac{O_{x_i} \,.\, O_{y_j}}{n}$

En las tablas de contingencia 2 x 2 se puede obtener la Chi-cuadrado únicamente con las frecuencias observadas:

$$\chi^2_1 = \frac{n \,.\, (a \,.\, c - b \,.\, d)^2}{(a+b) \,.\, (c+d) \,.\, (a+c) \,.\, (b+d)}$$

$p - valor > 0,05 \Rightarrow$ Se acepta la hipótesis nula con una fiabilidad del 95%

$\oplus$ Cuando el valor de todas las celdas es mayor que 5 se puede utilizar una aproximación normal de la distribución hipergeométrica con:

$$\mu = \frac{(a+b) \,.\, (a+c)}{(a+b+c+d)} \quad , \quad \sigma = \sqrt{\frac{(a+b) \,.\, (a+c) \,.\, (b+d) \,.\, (c+d)}{(a+b+c+d)^2 \,.\, (a+b+c+d-1)}}$$

con el test estadístico $z = \frac{a-\mu}{\sigma}$ y con la regla de decisión si $z \geq z_{\alpha/2}$ se rechaza la hipótesis nula de no asociación.

Este procedimiento es equivalente al estadístico χ^2 de Mantel-Haenszel, es decir, $z^2 \approx \chi^2_{M-H}$

⊕ **El estadístico Chi-cuadrado de Pearson se puede calcular con las frecuencias esperadas** $e_{ij} = \frac{O_{i\bullet} x O_{\bullet j}}{n}$

	Enfermedad		
Factor de riesgo	**Sí**	**No**	**Total**
Expuestos	$O_{11} = a$ $e_{11} = \frac{O_{1\bullet} x O_{\bullet 1}}{n}$	$O_{12} = b$ $e_{12} = \frac{O_{1\bullet} x O_{\bullet 2}}{n}$	$O_{1\bullet} = a + b$
No expuestos	$O_{21} = c$ $e_{21} = \frac{O_{2\bullet} x O_{\bullet 1}}{n}$	$O_{22} = d$ $e_{22} = \frac{O_{2\bullet} x O_{\bullet 2}}{n}$	$O_{2\bullet} = c + d$
Total	$O_{\bullet 1} = a + c$	$O_{\bullet 2} = b + d$	$n = a + b + c + d$

$$\chi_1^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{O_{ij}^2}{e_{ij}} - n$$

Además del estadístico Chi-cuadado de Pearson con una probabilidad asociada (p_valor = Sig. asint. = Significación asíntótica), la tabla con SPSS y en pruebas diagnósticas se muestra otro estadístico:

$$\text{Razón de verosimilitud} = 2\sum_{i=1}^{k}\sum_{j=1}^{m} O_{ij} \log\left(\frac{O_{ij}}{e_{ij}}\right)$$

CONTRASTE CHI-CUADRADO CON LA CORRECCIÓN DE YATES:

$$\chi_1^2 = \frac{n.\left(\left|a.c - b.d\right| - \frac{n}{2}\right)^2}{(a+b).(c+d).(a+c).(b+d)}$$

Cuando $\left|a.c - b.d\right| \le \frac{n}{2}$ **la corrección de Yates es contraproducente.**

$p - valor > 0,05 \Rightarrow$ **Se acepta la hipótesis nula de no asociación con una fiabilidad del 95%**

CONTRASTE EXACTO DE FISHER:

$$F = \frac{(a+b)!.(c+d)!.(a+c)!.(b+d)!}{n!.a!.b!.c!.d!}$$

Un estudio transversal para conocer la prevalencia de una enfermedad y su relación con algunos factores de riesgo observa a 400 mujeres. Se analizaron las variables dicotómicas osteoporosis y antecedentes de dieta pobre en calcio, obteniendo los siguientes datos:

	Osteoporosis		
Factor de riesgo	Enfermos	Sanos	Total
Expuestos	56	64	120
No expuestos	24	256	280
Total	80	320	400

PREVALENCIAS DE ENFERMEDAD:

Riesgo en expuestos: $p_1 = \frac{56}{120} = 0,466667$

Riesgo en no expuestos: $p_2 = \frac{24}{280} = 0,085714$

Razón de prevalencia: $RR = \frac{0,466667}{0,085714} = 5,444444$

Intervalo de confianza para la razón de prevalencia RR de enfermedad con distribución asintóticamente normal:

$$V\left[\ln RR\right] = \frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d} = \frac{1}{56} - \frac{1}{120} + \frac{1}{24} - \frac{1}{280} = 0,047619$$

$$e^{\pm\, z_{\alpha/2}\cdot\sqrt{V(\ln RR)}} = e^{\pm\, 1,96\,.\sqrt{0,047619}} = e^{\pm\, 0,427707}$$

$$IC_{0,95}(RR) = \left(5,444444\,.\,e^{-0,427707}\ ,\ 5,444444\,.\,e^{0,427707}\right) = (3,549819\ ,\ 8,350279)$$

Razón de Odds anticipado: $OR = \frac{(0,466667\,/\,0,533333)}{(0,085714\,/\,0,914286)} = 9,333333$

PREVALENCIAS DE EXPOSICIÓN:

Riesgo en expuestos: $p_1 = \frac{56}{80} = 0,7$

Riesgo en no expuestos: $p_2 = \frac{64}{320} = 0,2$

Razón prevalencia ≡ Riesgo relativo (RR) $= \frac{p_1}{p_2} = \frac{0,7}{0,2} = 3,5$

Intervalo de confianza para la razón de prevalencia RR de exposición con distribución asintóticamente normal:

$$\text{donde } V\left[\ln RR\right] = \frac{1}{a} - \frac{1}{a+c} + \frac{1}{b} - \frac{1}{b+d} = \frac{1}{56} - \frac{1}{80} + \frac{1}{64} - \frac{1}{320} = 0,017857$$

$$e^{\pm z_{\alpha/2} \cdot \sqrt{V(\ln R)}} = e^{\pm 1,96 \cdot \sqrt{0,017857}} = e^{\pm 0,261916}$$

$$IC_{0,95}(RR) = \left(3,5 \cdot e^{-0,261916} \ , \ 3,5 \cdot e^{0,261916}\right) = (2,693528 \ , \ 4,547939)$$

MEDIDAS DE ASOCIACIÓN:

Razón de Odds anticipado: $OR = \frac{56 \text{ x } 256}{64 \text{ x } 24} = 9,333333$

Intervalo de confianza aproximado para OR (Método de Woolf)

$$IC_{1-\alpha}(OR) = OR \cdot e^{\pm z_{\alpha/2} \cdot \sqrt{V(\ln OR)}} = \left(OR \cdot e^{-z_{\alpha/2} \cdot \sqrt{V(\ln OR)}} \ , \ OR \cdot e^{z_{\alpha/2} \cdot \sqrt{V(\ln OR)}}\right)$$

$$V\left[\ln OR\right] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} = \frac{1}{56} + \frac{1}{64} + \frac{1}{24} + \frac{1}{256} = 0,079055$$

$$e^{\pm z_{\alpha/2} \cdot \sqrt{V(\ln OR)}} = e^{\pm 1,96 \cdot \sqrt{0,079055}} = e^{\pm 0,551088}$$

$$IC_{0,95}(OR) = \left(9,333333 \cdot e^{-0,551088} \ , \ 9,333333 \cdot e^{0,551088}\right) = (5,379064 \ , \ 16,194474)$$

CONTRASTE CHI-CUADRADO DE ASOCIACIÓN:

$$\chi_1^2 = \frac{n \cdot (a \cdot c - b \cdot d)^2}{(a+b) \cdot (c+d) \cdot (a+c) \cdot (b+d)} = \frac{400 \cdot (56 \cdot 256 - 64 \cdot 24)^2}{120 \cdot 280 \cdot 80 \cdot 320} = 76,1905$$

Si el $p - valor > 0,05$ **se acepta la hipótesis nula con una fiabilidad del 95%**

⊕ **Como el valor de todas las celdas es mayor que 5 se utiliza una aproximación normal:**

$$\mu = \frac{(a+b) \cdot (a+c)}{(a+b+c+d)} = \frac{120 \cdot 80}{400} = 24$$

$$\sigma = \sqrt{\frac{(a+b) \cdot (a+c) \cdot (b+d) \cdot (c+d)}{(a+b+c+d)^2 \cdot (a+b+c+d-1)}} = \sqrt{\frac{120 \cdot 80 \cdot 320 \cdot 280}{400^2 \cdot 399}} = 3,670652$$

$z = \frac{a - \mu}{\sigma} = \frac{56 - 24}{3,670652} = 8,717797 > 1,96 \Rightarrow$ **Se rechaza la hipótesis nula de NO asociación**

Se observa que $z^2 = 8,717797^2 = 76 \approx 76,1905\ (\chi^2_{M-H})$

CONTRASTE CHI-CUADRADO CON LA CORRECCIÓN DE YATES:

$$\chi^2_1 = \frac{n.\left(\left|a.c-b.d\right|-\frac{n}{2}\right)^2}{(a+b).(c+d).(a+c).(b+d)} = \frac{400.\left(\left|56.256-64.24\right|-\frac{400}{2}\right)^2}{120.280.80.320} = 73,8281$$

Si el $p-valor > 0,05$ **se acepta la hipótesis nula con una fiabilidad del 95%**

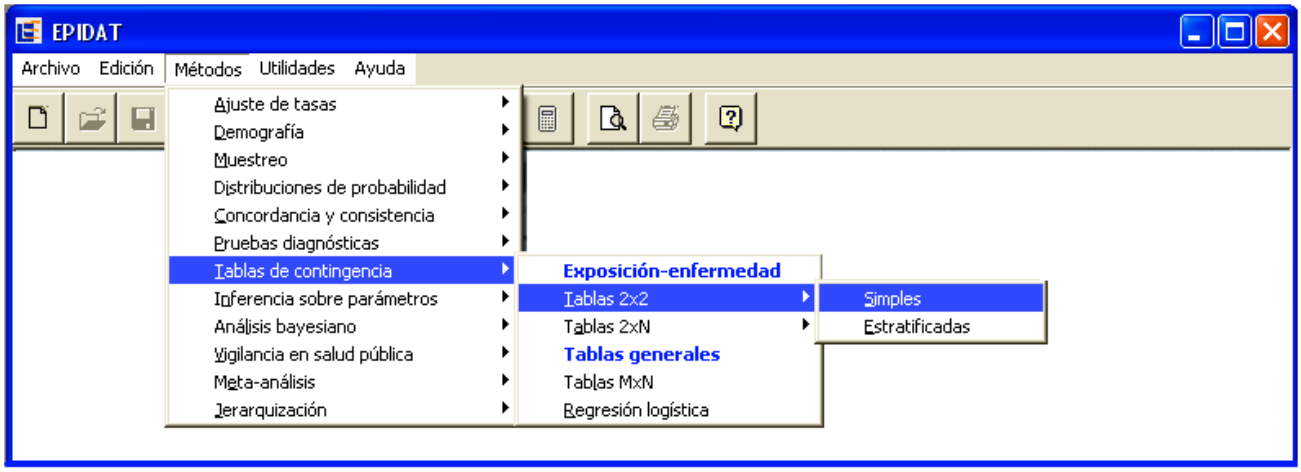
CONTRASTE EXACTO DE FISHER:

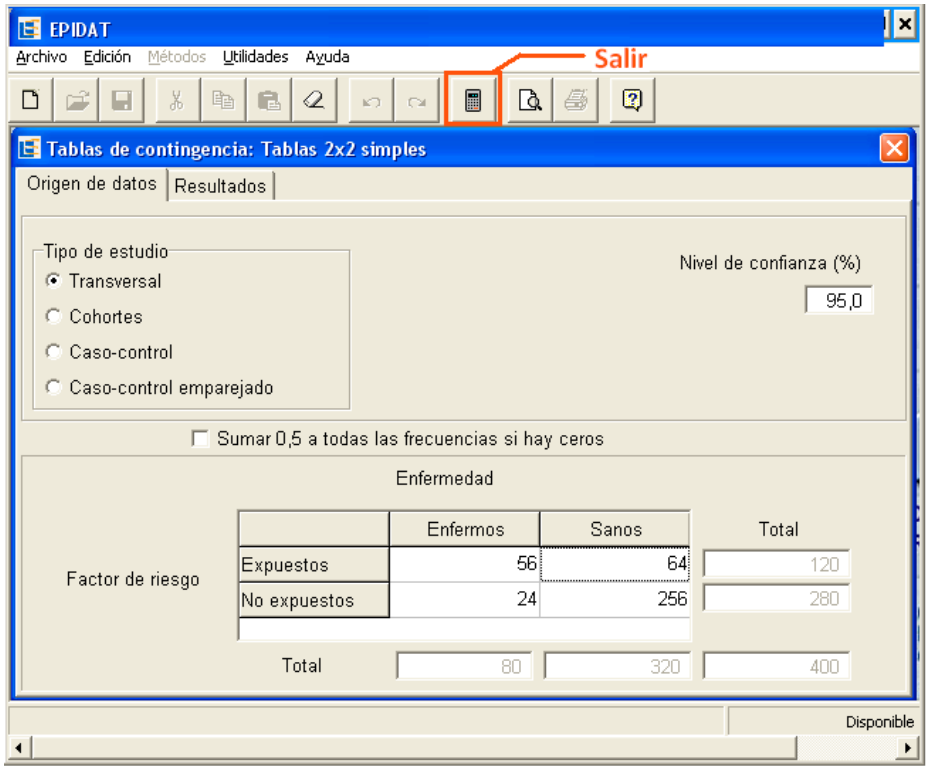
$$F = \frac{(a+b)!.(c+d)!.(a+c)!.(b+d)!}{n!.a!.b!.c!.d!} = \frac{120!.280!.80!.320!}{400!.56!.64!.24!.256!}$$

Si el $p-valor > 0,05$ **se acepta la hipótesis nula con una fiabilidad del 95%**

Epidat en el caso de un estudio transversal:

Métodos → ***Tablas de contingencia*** → ***Tablas 2 x 2*** → ***Simples***





[1] Tablas de contingencia: Tablas 2x2 simples

Tipo de estudio: Transversal
Nivel de confianza: 95,0%

Tabla

	Enfermos	Sanos	Total
Expuestos	56	64	120
No expuestos	24	256	280
Total	80	320	400

Prevalencia de la enfermedad	Estimación	IC(95,0%)		
En expuestos	0,466667	-	-	
En no expuestos	0,085714	-	-	
Razón de prevalencias	5,444444	3,549819	8,350279	(Katz)

Al tratarse de estudios transversales las frecuencias de los daños se presentan como tasas de prevalencia estimadas puntualmente. Las tasas cuantifican el número de personas que presentaban el daño en el momento del estudio en cada grupo (expuestos y no expuestos) en comparación con el total de la población en ambos grupos.
La prevalencia en expuestos fue del 46,6667%, mientras que la prevalencia en no-expuestos fue del 8,5714%.

La razón de prevalencias permite comparar la prevalencia de expuestos y no-expuestos. Si la razón de prevalencias es mayor que 1 indica que la exposición aumenta el riesgo de tener el daño, en caso de ser menor que 1 indicaría que la exposición sería un factor de protección. Cuando la razón de prevalencias es 1 sugiere que la exposición no se encuentra relacionada con el daño.

La razón de prevalencias es de 5,444444 > 1 indicando que la exposición aumenta el riesgo de tener el daño (existe una relación entre el antecedente y el daño).

El intervalo de confianza $\left[3,549819\ ,\ 8,350279\right]$ no cubre el 1 con lo que no es una prueba de significación, rechazando la hipótesis nula de que no hay asociación y de que los resultados son producto del azar.

Prevalencia de exposición	Estimación	IC(95,0%)		
En enfermos	0,700000	-	-	
En no enfermos	0,200000	-	-	
Razón de prevalencias	3,500000	2,693528	4,547939	(Katz)

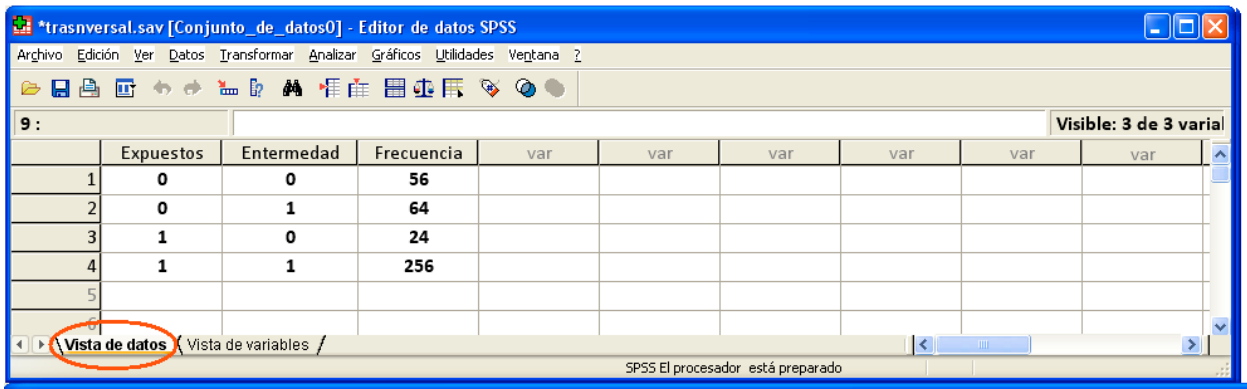
OR	IC(95,0%)		
9,333333	5,379064	16,194474	(Woolf)
	5,395274	16,140783	(Cornfield)

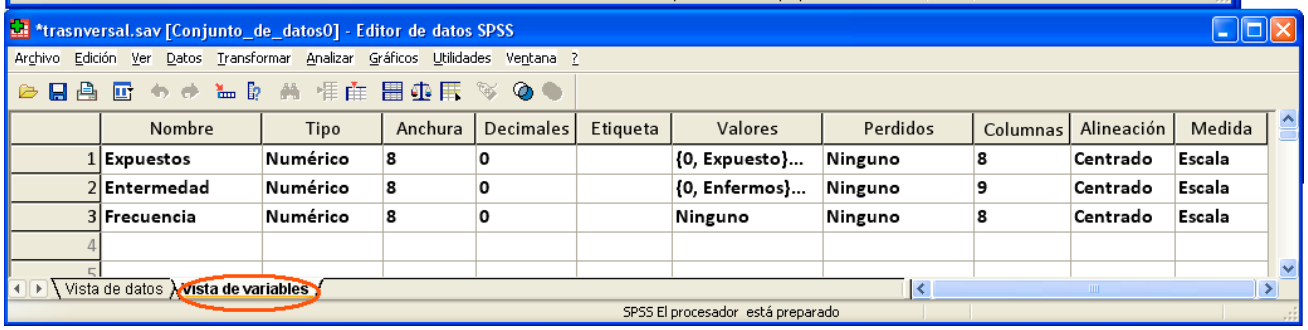
✍ **Un Odds ratio (OR) mayor que 1 indica que la exposición aumenta la probabilidad de desarrollar la enfermedad, en este caso 9 veces más, con un intervalo de confianza que no cure 1 por lo que se rechaza la hipotesis nula $H_0 : OR = 1$ que indica que la exposición no se asocia con la enferemedad.**

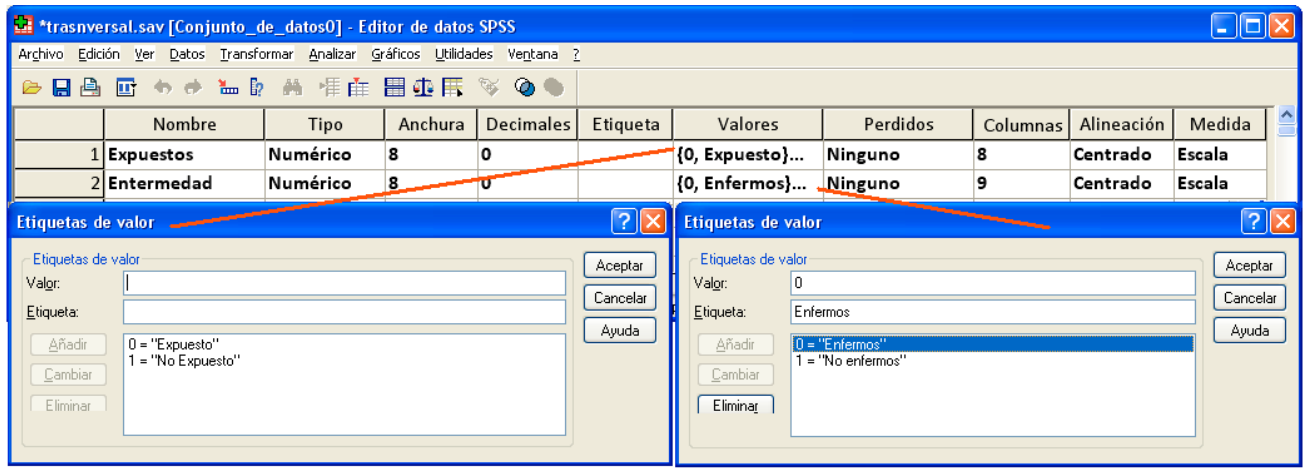
Prueba Ji-cuadrado de asociación	Estadístico	Valor p
Sin corrección	76,1905	0,0000
Corrección de Yates	73,8281	0,0000

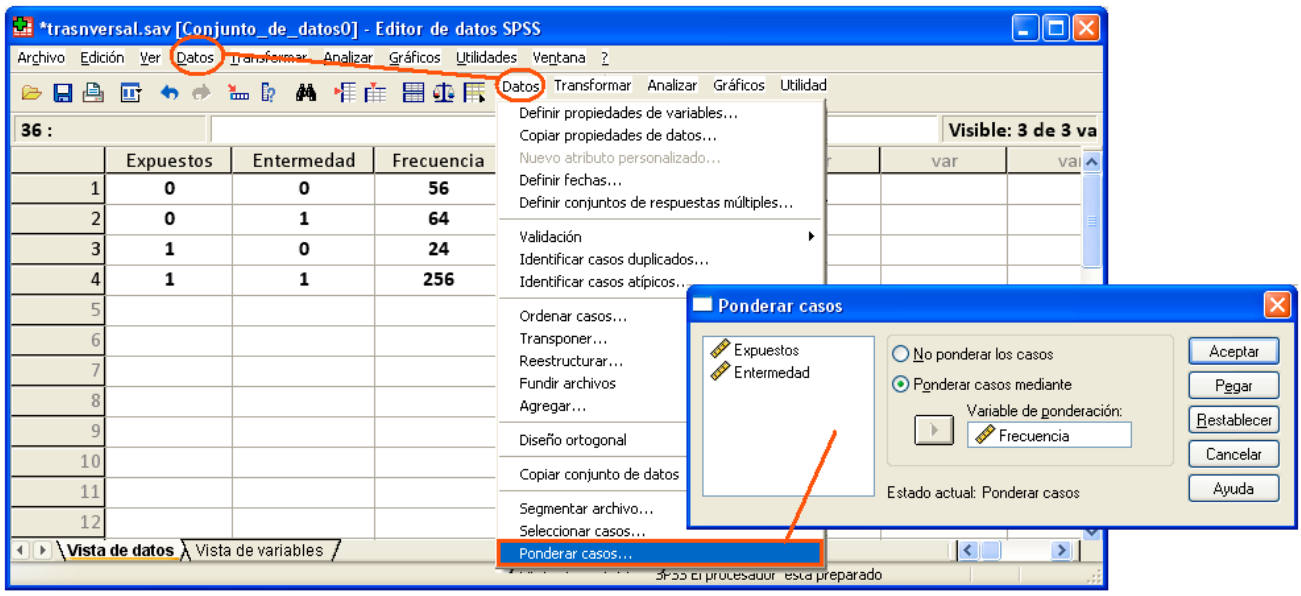
Prueba exacta de Fisher	Valor p
Unilateral	0,0000
Bilateral	0,0000

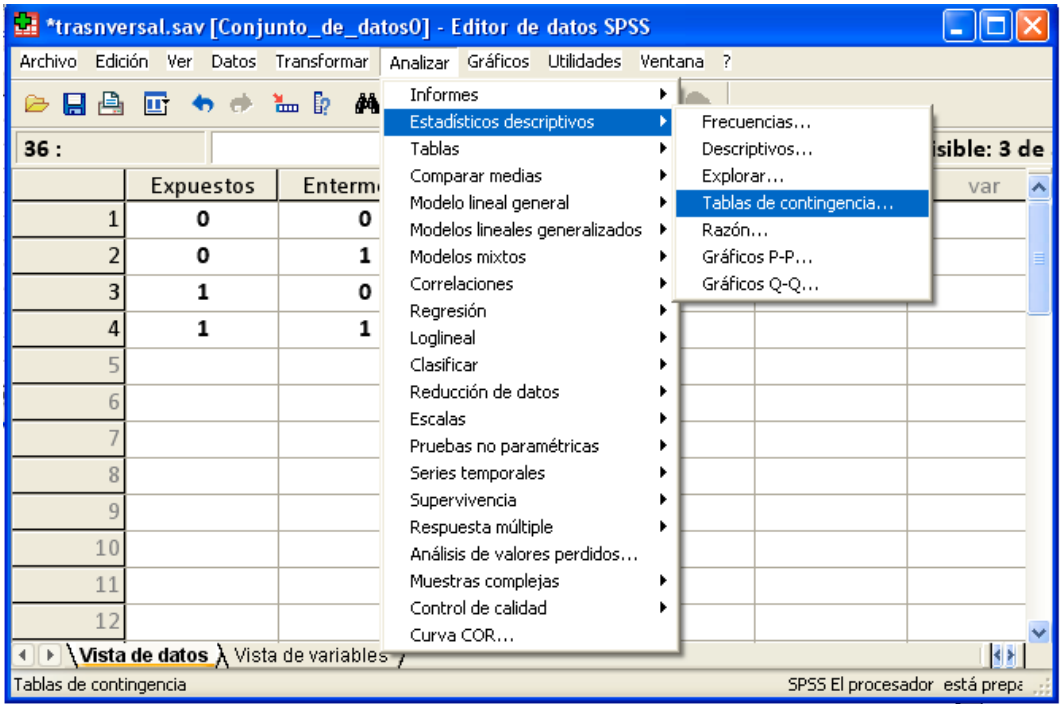
✍ **Las medidas de significación estadística se deciden según el p-valor, cuando $p\_valor < 0,05$ se rechaza la hipótesis nula de no-asociación con un nivel de confianza del 95%.**

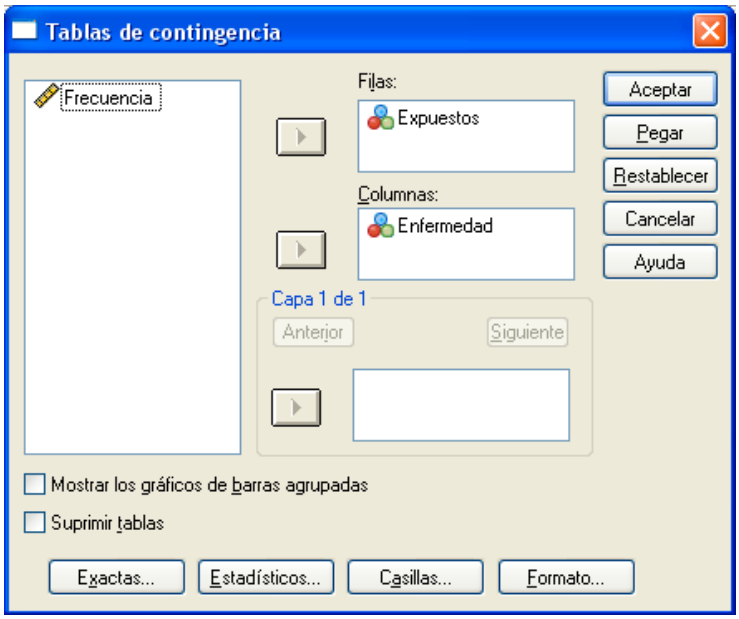

*trasnversal.sav [Conjunto_de_datos0] - Editor de datos SPSS
Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?
9 :
Visible: 3 de 3 varial
Expuestos Entermedad Frecuencia var var var var var var
1 0 0 56
2 0 1 64
3 1 0 24
4 1 1 256
Vista de datos Vista de variables
SPSS El procesador está preparado


*trasnversal.sav [Conjunto_de_datos0] - Editor de datos SPSS
Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?
Nombre Tipo Anchura Decimales Etiqueta Valores Perdidos Columnas Alineación Medida
1 Expuestos Numérico 8 0 {0, Expuesto}... Ninguno 8 Centrado Escala
2 Entermedad Numérico 8 0 {0, Enfermos}... Ninguno 9 Centrado Escala
3 Frecuencia Numérico 8 0 Ninguno Ninguno 8 Centrado Escala
Vista de datos Vista de variables
SPSS El procesador está preparado


*trasnversal.sav [Conjunto_de_datos0] - Editor de datos SPSS
Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?
Nombre Tipo Anchura Decimales Etiqueta Valores Perdidos Columnas Alineación Medida
1 Expuestos Numérico 8 0 {0, Expuesto}... Ninguno 8 Centrado Escala
2 Entermedad Numérico 8 0 {0, Enfermos}... Ninguno 9 Centrado Escala
Etiquetas de valor
Etiquetas de valor
Valor:
Etiqueta:
Añadir
Cambiar
Eliminar
0 = "Expuesto"
1 = "No Expuesto"
Aceptar
Cancelar
Ayuda
Etiquetas de valor
Valor: 0
Etiqueta: Enfermos
0 = "Enfermos"
1 = "No enfermos"


*trasnversal.sav [Conjunto_de_datos0] - Editor de datos SPSS
Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?
Datos Transformar Analizar Gráficos Utilidad
Definir propiedades de variables...
Copiar propiedades de datos...
Nuevo atributo personalizado...
Definir fechas...
Definir conjuntos de respuestas múltiples...
Validación
Identificar casos duplicados...
Identificar casos atípicos...
Ordenar casos...
Transponer...
Reestructurar...
Fundir archivos
Agregar...
Diseño ortogonal
Copiar conjunto de datos
Segmentar archivo...
Seleccionar casos...
Ponderar casos...
36 :
Visible: 3 de 3 va
Expuestos Entermedad Frecuencia
1 0 0 56
2 0 1 64
3 1 0 24
4 1 1 256
Vista de datos Vista de variables
Ponderar casos
Expuestos
Entermedad
No ponderar los casos
Ponderar casos mediante
Variable de ponderación:
Frecuencia
Estado actual: Ponderar casos
Aceptar
Pegar
Restablecer
Cancelar
Ayuda





En el Visor de SPSS salen los resultados:

Tabla de contingencia Expuestos * Entermedad

Recuento

		Entermedad		
		Enfermos	No enfermos	Total
Expuestos	Expuesto	56	64	120
	No Expuesto	24	256	280
Total		80	320	400

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	76,190476[b]	1	,000		
Corrección por continuidad[a]	73,828125	1	,000		
Razón de verosimilitudes	70,695602	1	,000		
Estadístico exacto de Fisher	76,190476			,000	,000
Asociación lineal por lineal	76,000000	1	,000		
N de casos válidos	400				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 24,00.

Pruebas de homogeneidad de la razón de las ventajas

	Chi-cuadrado	gl	Sig. asintótica (bilateral)
Breslow-Day	,000	0	.
De Tarone	,000	0	.

Pruebas de independencia condicional

	Chi-cuadrado	gl	Sig. asintótica (bilateral)
De Cochran	76,190476	1	,000
Mantel-Haenszel	73,643555	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.

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

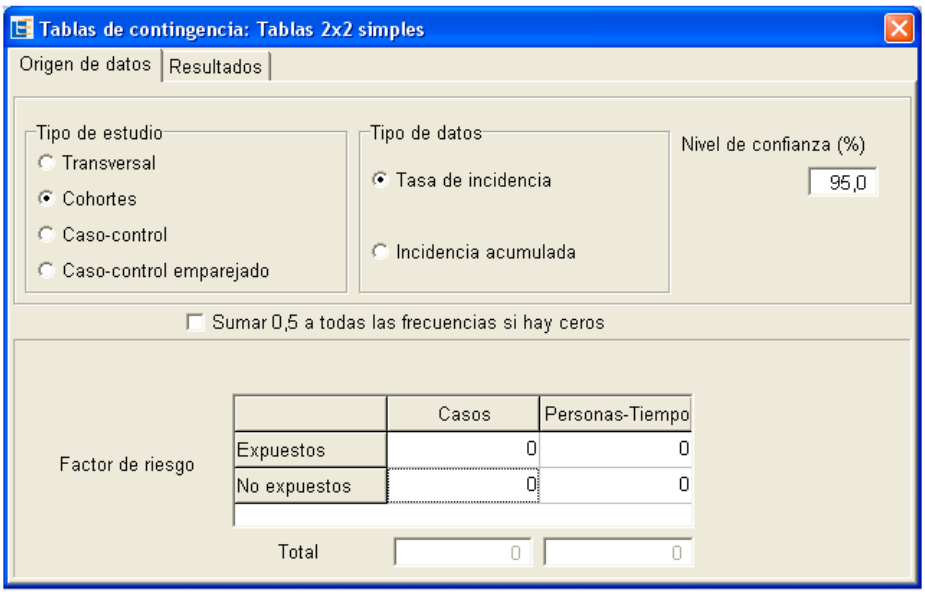
Estimación			9,33333
ln(estimación)			2,23359
Error típ. de ln(estimación)			,281167
Sig. asintótica (bilateral)			,000000
Intervalo de confianza asintótico al 95%	Razón de ventajas común	Límite inferior	5,37906
		Límite superior	16,1945
	ln(Razón de ventajas común)	Límite inferior	1,68251
		Límite superior	2,78467

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.

Estudio de Cohortes

Es un estudio epidemiológico, observacional, analítico, longitudinal prospectivo o retrospectivo, en el que se hace una comparación de la frecuencia de enfermedad (o de un determinado desenlace) entre dos poblaciones, una población está expuesta a un determinado factor de exposición o factor de riesgo al que no está expuesta la otra.

Los individuos que componen los grupos de estudio (dos o más) se seleccionan en función de la presencia de una determinada característica o exposición.



El estudio de *cohorte prospectivo (estudio de seguimiento, de proyección o de incidencia)*, los individuos no tienen la enfermedad de interés y son seguidos durante un determinado período de tiempo para observar la frecuencia con que la enfermedad aparece en cada uno de los grupos. Tiene por objetivo medir la *causalidad* entre factores de riesgo y la enfermedad en estudio.

El estudio de *cohorte retrospectivo* selecciona individuos que tienen una determinada enfermedad y analiza qué factores de riesgo han tenido en el pasado que pudieran provocar dicha enfermedad.

El *estudio de seguimiento* (prospectivo) se plantea con anterioridad al desarrollo de la enfermedad, ninguno de los individuos incluidos en los grupos tiene la enfermedad o característica en estudio.
Para analizar si la exposición influye en el desarrollo hay que comparar la incidencia de nuevos casos entre los grupos.

Las incidencias se pueden calcular de dos formas:

- **Número de casos nuevos en relación con la población que integra la cohorte (*incidencia acumulada*).**
- **Considerar el período que cada individuo permaneció en el grupo (*tasa de incidencia o densidad de incidencia*).**

La *incidencia acumulada es* más sencilla de calcular porque como denominador solo requiere el número de individuos que se incluyó en cada grupo, aunque es más precisa la *tasa de incidencia* ya que considera el momento en que se producen los casos y los períodos de seguimiento de los individuos, que tipicamente no son iguales para todos los individuos.

El formato de la tabla para el análisis de los estudios de cohorte es similar a las tablas de contingencia vistas, para el cálculo de las tasas de incidencia se requiere otra tabla de contingencia que considere el período Personas-Tiempo.

Factor de riesgo o Factor de protección	Enfermedad o daño	Personas-Tiempo
Expuestos	a	t_1
No expuestos	c	t_0
Total	a + c	$t = t_1 + t_0$

En metro Madrid para evaluar la exposición al amianto sobre el riesgo de padecer cáncer de pulmón un estudio comparó un grupo de 2435 trabajadores expuestos al amiento con otro grupo de 3895 empleados sin exposición a este factor.
A lo largo de 19 años, el primer grupo presenta 45 defunciones por cáncer de pulmón con un tiempo de seguimiento de 25219 personas/año, en tanto que en el grupo de no expuestos el número de defunciones por este causa fue 18 con un tiempo de seguimiento de 48343 personas/año.

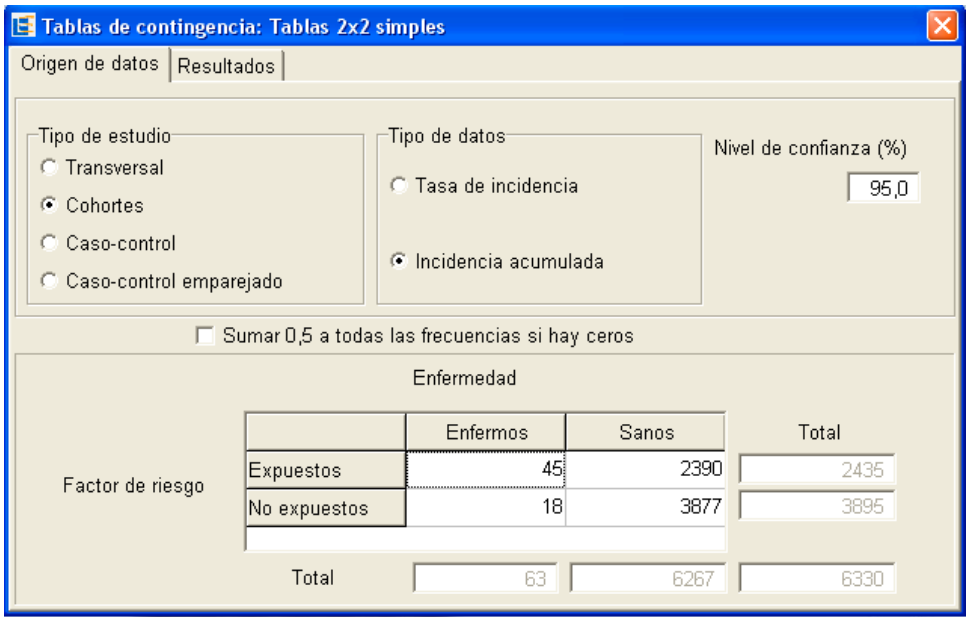
Tabla incidencia acumulada

	Defunción por cáncer		
Exposición al amianto	**Sí**	**No**	**Total**
Expuestos	**45**	**2390**	**2435**
No expuestos	**18**	**3877**	**3895**
Total	**63**	**6267**	**6330**

Tabla tasas de incidencia

Exposición al amianto	Defunciones	Personas-Año
Expuestos	**45**	**25219**
No expuestos	**18**	**48343**
Total	**63**	**73562**

Introduciendo los datos en EPIDAT:



Tablas de contingencia: Tablas 2x2 simples

Tipo de estudio: Cohortes
Tipo de datos: Incidencia acumulada
Nivel de confianza: 95,0%

Tabla

	Enfermos	Sanos	Total
Expuestos	45	2390	2435
No expuestos	18	3877	3895
Total	63	6267	6330

	Estimación	IC(95,0%)		
Riesgo en expuestos	0,018480	-	-	
Riesgo en no expuestos	0,004621	-	-	
Riesgo relativo	3,998973	2,320494	6,891544	(Katz)
Diferencia de riesgos	0,013859	0,008101	0,019617	
Odds ratio	4,055439	2,342174	7,021932	(Woolf)
		2,357461	6,975982	(Cornfield)

✍ **El Riesgo absoluto para el total del período en estudio se cuantifica mediante la incidencia acumulada: Tiene un riesgo en expuestos de 0,018480, que se interpreta como una incidencia de 1,848 %. El riesgo es considerablemente más alto que en los no-expuestos de 0,004621 (incidencia de 0,4621%). Por tanto, la exposición al amianto estaría causando un mayor riesgo de morir de cáncer de pulmón.**

✍ **El Riesgo relativo, con interpretación similar a la razón de prevalencias, de $3{,}998973 \approx 4$ indica que en los expuestos la incidencia es 4 veces la de los no expuestos, pudiendo interpretar que en los expuestos hay 3 veces más riesgo de tener cáncer de pulmón que en los no-expuestos.**

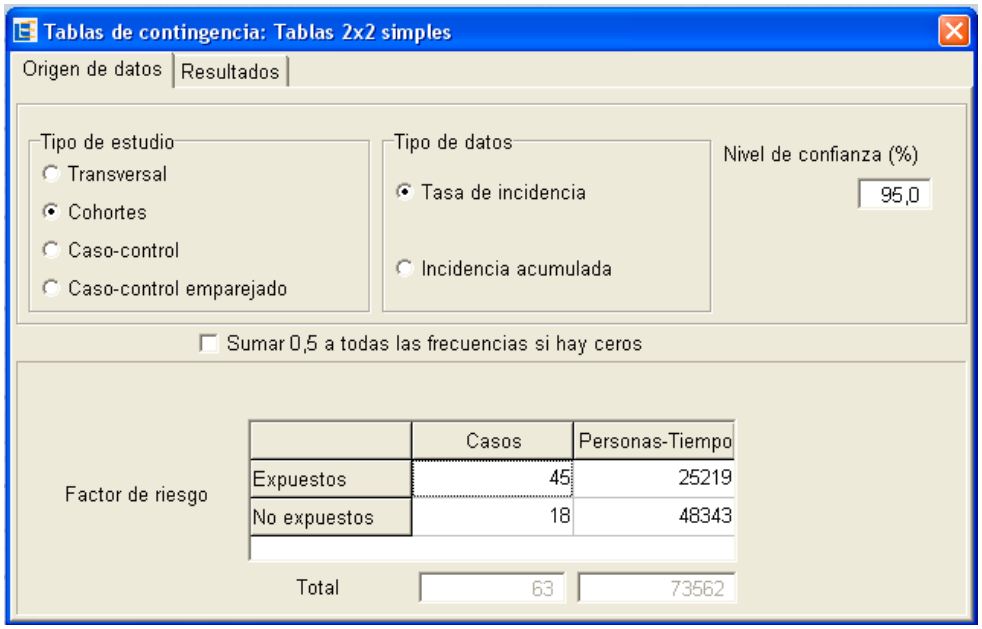
✍ **El Odds ratio tiene un valor de 4,055439 muy próximo al del Riesgo relativo por tratarse de una enfermedad poco frecuente.**

Fracción atribuible en expuestos	0,749936	0,569057	0,854895
Fracción atribuible poblacional	0,535668	0,314227	0,685605

✍ **La Función atribuible o previsora entre los expuestos representa la fracción del daño que podría evitarse entre los expuestos al elminarse esa exposición, es decir, si no hubiera ocurrido la exposición la fracción del daño no hubiera ocurrido.**
La Función atribuible es aplicada a un análisis prospectivo para contestar a la pregunta ¿cuánto daño se puede evitar si la población no se expusiera a un factor determinado?.
Si la exposición ya existe y se desea estimar la reducción del daño al eliminar la exposición, solo se puede aplicar cuando la exposición sea totalmente reversible.
Este indicador tiene una virtualidad teórica al cuantificar supuestamente el peso etiológico de determinado factor en la salud pública.
Un 0,749936 (75%) de los casos de cáncer de pulmón de los trabajadores del metro de Madrid han sido por su exposición al amianto, con un intervalo de confianza [57% - 85%]

Prueba Ji-cuadrado de asociación	Estadístico	Valor p
Sin corrección	29,2069	0,0000
Corrección de Yates	27,8173	0,0000

✍ **En las pruebas de Chi-cuadrado $p-\text{valor} = 0{,}0000 < 0{,}05$ admitiendo la asociación entre el amianto y el cáncer de pulmón.**



Tablas de contingencia: Tablas 2x2 simples

Tipo de estudio: Cohortes
Tipo de datos: Tasa de incidencia
Nivel de confianza: 95,0%

Tabla

	Casos	Personas-Tiempo
Expuestos	45	25219
No expuestos	18	48343
Total	63	73562

	Estimación	IC(95,0%)	
Tasa de incidencia en expuestos	0,001784	-	-
Tasa de incidencia en no expuestos	0,000372	-	-
Razón de tasas de incidencia	4,792319	2,774323	8,278173
Diferencia de tasas de incidencia	0,001412	0,000863	0,001961

La Tasa de incidencia en expuestos y no expuestos al considerar el tiempo real de seguimiento evita los errores que se pueden producir por las diferencias de seguimiento entre los grupos. En este sentido, el tiempo promedio de seguimiento en los expuestos fue de $25219 / 2435 = 10,36$ años menor que el de no-expuestos $48343 / 3895 = 12,41$ años .
La tasa de incidencia anual en expuestos es de 0,001784 (0,1784%) mayor que la tasa de incidencia en no expuestos de 0,000372 (0,0372%)

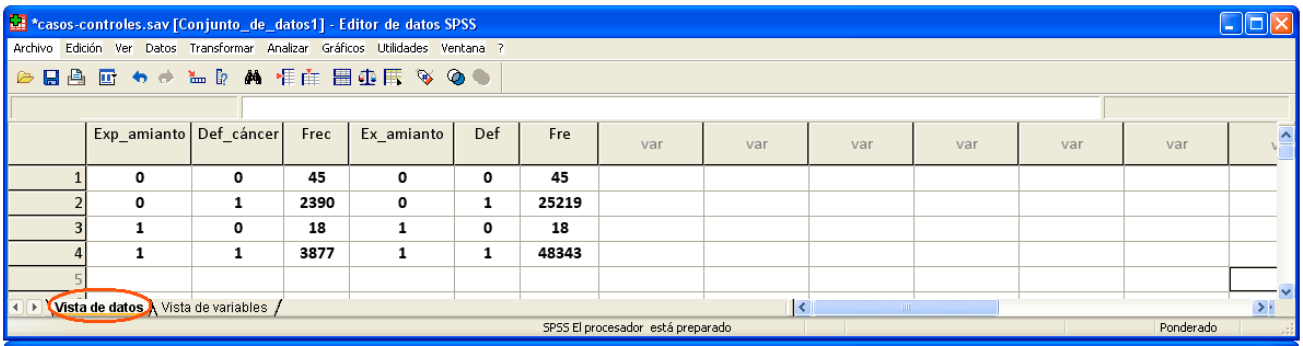
La Razón de tasas de incidencia es de 4,792319 consecuencia del diferente tiempo promedio de seguimiento entre los grupos, por lo que conviene considerar esta medida y no solo el Riesgo relativo.

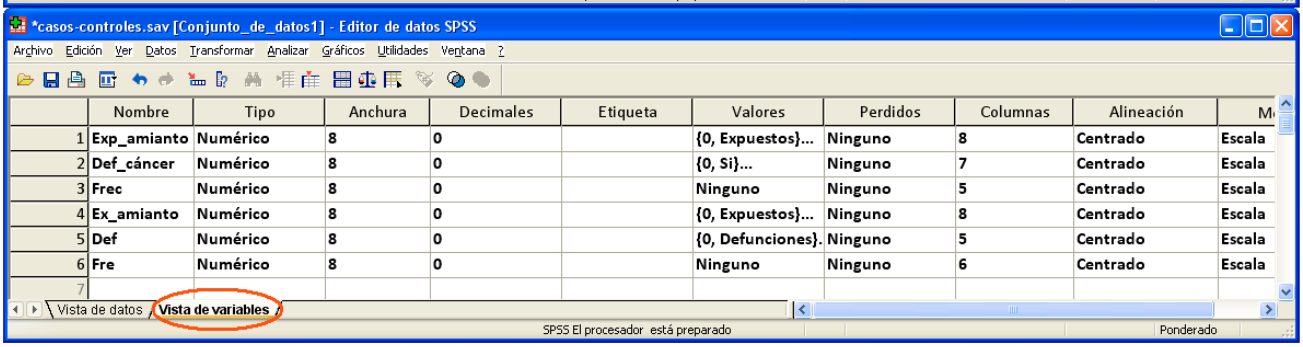
Fracción atribuible en expuestos	**0,791333**	**0,639552**	**0,879200**
Fracción atribuible poblacional	**0,565238**	**0,357586**	**0,705769**

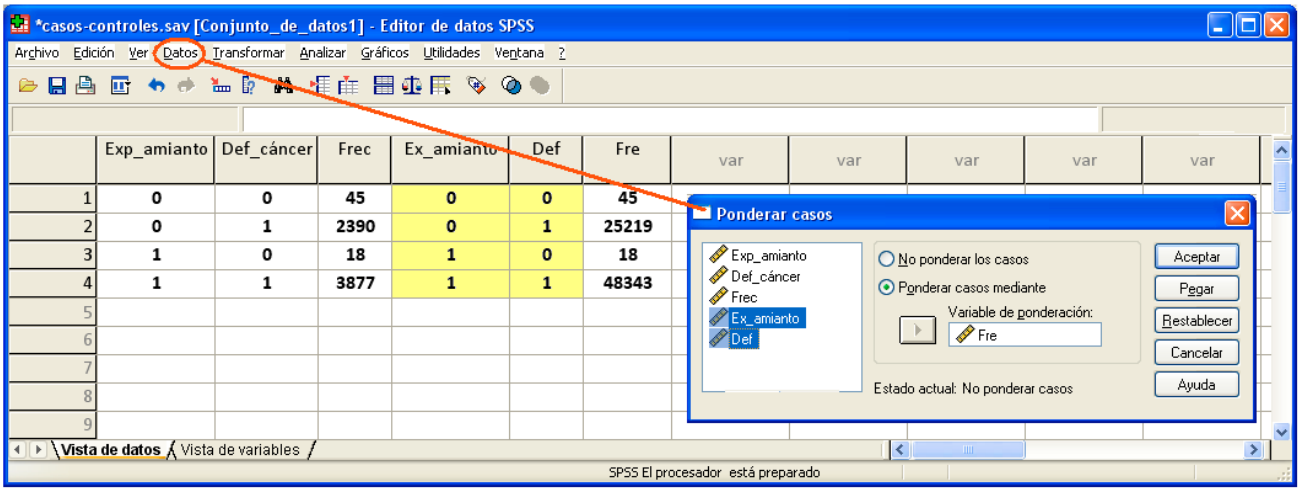
Prueba de asociación

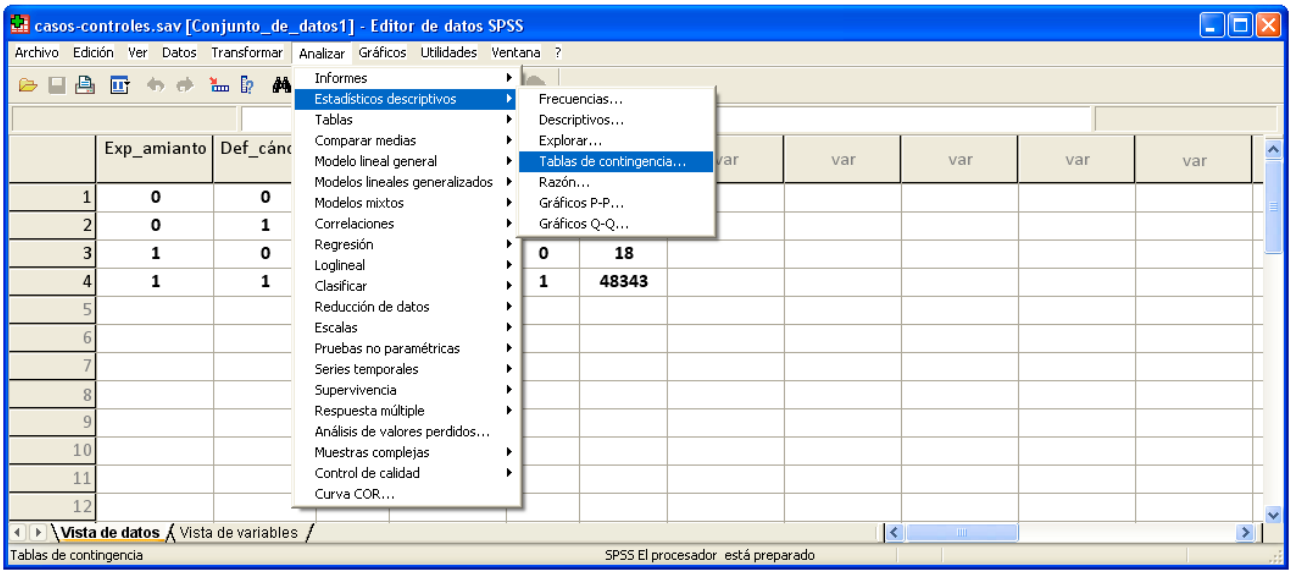
Estadístico Z	Valor p
6,0789	0,0000

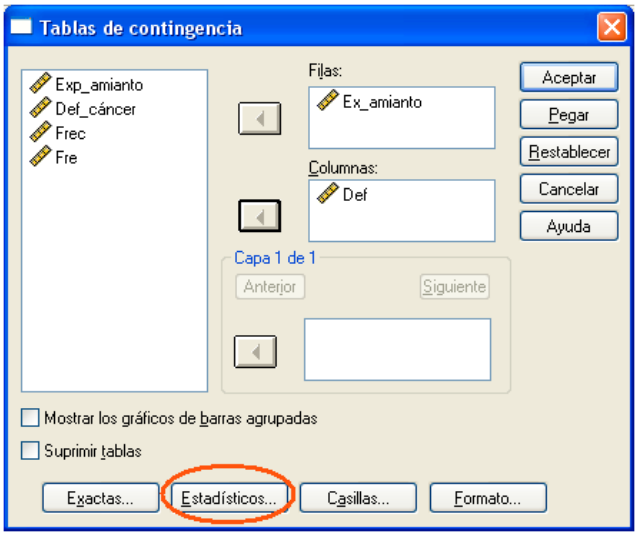
Tasas de incidencia

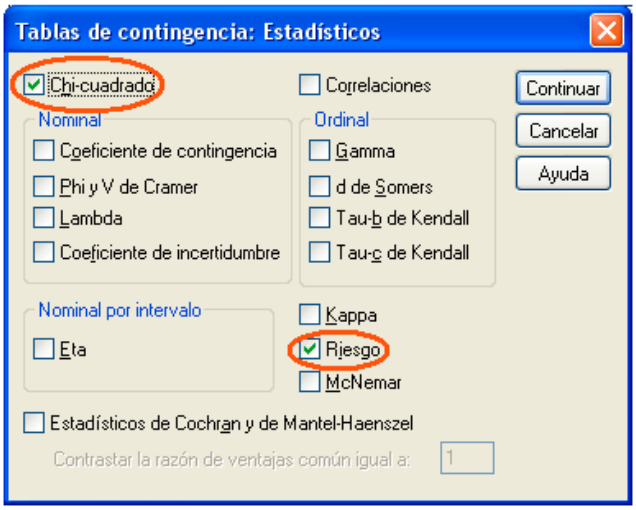












En el Visor de SPSS:

Tabla de contingencia Ex_amianto * Def

Recuento

		Def		
		Defunciones	Personas_año	Total
Ex_amianto	Expuestos	45	25219	25264
	No Expuestos	18	48343	48361
Total		63	73562	73625

La ratio entre Expuestos/No-expuestos en el grupo de sujetos con problemas de morir por amianto es $45 / 18 = 2{,}5$ y en el grupo de personas al año es $25219 / 48343 = 0{,}521668$.
El *índice de riesgo* en un diseño *caso-control* se obtiene dividiendo ambos ratios $2{,}5 / 0{,}521668 = 4{,}792320$. Este valor se interpreta de la misma forma que el índice de *riesgo relativo* (es una estimación del mismo), pero también admite otra interpretación: entre sujetos con problemas de morir por el amianto es 4,792319 más probable encontrar expuestos a no-expuestos. Un índice de riesgo de 1 indica que la probabilidad de encontrarse con el factor deseencadente es la misma en las dos cohortes estudiadas.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	38,534024[b]	1	,000		
Corrección por continuidad [a]	36,903618	1	,000		
Razón de verosimilitudes	36,045860	1	,000		
Estadístico exacto de Fisher				,000	,000
Asociación lineal por lineal	38,533501	1	,000		
N de casos válidos	73625,000				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 21,62.

Estimación de riesgo

	Valor	Intervalo de confianza al 95%	
		Inferior	Superior
Razón de las ventajas para Ex_amianto (Expuestos / No Expuestos)	4,792319	2,773735	8,279928
Para la cohorte Def = Defunciones	4,785564	2,770999	8,264755
Para la cohorte Def = Personas_año	,998590	,998043	,999138
N de casos válidos	4,792319		

La primera fila de la tabla indica que el riesgo estimado se refiere al de expuestos sobre los no-expuestos (Expuestos/No-expuestos) en un diseño de *caso-control* (*Razón de las ventajas*). Su valor 4,79232 indica que, entre los sujetos con exposición al amianto, la probabilidad (el riesgo) de encontrar sujetos expuestos al amianto es 4,792319 veces mayor que la de no encontrar sujetos expuestos.

La Razón de ventajas también puede interpretarse como una *estimación del riesgo relativo* (particularmente si la proporción de desenlaces es pequeña): El riesgo de morir es 4,792319 más entre expuestos que entre no-expuestos.
Los límites del intervalo de confianza no abarcan el 1 con lo que el riesgo es significativamente mayor.

Las dos filas siguientes ofrecen dos índices de riesgo para un diseño de cohortes (dos índices porque el desenlace que interesa evaluar puede encontrarse en cualquiera de las dos categorías de la variable).

Si el desenlace que interesa analizar es la presencia de defunciones (Cohorte con defunciones), la probabilidad o riesgo de encontrar tal desenlace entre los expuestos es 4,785564 veces mayor que la de encontrarlo entre los no-expuestos, es decir, por cada sujeto difunto entre los no-expuestos se puede encontrar 4,785564 sujetos difuntos entre los expuestos.

Si el desenlace que interesa analizar es la presencia de personas (Cohorte personas año), la probabilidad o riesgo de encontrar tal desenlace entre los expuestos es menor que entre los no-expuestos, por cada persona-año entre los no-expuestos se encuentra 0,998590 entre los expuestos.

Estudio de Casos y Controles

Los sujetos proceden típicamente de dos grupos, según sean casos (enfermedad o daño) o controles (no enfermedad o sin daño). El estudio trata de comparar los antecendentes de los enfermos (casos) de una población con los sanos (controles) de la misma población, los resultados se presentan utilizando los Odds ratio (OR: cociente entre la probabilidad de enfermar y la probabilidad de no enfermar y la razón de Odds de adquirir una enfermedad entre expuestos y entre no-expuestos.

	Casos	Controles	Total
Expuestos	a	b	a + b
No expuestos	c	d	c + d
Total	a + c	b + d	a + b + c + d

El número de controles por cada caso difiere entre un estudio y otro, en general oscila entre uno y tres, a lo sumo se toman cuatro controles por cada caso, no tiene sentido tomar más porque la potencia de la prueba no crece significativamente.

Este diseño es más eficiente que los estudios de seguimiento en términos de costo y tiempo, especialmente para enfermedades poco comunes y/o de largos períodos de incubación. Adviértase que una vez diagnosticada la enfermedad o evento, solo se necesita incluir en el estudio un número pequeño de caso, y especialmente de controles.

Otra ventaja del estudio de este diseño, respecto a los estudios de seguimiento, es la posibilidad de estudiar varias exposiciones de manera simultánea.

El estudios de casos y controles presenta dos grandes inconvientes:

(a) El sesgo de la selección que puede introducirse al elegir los controles (sin daño).

(b) Cuando se incluyen individuos en el estudio, tanto las exposiciones como el daño ya han ocurrido.

Esto dificulta establecer la precisión y similitud de criterio con que exposiciones y daños han sido cuantificados en los individuos. Presentando incluso el potencial problema de los estudios transversales donde la secuencia exposición-daño no puede conocerse adecuadamente o incluso estar invertida en algunos casos (la exposición se ha modificado como consecuencia del daño) sin que el investigador se encuentre informado.

El estudio de casos y controles permite estimar directamente las medidas de riesgo dentro de cada grupo, dado que la porporción de enfermos dentro del grupo de expuestos y en grupo de no-expuestos depende de la decisión del investigador en cuanto al número de casos y controles que subyacen en el estudio.

En otras palabras, la muestra típicamente no es representativa de la población en cuanto a la proporción enfermos/no-enfermos, eliminando con ello la posibilidad de estimar adecuadamente las tasas de enfermos entre expuestos y de enfermos que quedan libres de la exposición.

Los casos y controles se emparejan para hacer frente a diferentes factores de confusión (género, edad, tabaquismo, consumo de alcohol, etc.).
Cuando se realiza durante el análisis, los datos pueden procesarse como si el emparejamiento no se hubiera realizado, o bien a través de una tabla especial que busca comparar las diferencias entre estos *pares*.

ANÁLISIS ESTRATIFICADO

El análisis estratificado engloba a un conjunto de métodos que se sustentan en la creación de subgrupos (estratos) con el objetivo de analizar determinados problemas. Entre las utilidades más corrientes se encuentran la identificación y el control de factores de confusión y la modificación del efecto.

La estrategia general consiste en crear subgrupos (estratos) formados por las categorias de la tercera variable (confusora o modificadora del efecto) y medir el efecto (riesgo relativo, odds ratio, diferencia de tasas) en esos subgrupos.

Si se observan diferencias relevantes en el efecto en los estratos es posible que se trate de una variable modificadora. Las diferencias entre el odds ratio crudo (sin considerar los estratos) y el odds ratio combinado de Mantel-Haenszel produce un efecto de confusión:

$$OR_{crudo} = \frac{a \times d}{b \times c} \neq OR_{M-H} = \frac{\sum_{i=1}^{k} \frac{a_i \cdot d_i}{n_i}}{\sum_{i=1}^{k} \frac{b_i \cdot c_i}{n_i}} \quad \rightarrow \quad \text{Confusión}$$

Cuando el sesgo de confusión: $\left(\frac{OR_{M-H}}{OR_{crudo}} - 1 \right) . 100$ **es superior al 10% se puede afirmar que la variable por la que se estratifica produce un efecto de confusión.**

Cuando una variable es confusora debe excluirse de los resultados del estudio.

La modificación del efecto (interacción) se presenta cuando la estimación del riesgo difiere entre estratos definido por una tercera variable (por ejemplo la OR), es decir, cuando la presencia o diferentes valores de una tercera variable modifican el efecto de la exposición sobre la enfermedad.
El fenómeno de la modificación del efecto hay que describirlo siempre. En esta línea, en caso de rechazar la hipótesis nula H_o de homogeneidad de efectos $(p\_valor < 0,05)$, los resultados deben presentarse de forma estratificada para cada uno de los niveles de la variable modificadora del efecto.

TABLAS 2x 2 ESTRATIFICADAS

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 sé 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 sí lo habían hecho y 173 no).
Los datos quedan reflejados en las dos tablas siguientes:

Casos y Controles

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

Casos y Controles emparejados

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

$$OR = \frac{255.268}{500.487} = 0,281$$

$$OR = \frac{160.173}{327.95} = 0,891$$

$$RR_{\text{No Lactancia/No Lactancia}} = \frac{d.(a+b)}{b.(c+d)} = \frac{268.742}{487.768} = 0,532$$

$$RR_{\text{Lactancia/No Lactancia}} = \frac{d.(a+b)}{b.(c+d)} = \frac{173.255}{95.500} = 0,929$$

$$RR_{\text{No Lactancia/Lactancia}} = \frac{c.(a+b)}{a.(c+d)} = \frac{500.742}{255.768} = 1,894$$

$$RR_{\text{Lactancia/Lactancia}} = \frac{c.(a+b)}{a.(c+d)} = \frac{327.255}{160.500} = 1,042$$

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).

El riesgo relativo (RR) es el cociente entre el riesgo en el grupo con factor de exposición o factor de riesgo y el riesgo en el grupo de referencia (que no tiene el factor de exposición) como índice de asociación.

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 0,532 más alta que entre las mujeres no expuestas.

En el emparejamiento de casos y controles, la proporción aumenta hasta 0,929.

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:

$$IC_{1-\alpha}(RR) = RR \,.\, e^{\pm\, z_{\alpha/2}\cdot\sqrt{V(\ln RR)}} = \left(RR \,.\, e^{-\, z_{\alpha/2}\cdot\sqrt{V(\ln RR)}} \;,\; RR \,.\, e^{\,z_{\alpha/2}\cdot\sqrt{V(\ln RR)}} \right)$$

$$V\left[\ln RR\right] = \frac{1}{a} - \frac{1}{a+c} + \frac{1}{b} - \frac{1}{b+d} = \frac{1}{255} - \frac{1}{755} + \frac{1}{487} - \frac{1}{755} = 0{,}003326$$

$$IC(\text{RR caso - control})_{\text{No Lactancia/No Lactancia}} = \left[0{,}532 \,.\, e^{-1{,}96\,\cdot\,\sqrt{0{,}003326}} \;,\; 0{,}532 \,.\, e^{\,1{,}96\,\cdot\,\sqrt{0{,}003326}} \right] =$$

$$= \left[0{,}476 \;,\; 0{,}593 \right]$$

$$IC(\text{RR emparejados})_{\text{Lactancia/No Lactancia}} = \left[0{,}929 \,.\, e^{-1{,}96\,\cdot\,\sqrt{0{,}00912}} \;,\; 0{,}929 \,.\, e^{1{,}96\,\cdot\,\sqrt{0{,}00912}} \right] =$$

$$= \left[0{,}761 \;,\; 1{,}134 \right]$$

siendo $V\left[\ln RR\right] = \frac{1}{a} - \frac{1}{a+c} + \frac{1}{b} - \frac{1}{b+d} = \frac{1}{160} - \frac{1}{255} + \frac{1}{95} - \frac{1}{268} = 0{,}00912$

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*):

$$OR = \frac{n_{11} \,.\, n_{22}}{n_{12} \,.\, n_{21}} = \frac{a.d}{b.c}$$

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:

$$IC_{1-\alpha}\,(OR) = \left[OR \cdot e^{-z_{\alpha/2} \cdot \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}} \;,\; OR \cdot e^{z_{\alpha/2} \cdot \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}} \right]$$

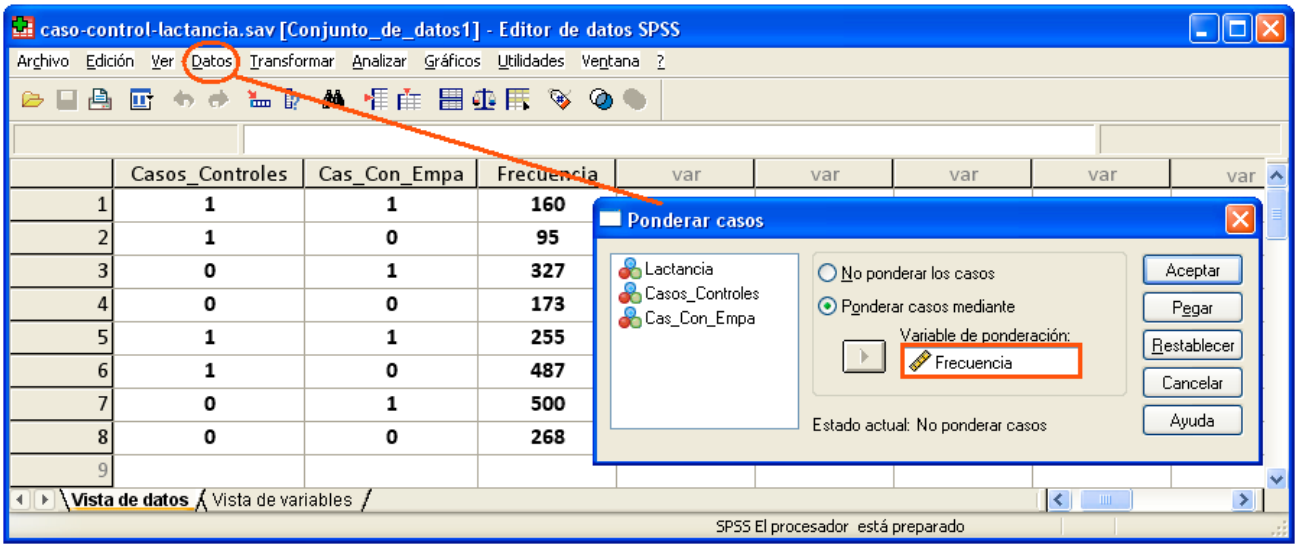
$$IC\,(OR\ \text{No Lactancia})_{\text{No Lactancia/Lactancia}} = \left[0{,}281 \cdot e^{-1{,}96 \cdot \sqrt{\frac{1}{255}+\frac{1}{487}+\frac{1}{500}+\frac{1}{268}}} \;,\; 0{,}281 \cdot e^{1{,}96 \cdot \sqrt{\frac{1}{255}+\frac{1}{487}+\frac{1}{500}+\frac{1}{268}}} \right] =$$

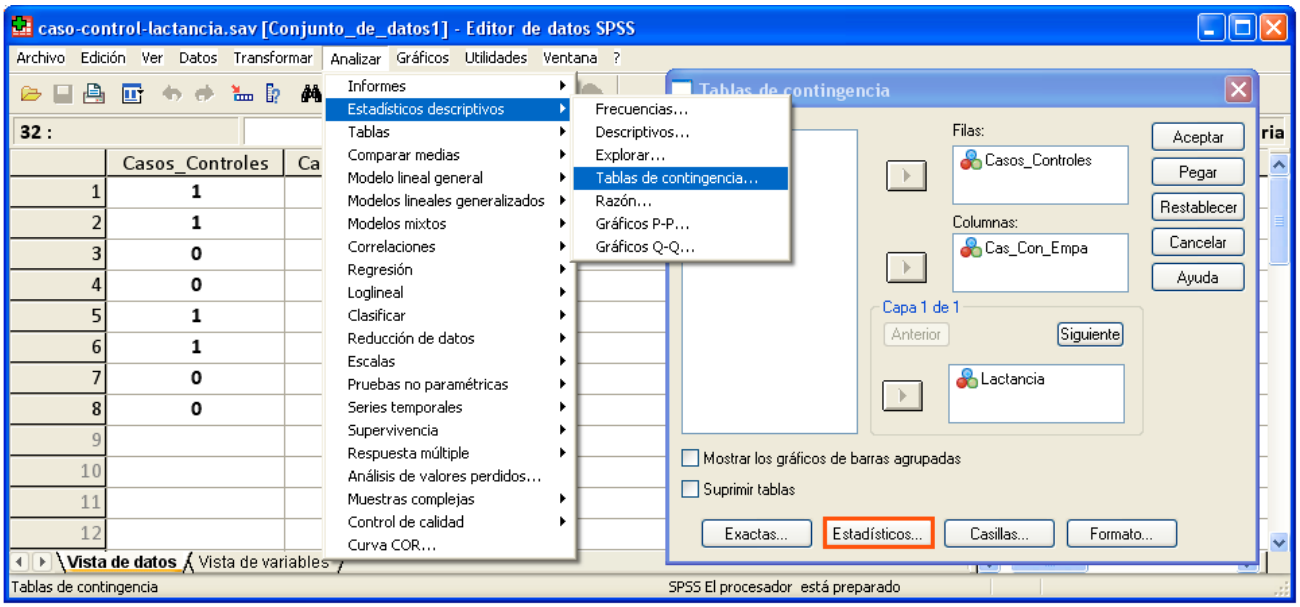
$$= [0{,}227 \;,\; 0{,}347]$$

$$IC\,(OR\ \text{Lactancia})_{\text{No Lactancia/Lactancia}} = \left[0{,}891 \cdot e^{-1{,}96 \cdot \sqrt{\frac{1}{160}+\frac{1}{95}+\frac{1}{327}+\frac{1}{173}}} \;,\; 0{,}891 \cdot e^{1{,}96 \cdot \sqrt{\frac{1}{160}+\frac{1}{95}+\frac{1}{327}+\frac{1}{173}}} \right] =$$

$$= [0{,}651 \;,\; 1{,}219]$$

Se introducen los datos de las variables dicotómicas en SPSS:





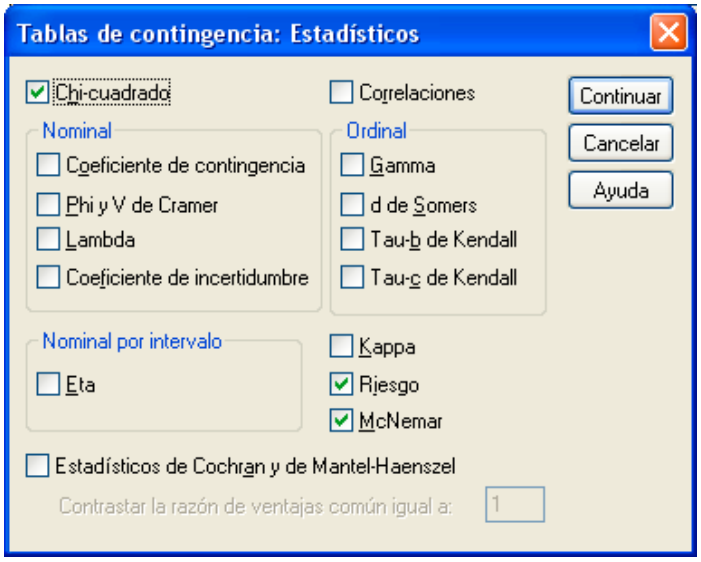


Tabla de contingencia Casos_Controles * Cas_Con_Empa * Lactancia

Estadísticos | Recuento

Lactancia			Cas_Con_Empa		Total
			No Lactancia	Lactancia	
No Lactancia	Casos_Controles	No Lactancia	268	500	768
		Lactancia	487	255	742
	Total		755	755	1510
Lactancia	Casos_Controles	No Lactancia	173	327	500
		Lactancia	95	160	255
	Total		268	487	755

Pruebas de chi-cuadrado

Lactancia		Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
No Lactancia	Chi-cuadrado de Pearson	142,6224[b]	1	,000		
	Corrección por continuidad[a]	141,3956	1	,000		
	Razón de verosimilitudes	144,9600	1	,000		
	Estadístico exacto de Fisher				,000	,000
	Asociación lineal por lineal	142,5280	1	,000		
	Prueba de McNemar				,7025[c]	
	N de casos válidos	1510				
Lactancia	Chi-cuadrado de Pearson	,5199[d]	1	,4709		
	Corrección por continuidad[a]	,4104	1	,5218		
	Razón de verosimilitudes	,5181	1	,4717		
	Estadístico exacto de Fisher				,4706	,2604
	Asociación lineal por lineal	,5192	1	,4712		
	Prueba de McNemar				,000[c]	
	N de casos válidos	755				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 371,00.

c. Utilizada la distribución binomial

d. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 90,52.

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).
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 no expuestos o No Lactancia, el p_value es menor que 0,05, se rechaza la hipótesis nula, afirmando que las variables No lactancia y cáncer de mama no son independientes, esto es, hay una asociación entre ellas.

En el caso de los expuestos o Lactancia, p_value es mayor que 0,05, se acepta la hipótesis nula, y se concluye que las variables lactancia y cáncer de mama son independientes, es decir, no existe una aociaciación entre ellas.

Estimación de riesgo

			Intervalo de confianza al 95%	
Lactancia		Valor	Inferior	Superior
No Lactancia	Razón de las ventajas para Casos_Controles (No Lactancia / Lactancia)	,281	,227	,347
	Para la cohorte Cas_Con_Empa = No Lactancia	,532	,476	,593
	Para la cohorte Cas_Con_Empa = Lactancia	1,894	1,693	2,119
	N de casos válidos	1510		
Lactancia	Razón de las ventajas para Casos_Controles (No Lactancia / Lactancia)	,891	,651	1,219
	Para la cohorte Cas_Con_Empa = No Lactancia	,929	,761	1,134
	Para la cohorte Cas_Con_Empa = Lactancia	1,042	,930	1,168
	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 No 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 y no es significativo porque su intervalo de confianza cubre el uno, indicando que la probabilidad 0,891 de encontrarse con el cáncer de mama en los grupos estudiados es la misma.

Observando la segunda fila, para la cohorte Emparejamiento = No lactancia , la proporción de cáncer de mama entre las mujeres no lactantes es RR = 0,532 veces más alta que entre las mujeres lactantes. En el emparejamiento, la proporción aumenta hasta 0,929 , su intervalo de confianza cubre el 1 por lo que no es significativo, indicando que la probabilidad de encontrar cáncer de mama es la misma en madres lactantes y no lactantes.

TABLAS GENERALES M x N

En una tabla de contingencia M x N los sujetos de una muestra se clasifican, respectivamente, respecto a dos variables cualitativas con M y N categorías. La clasificación debe de ser exhaustiva y mutuamente exclusiva, es decir, todos los individuos pueden asignarse a una y solo una categoría.

X \ Y	y_1	y_2	...	y_j	...	y_m	$\sum_{i=1}^{k} O_{i\bullet}$
x_1	O_{11}	O_{12}	...	O_{1j}	...	O_{1m}	$O_{1\bullet}$
x_2	O_{21}	O_{22}	...	O_{2j}	...	O_{2m}	$O_{2\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	O_{i1}	O_{i2}	...	O_{ij}	...	O_{im}	$O_{i\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_k	O_{k1}	O_{k2}	...	O_{kj}	...	O_{km}	$O_{k\bullet}$
$\sum_{j=1}^{m} O_{\bullet j}$	$O_{\bullet 1}$	$O_{\bullet 2}$	...	$O_{\bullet j}$	...	$O_{\bullet m}$	N

Generalmente, la cuestión más importante que se plantea ante una tabla de contingencia es si las variables son independientes o no lo son. Para contestar a esta pregunta se plantean diversos contrastes de hipótesis: La prueba Chi-cuadrado de Pearson, la prueba de razón de verosimilitudes y, para tablas 2x2 la prueba Chi-cuadrado con corrección de Yates y la prueba exacta de Fisher.

La prueba Chi-cuadrado de Pearson es basa en la hipótesis nula de que en la tabla no hay discrepancias entre las frecuencias observadas y las frecuencias esperadas, el estadístico de este contraste sigue aproximadamente una distribución Chi-cuadrado con $(k-1)\,x\,(m-1)$ grados de libertad. Analizando la validez de la aproximación, Cochran recomienda que sólo se utilice esta prueba cuando a lo sumo un 20% de las celdas presentan frecuencia esperada menor que 5 y ninguna frecuencia esperada es menor que 1.

En las tablas para mejorar la aproximación por continuidad se incorpora la corrección de Yates. En muestras grandes se obtienen resultados similares con y sin corrección. En el caso de muestras pequeñas se recomienda utilizar la prueba exacta de Fisher, esta prueba calcula la probabilidad exacta de obtener los resultados observados si las dos variables son independientes y los totales marginales son fijos.

La prueba de Razón de verosimilitudes, basada en la máxima verosimilitud, es una alternativa a la prueba Chi-cuadrado para contrastar la hipótesis nula de que las dos variables son independientes. El estadístico de la prueba, sigue una

distribución Chi-cuadrado con $(k-1)\times(m-1)$ grados de libertad, compara la probabilidad de los datos observados con la probabilidad de los datos esperados, en caso de ser cierta la hipótesis nula de independencia. En consecuencia, valores altos del estadístico son indicativos de asociación entre las variables. La distribución del estadístico también es aproximada, por lo que no es apropiado si el tamaño de la muestra es pequeño.
Hay varias medidas que cuantifican la intensidad de asociación entre dos variables que determinan la tabla de contingencia, algunas son válidas en general para variables nominales; otras son específicas de variables ordinales.

MEDIDAS DE ASOCIACIÓN DE VARIABLES NOMINALES:

Las medidas nominales sólo aprovechan información nominal, únicamente informan del grado de asociación existente, no de la dirección o naturaleza de tal asociación.
Para los datos nominales (sin orden intrínseco) se puede seleccionar los estadísticos: Phi y V de Cramer, Coeficiente de contingencia, Lambda (simétricas y asimétricas), Tau de Goodman y Kruskal, y Coeficiente de incertidumbre.
En programas estadísticos, las medidas (Phi, V de Cramer, Coeficiente de contingencia, Lambda, Tau de Goodman-Kruskal, Coeficiente de incertidumbre) figuran acompañadas de su nivel crítico (*Sig. aproximada, p_valor*) , que cuando es menor de 0,05 conduce a rechazar la hipótesis nula de independencia con un nivel de confianza del 95%, concluyendo que las variables en estudio están relacionadas.

- **Coeficiente Phi de Pearson: Toma valores en el intervalo $0 \le \phi \le 1$, el valor 0 indica que no hay dependencia, el valor 1 es cuando la dependencia es directa y perfecta. Se define: $\phi = \sqrt{\frac{\chi^2_{exp}}{N}}$**

  En tablas de contingencia 2x2 adopta valores entre 0 y 1, y su valor es idéntico al del coeficiente de correlación de Pearson.
  En tablas en las que una de las variables tiene más de dos niveles, phi puede tomar valores mayores que 1 (pues el valor de Chi-cuadrado puede ser mayor que el tamaño muestral).

- **Coeficiente de contingencia (C): Toma valores entre 0 y 1. Vale 0 en caso de independencia completa, nunca puede tomar el valor máximo 1, porque incluso en caso de asociación completa, el valor C depende del número de filas y de columnas de la tabla. Se define: $C = \sqrt{\frac{\chi^2_{exp}}{\chi^2_{exp} + N}}$ Si el número de filas y de columnas es el mismo (k), el valor máximo de C se obtiene: $C_{máx} = \sqrt{\frac{k-1}{k}}$**

- **Coeficiente Lambda de Goodman y Kruskal:** Estadístico utilizado para determinar si partiendo de los resultados de una de las variables se pueden predecir los resultados de la otra variable.
  Lambda toma valores entre 0 y 1, $0 \leq \lambda \leq 1$, donde 0 indica que la variable independiente no contribuye en absoluto en reducir el error de predicción (independencia entre las variables) : $\lambda = 0 \not\Rightarrow$ 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 la otra.

  El valor 1 indica que se ha conseguido por completo reducir el error de predicción (total dependencia), es decir, que la variable independiente predice con toda precisión a qué categoría de la variable dependiente pertenecen los casos clasificados.

  Lambda igual a 1 implicaría (extremo que no se suele dar) que la variable independiente consigue reducir a 0 el error de la variable dependiente.
  Suponiendo que Y es el factor explicado y X el explicativo , se evalúa la capacidad de X para predecir Y mediante:

$$\lambda_Y = \frac{\sum_{i=1}^{k} \text{máx}_j\, O_{ij} - \text{máx}_j\, O_{\bullet j}}{N - \text{máx}_j\, O_{ij}} \qquad 0 \leq \lambda_Y \leq 1$$

  De forma análoga, cuando X es el factor explicado e Y el explicativo, se evalúa

$$\lambda_X = \frac{\sum_{j=1}^{m} \text{máx}_i\, O_{ij} - \text{máx}_i\, O_{i\bullet}}{N - \text{máx}_i\, O_{ij}} \qquad 0 \leq \lambda_X \leq 1$$

  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).
  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:

$$\lambda_X = \frac{\sum_{i=1}^{k} \text{máx}_j\, O_{ij} + \sum_{j=1}^{m} \text{máx}_j\, O_{ij} - \text{máx}_i\, O_{i\bullet} - \text{máx}_j\, O_{\bullet j}}{2N - \text{máx}_i\, O_{i\bullet} - \text{máx}_j\, O_{\bullet j}} \qquad 0 \leq \lambda \leq 1$$

- **Coeficiente V de Cramer:** Corrige el problema de la dependencia que tiene el coeficiente de contingencia del número de filas y columnas.
  Para las tablas 2x2 toma valores entre -1 y 1 y, en otro caso, varía entre 0 y 1, alcanzando el -1 o 0, respectivamente, en caso de independencia completa, y

el valor 1 en caso de asociación completa. En estas tablas el valor de *phi* (ϕ) y el valor de V_{Cramer} es el mismo.

Las pruebas del Coeficiente de contingencia (C) y el Coeficiente V de Cramer se basan en el estadístico *Chi-cuadrado de Pearson* y son *simétricas* (no dependen de cuál es la variable filas y cuál es la variable columnas), en caso de intercambiar las variables se obtiene el mismo valor para el coeficiente.

La *V de Cramer* incluye una pequeña modificación de phi:

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

donde k y m se refiere al número de filas y de columnas.

En consecuencia, miden el grado de asociación entre las variables, pero no el grado en que una variable predice a la otra. Tiende a subestimar el grado de asociación entre las variables.

- **Tau de Goodman y Kruskal:** Es una medida de asociación asimétrica, que permite considerar una de las variables como independiente y a la otra como dependiente. Se presenta en dos situaciones: (a) Considerando las filas como categorías de la variable dependiente (Filas/Columnas). (b) Considerando las columnas como categorías de la variable dependiente (Columnas/Filas). Ambos coeficientes toman valores entre 0 y 1, serán más cercanos a 1 cuanto mayor sea la capacidad de predecir sin error la variable dependiente, mientras que un valor de 0 indica que la variable independiente no tiene capacidad de predecir a la variable dependiente.

$$\tau = \frac{\frac{1}{N}\sum_{i=1}^{k}(N - O_{i\bullet}) \cdot O_{i\bullet} - \sum_{j=1}^{m}\frac{(O_{\bullet j} - O_{ij}) \cdot O_{ij}}{O_{\bullet j}}}{\frac{1}{N}\sum_{i=1}^{k}(N - O_{i\bullet}) \cdot O_{i\bullet}}$$

- **Coeficiente de incertidumbre:** Es una medida semejante a *Lambda* y *Tau* en cuánto a su concepción de la asociación de las variables, en relación a su capacidad predictiva y la disminución 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 (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 la variable independiente y la variable dependiente).

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

$$I(X) = -\sum_{i=1}^{k}\left(\frac{O_{i\bullet}}{N}\right).\mathrm{Ln}\left(\frac{O_{i\bullet}}{N}\right) \quad I(Y) = -\sum_{j=1}^{m}\left(\frac{O_{\bullet j}}{N}\right).\mathrm{Ln}\left(\frac{O_{\bullet j}}{N}\right) \quad I(XY) = -\sum_{i=1}^{k}\sum_{j=1}^{m}\left(\frac{O_{ij}}{N}\right).\mathrm{Ln}\left(O_{ij}\right)$$

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

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

MEDIDAS DE ASOCIACIÓN DE VARIABLES ORDINALES:

Tienen en cuenta la ordenación, toman valores positivos cuando una de las variables tiende a aumentar a medida que lo hace la otra, y valores negativos cuando los valores altos de una variable se asocian con valores bajos de la otra. Para las variables ordinales, en las que tanto las filas como las columnas contienen valores ordenados, se pueden seleccionar los estadísticos: Gamma (orden cero para tablas de doble clasificación y condicional para tablas cuyo factor de clasificación va de 3 a 10), Tau-b de Kendall, Tau-c de Kendall y d de Somers.

- **.Tau-b de Kendall: Se basa en el número de concordancias, discordancias y empates entre pares de casos. Un par es concordante si los valores de ambas variables para un caso son menores/mayores que los correspondientes para el otro caso. Un par es discordante si ocurre lo contrario.**

  Ejemplo: **En una muestra de pacientes se tiene la relación entre la variable clase social CS con tres categorías en orden creciente y la variable diagnóstico D con cuatro categorías de menor a mayor gravedad.**
  Un par de pacientes con valores: $CS=2, D=2$ y $CS=3, D=4$ es concordante, mientras que un par discordante estaría formado por dos pacientes con valores $CS=2, D=2$ y $CS=3, D=1$.

  El *coeficiente Tau-b de Kendall* puede tomar valores entre -1 y 1 alcanzando los extremos sólo en el caso de tablas cuadradas. Si el predominio de los pares es concordante, el valor es próximo a 1 e indica que la asociación es positiva. Si el predominio de los pares es discordante, el valor se acerca a -1 y la asociación es negativa. El valor 0 indica que no hay relación entre las variables, ocurriendo cuando los pares concordantes y discordantes son igualmente probables (empate).

- **Tau-c de Kendall: Es una variante de la Tau-b, se diferencia en que puede alcanzar los valores mínimo y máximo, -1 y 1, en tablas de cualquier dimensión, salvo pequeñas discrepancias cuando el tamaño de la muestra no es un múltiplo mínimo entre M y N (número de filas y columnas, respectivamente).**

- **Gamma de Goodman y Kruskal:** Se basa en pares concordantes y discordantes, toma valores entre -1 y 1. El valor 0 se alcanza en el caso de que las variables sean independientes y la asociación es tanto mayor cuanto más se aproxima gamma a -1 ó a 1.

- **d de Somers:** Medida asimétrica que, por tanto, permite realizar un análisis de relación entre dos variables tomando una de ellas como dependiente, por lo que se obtendrán dos índices, como ocurre en el caso de la Tau de Goodman y Kruskal, uno cuando la variable independiente es la situada en las filas y otro en el caso de que dicha variable sea la de columnas.
  Toma valores entre -1 y 1.
  Para las medidas de asociación con datos ordinales se realiza una prueba de significación:

- **Coeficiente de correlación por rangos de Spearman:** Los valores de cada una de las variables se clasifican de menor a mayor y se calcula el coeficiente de correlación de Pearson en base a los rangos. Los valores varían entre -1 y 1, el valor 0 indica que no existe ninguna relación lineal entre las variables.

NOMINAL POR INTERVALO:

Cuando una variable es categórica y la otra cuantitativa se selecciona el estadístico *Eta*. La variable categórica debe codificarse numéricamente.

- **Eta:** Es una medida de asociación con valor comprendido entre 0 y 1. El valor 0 indica que no hay asociación entre las variables de fila y columna, los valores cercanos a 1 indican que hay una gran relación entre las variables.
  Es un estadístico apropiado para una variable dependiente medida en escala de intervalos (por ejemplo, ingresos) y una variable independiente con un número limitado de categorías (por ejemplo, género).
  Se calculan dos valores de eta: uno trata la variable de las filas como una variable de intervalo; el otro trata la variable de las columnas como una variable de intervalo.

- **Kappa:** Estadístico que mide el acuerdo entre las evaluaciones de dos jueces cuando ambos están valorando el mismo objeto. Toma valores entre 0 y 1, sólo está disponible para las tablas cuadradas (donde ambas variables tienen el mismo número de categorías).
  Un valor igual a 1 indica un acuerdo perfecto, un valor igual a 0 indica que el acuerdo no es mejor que el que se obtendría por azar.

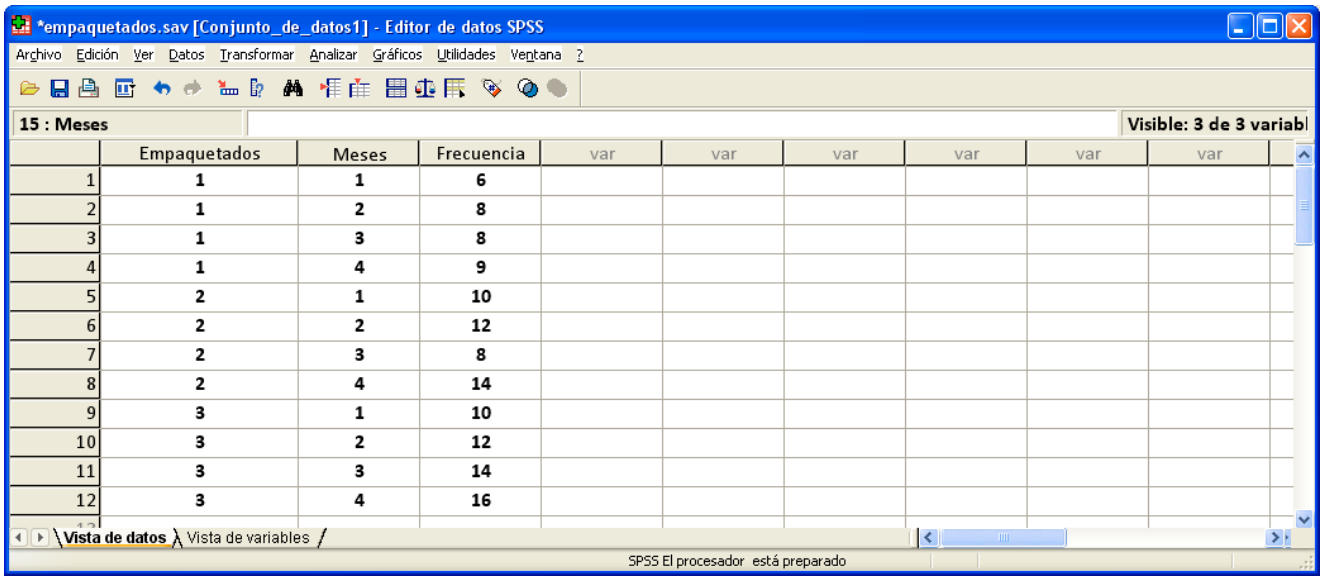
- **Riesgo:** Para las tablas 2x2 es una medida de asociación entre la presencia de un factor y la ocurrencia del evento. Si el intervalo de confianza para el estadístico incluye un valor de 1, no se puede asumir que el factor está asociado con el evento. Cuando la ocurrencia del factor es rara, se puede utilizar la razón de las ventajas (odds ratio) como estimación del riesgo relativo.

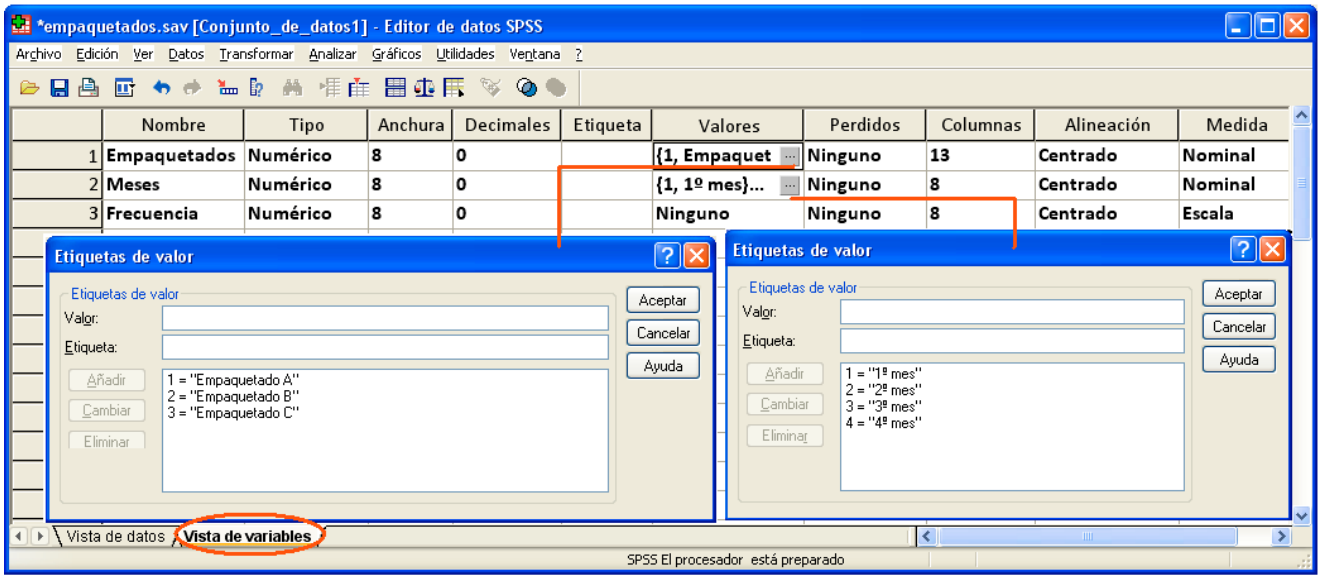
- **McNemar:** **Prueba no paramétrica para dos variables dicotómicas relacionadas. Contrasta los cambios en las respuestas utilizando la distribución de Chi-cuadrado. Es útil para detectar cambios en las respuestas debidas a la intervención experimental en los diseños del tipo "antes-después". Para las tablas cuadradas de mayor orden se informa de la prueba de simetría de McNemar-Bowker.**

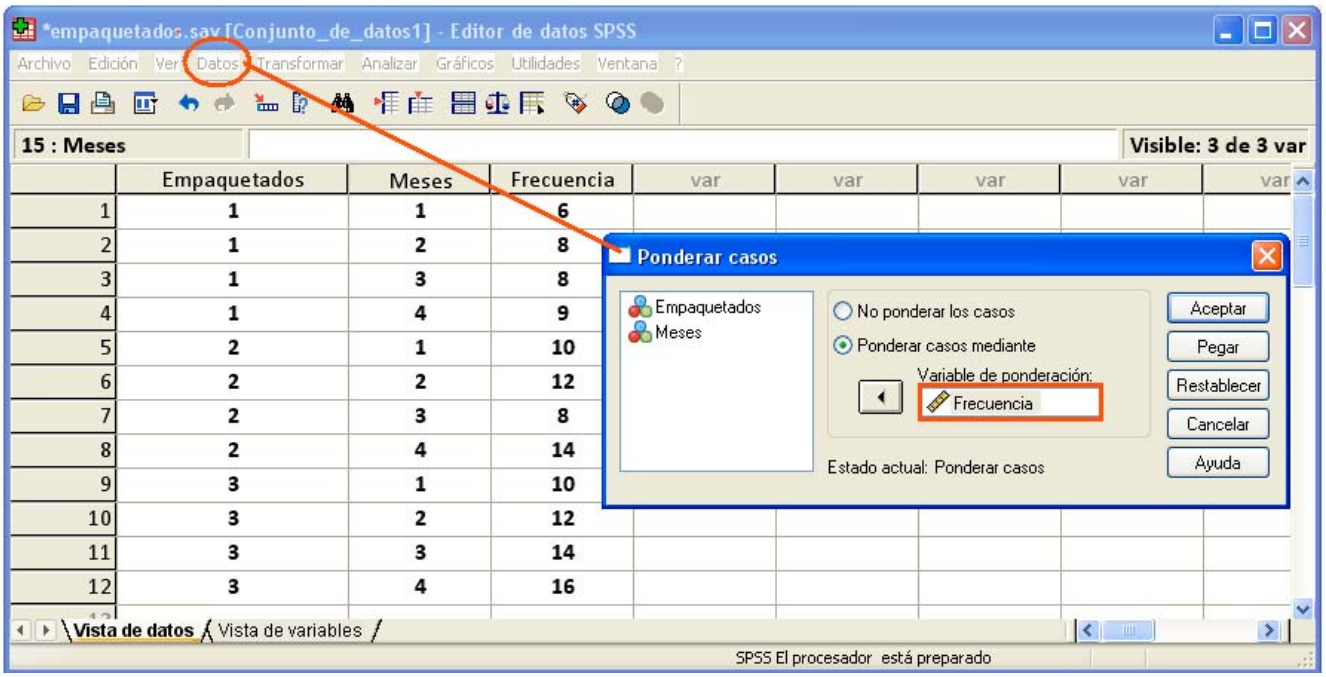
- **Cochran y Mantel-Haenszel:** **Los estadísticos de Cochran y Mantel-Haenszel pueden utilizarse para contrastar la independencia entre una variable de factor dicotómica y una variable de respuesta dicotómica, condicionada por los patrones en las covariables, los cuales vienen definidos por la variable o variables de las capas (variables de control).**
  Mientras que otros estadísticos se calculan capa por capa, los estadísticos de Cochran y Mantel-Haenszel se calculan una sola vez para todas las capas.

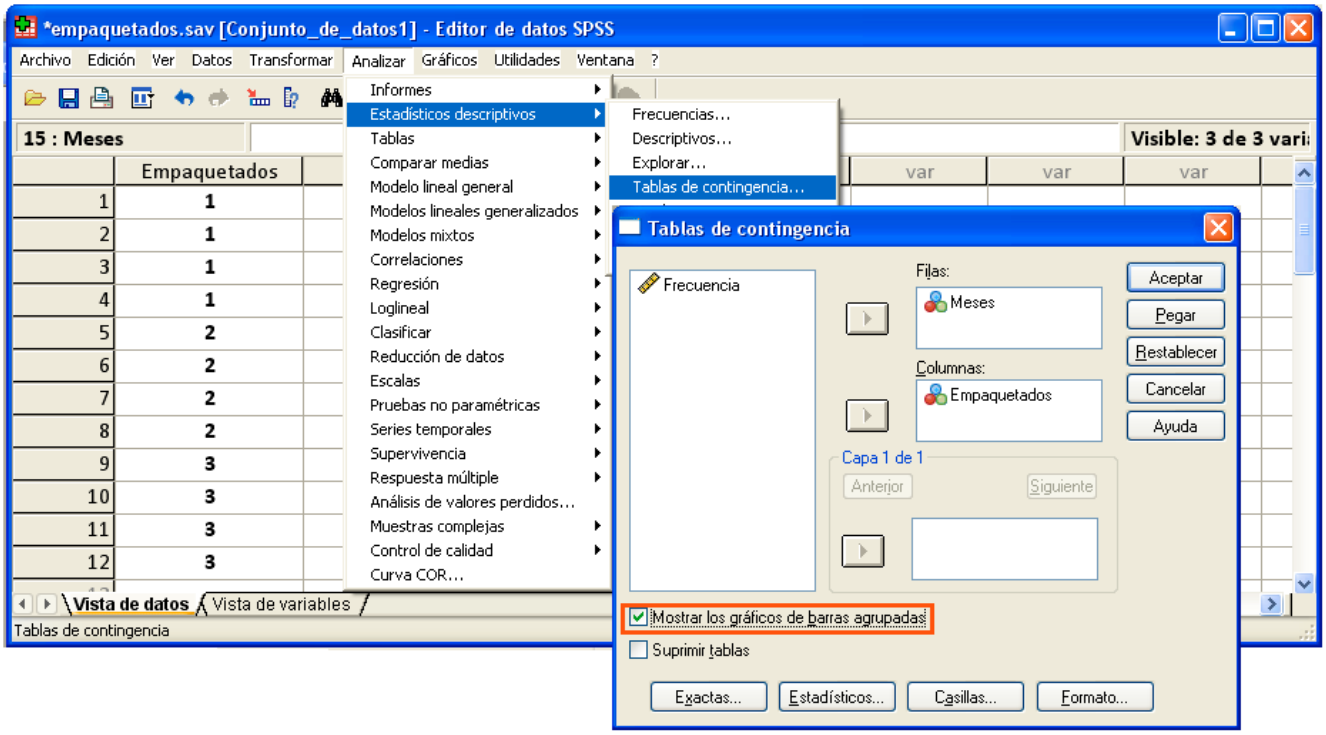
Tres métodos de empaquetado de tomates fueron probados durante un período de cuatro meses; se hizo un recuento del número de kilos por 1000 que llegaron estropeados, obteniéndose la tabla adjunta. Con un nivel de significación de 0,05, ¿tienen los tres métodos la misma eficacia?.

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

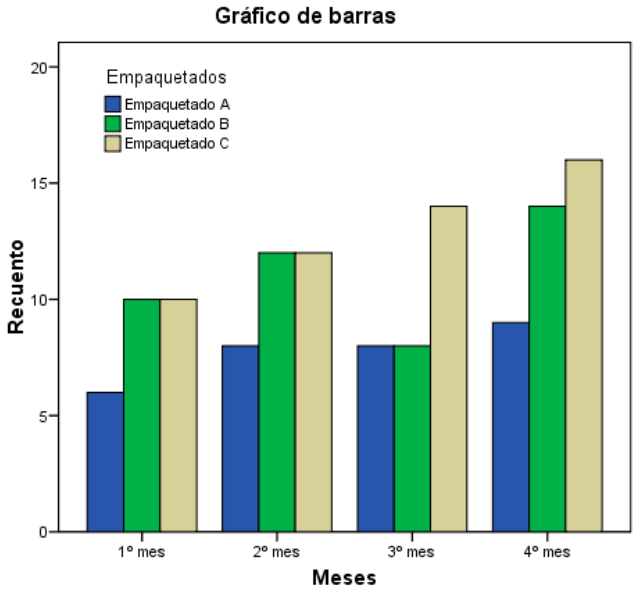



*empaquetados.sav [Conjunto_de_datos1] - Editor de datos SPSS
Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?
Nombre Tipo Anchura Decimales Etiqueta Valores Perdidos Columnas Alineación Medida
1 Empaquetados Numérico 8 0 {1, Empaquet Ninguno 13 Centrado Nominal
2 Meses Numérico 8 0 {1, 1º mes}... Ninguno 8 Centrado Nominal
3 Frecuencia Numérico 8 0 Ninguno Ninguno 8 Centrado Escala
Etiquetas de valor
Etiquetas de valor
Valor:
Etiqueta:
Añadir
Cambiar
Eliminar
1 = "Empaquetado A"
2 = "Empaquetado B"
3 = "Empaquetado C"
Aceptar
Cancelar
Ayuda
1 = "1º mes"
2 = "2º mes"
3 = "3º mes"
4 = "4º mes"
Vista de datos
Vista de variables
SPSS El procesador está preparado


*empaquetados.sav [Conjunto_de_datos1] - Editor de datos SPSS
Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?
15 : Meses
Visible: 3 de 3 var
Empaquetados Meses Frecuencia
1 1 1 6
2 1 2 8
3 1 3 8
4 1 4 9
5 2 1 10
6 2 2 12
7 2 3 8
8 2 4 14
9 3 1 10
10 3 2 12
11 3 3 14
12 3 4 16
Ponderar casos
Empaquetados
Meses
No ponderar los casos
Ponderar casos mediante
Variable de ponderación:
Frecuencia
Estado actual: Ponderar casos
Aceptar
Pegar
Restablecer
Cancelar
Ayuda
Vista de datos
Vista de variables
SPSS El procesador está preparado


*empaquetados.sav [Conjunto_de_datos1] - Editor de datos SPSS
Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?
15 : Meses
Visible: 3 de 3 vari
Empaquetados
1 1
2 1
3 1
4 1
5 2
6 2
7 2
8 2
9 3
10 3
11 3
12 3
Informes
Estadísticos descriptivos
Tablas
Comparar medias
Modelo lineal general
Modelos lineales generalizados
Modelos mixtos
Correlaciones
Regresión
Loglineal
Clasificar
Reducción de datos
Escalas
Pruebas no paramétricas
Series temporales
Supervivencia
Respuesta múltiple
Análisis de valores perdidos...
Muestras complejas
Control de calidad
Curva COR...
Frecuencias...
Descriptivos...
Explorar...
Tablas de contingencia...
Tablas de contingencia
Frecuencia
Filas:
Meses
Columnas:
Empaquetados
Capa 1 de 1
Anterior
Siguiente
Aceptar
Pegar
Restablecer
Cancelar
Ayuda
Mostrar los gráficos de barras agrupadas
Suprimir tablas
Exactas...
Estadísticos...
Casillas...
Formato...
Vista de datos
Vista de variables
Tablas de contingencia

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.



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.

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

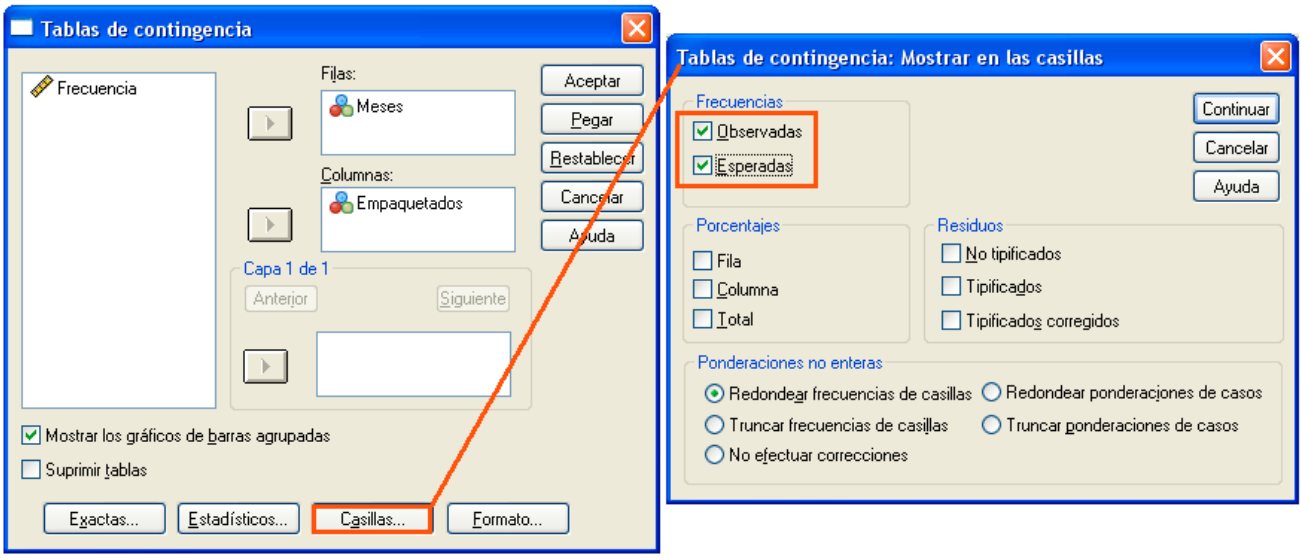


Tabla de contingencia Meses * Empaquetados

			Empaquetados			
			Empaquetado A	Empaquetado B	Empaquetado C	Total
Meses	1º mes	Recuento	6	10	10	26
		Frecuencia esperada	6,35	9,01	10,65	26
	2º mes	Recuento	8	12	12	32
		Frecuencia esperada	7,81	11,09	13,10	32
	3º mes	Recuento	8	8	14	30
		Frecuencia esperada	7,32	10,39	12,28	30
	4º mes	Recuento	9	14	16	39
		Frecuencia esperada	9,52	13,51	15,97	39
Total		Recuento	31	44	52	127
		Frecuencia esperada	31	44	52	127

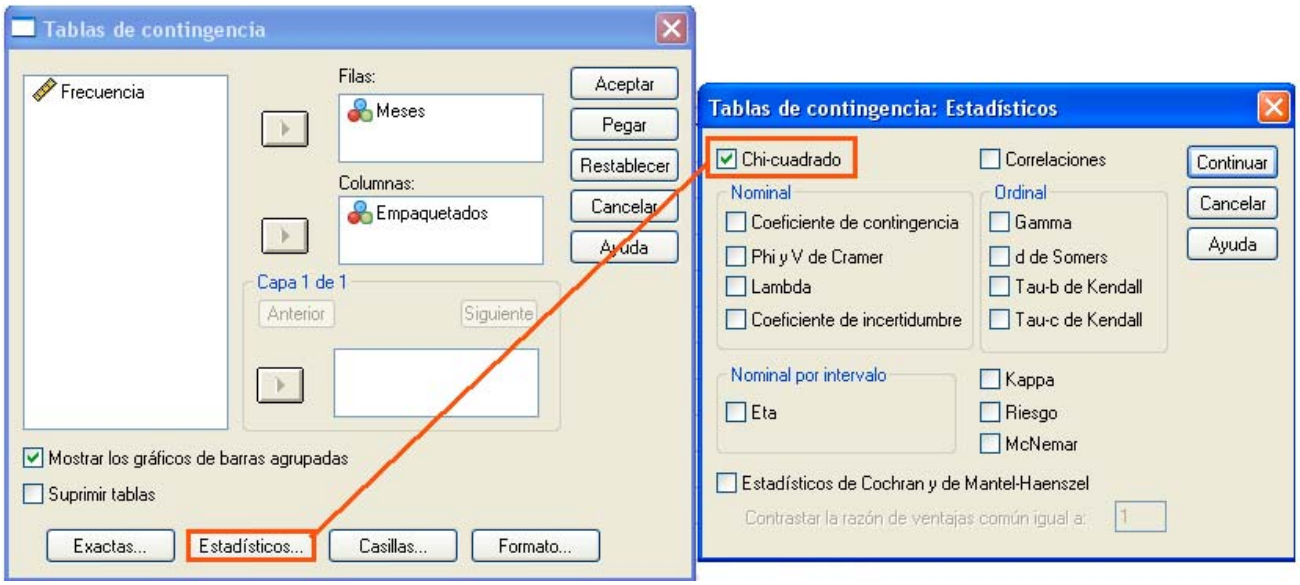
Empaquetado Meses	A	B	C	$O_{i\bullet}$
1	6 $e_{11}=6,35$	10 $e_{12}=9,01$	10 $e_{13}=10,62$	26 26
2	8 $e_{21}=7,81$	12 $e_{22}=11,09$	12 $e_{23}=13,10$	32 32
3	8 $e_{31}=7,32$	8 $e_{32}=10,39$	14 $e_{33}=12,28$	30 30
4	9 $e_{41}=9,52$	14 $e_{42}=13,51$	16 $e_{43}=15,97$	39 39
$O_{\bullet j}$	31	44	52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

$$e_{11}=\frac{26.31}{127}=6,35 \quad e_{12}=\frac{26.44}{127}=9,01 \quad e_{13}=\frac{26.52}{127}=10,65$$
$$e_{21}=\frac{32.31}{127}=7,81 \quad e_{22}=\frac{32.44}{127}=11,09 \quad e_{23}=\frac{32.52}{127}=13,10$$
$$e_{31}=\frac{30.31}{127}=7,32 \quad e_{32}=\frac{30.44}{127}=10,39 \quad e_{33}=\frac{30.52}{127}=12,28$$
$$e_{41}=\frac{39.31}{127}=9,52 \quad e_{42}=\frac{39.44}{127}=13,51 \quad e_{43}=\frac{39.52}{127}=15,97$$

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

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, p_valor*) 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.

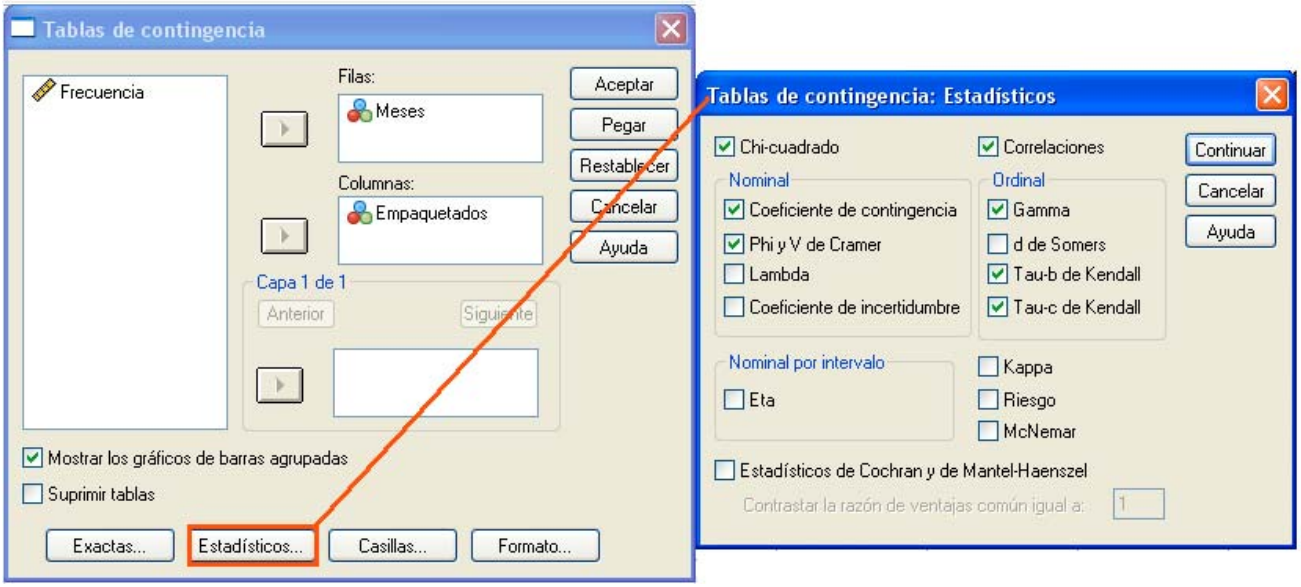
- ***Razón de verosimilitud Chi-cuadrado:*** $RV = 2\sum_{i=1}^{4}\sum_{j=1}^{3} O_{ij} \log\left(\frac{O_{ij}}{e_{ij}}\right) = 1,274 < 12,592 = \chi^2_{0,05\,;\,6}$,

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:*** $\chi^2_c = \sum_{i=1}^{4}\sum_{j=1}^{3} \frac{\left(\left|n - e_{ij}\right| - 0,5\right)^2}{e_{ij}} = 0,059$

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, Lambda y el Coeficiente de incertidumbre.

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 ordina	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	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):**

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

$$\underbrace{0}_{\text{independencia}} \leq C \leq \underbrace{1}_{\text{asociación perfecta}}$$

- **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	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

✓ ***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:* $E_{XY} = \sum_{i,j} \frac{O_{ij}(O_{ij}-1)}{2}$

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)

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

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:

$$T = \frac{N(N-1)}{2} = \frac{127 \text{ x } 126}{2} = 8001$$

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

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

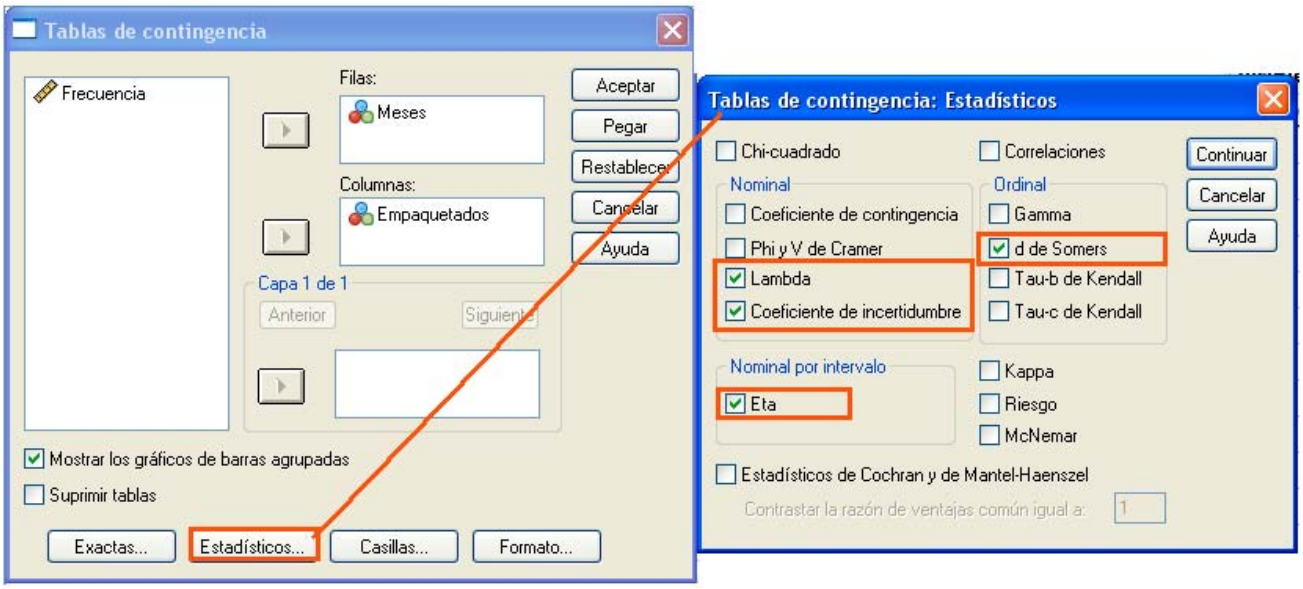
- **Tau-a de Kendall:** $\tau_a = \dfrac{(C-D)}{T} = \dfrac{(2020-1906)}{8001} = 0,0142$

- **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:**

$$\tau_c = \frac{2\,m\,(C-D)}{N^2\,(m-1)} = \frac{2.3(2020-1906)}{127^2.\ 2} = 0,021 \quad \text{donde} \quad m = \text{mín}\left\{ n^o \text{ filas}, n^o \text{ columnas} \right\}$$



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	$O_{1\bullet}=26$	$\text{máx } O_{1j}=10$
2	8	12	12	$O_{2\bullet}=32$	$\text{máx } O_{2j}=12$
3	8	8	14	$O_{3\bullet}=30$	$\text{máx } O_{3j}=14$
4	9	14	16	$O_{4\bullet}=39$	$\text{máx } O_{4j}=16$
Total marginal	$O_{\bullet 1}=31$	$O_{\bullet 2}=44$	$O_{\bullet 3}=52$	$N=127$	$\sum_{j}^{4}\text{máx } O_{ij}=52$
	$\text{máx } O_{i1}=9$	$\text{máx } O_{i2}=14$	$\text{máx } O_{i3}=16$	$\sum_{i=1}^{3}\text{máx}_{i} O_{ij}=39$	

♦ **Coeficiente Lambda:**

$$\lambda=\frac{\sum_{i}\text{máx}_{j} O_{ij}+\sum_{j}\text{máx}_{i} O_{ij}-\text{máx}_{i} O_{i\bullet}-\text{máx}_{j} O_{\bullet j}}{2N-\text{máx}_{i} O_{i\bullet}-\text{máx}_{j} O_{\bullet j}}=\frac{52+39-39-52}{2\,.\,127-39-52}=0$$

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

♦ **Tau de Goodman y Kruskal (*variable X dependiente*):**

$$\tau = \frac{\frac{1}{N}\sum_i (N - O_{i\bullet})O_{i\bullet} - \sum_j \frac{(O_{\bullet j} - O_{ij})O_{ij}}{O_{\bullet j}}}{\frac{1}{N}\sum_i (N - O_{i\bullet})O_{i\bullet}} = \frac{94{,}551 - 94{,}258}{94{,}551} = 0{,}003$$

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

$$\sum_{j=1}^{3}\frac{(O_{\bullet j} - O_{ij})O_{ij}}{O_{\bullet j}} = \left[\frac{(31-6)6 + (31-8)8 + (31-8)8 + (31-9)9}{31}\right] +$$
$$+ \left[\frac{(44-10)10 + (44-12)12 + (44-8)8 + (44-14)14}{44}\right] +$$
$$+ \left[\frac{(52-10)10 + (52-12)12 + (52-14)14 + (52-16)16}{52}\right] = 94{,}258$$

♦ **Coeficiente de Goodman y Kruskal (variable Y dependiente):**

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

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

$$\sum_{i=1}^{4}\frac{(O_{i\bullet} - O_{ij})O_{ij}}{O_{i\bullet}} = \left[\frac{(26-6)6 + (26-10)10 + (26-10)10}{26}\right] + \left[\frac{(32-8)8 + (32-12)12 + (32-12)12}{32}\right] +$$
$$+ \left[\frac{(30-8)8 + (30-8)8 + (30-14)14}{30}\right] + \left[\frac{(39-9)9 + (39-14)14 + (39-16)16}{39}\right] = 82{,}456$$

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

donde, $I(X) = -\sum_{i=1}^{k}\left(\frac{O_{i\bullet}}{N}\right)Ln\left(\frac{o_{i\bullet}}{N}\right)$ $\quad I(Y) = -\sum_{j=1}^{m}\left(\frac{O_{\bullet j}}{N}\right)Ln\left(\frac{O_{\bullet j}}{N}\right)$ $\quad I(XY) = -\sum_{i=1}^{k}\sum_{j=1}^{m}\left(\frac{O_{ij}}{N}\right)Ln\left(\frac{O_{ij}}{N}\right)$

$O_{i\bullet}$	$O_{i\bullet} / N$	$Ln(O_{i\bullet} / N)$	$(O_{i\bullet} / N)\, Ln(O_{i\bullet} / N)$
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

$$I(X) = -\sum_{i=1}^{4}\left(\frac{O_{i\bullet}}{N}\right)Ln\left(\frac{o_{i\bullet}}{N}\right) = 1,3755$$

$O_{\bullet j}$	$O_{\bullet j} / N$	$Ln(O_{\bullet j} / N)$	$(O_{\bullet j} / N)\, Ln(O_{\bullet j} / N)$
31	0,2441	−1,4102	−0,3442
44	0,3465	−1,0600	−0,3673
52	0,4094	−0,8929	−0,3655
N = 127			−1,0771

$$I(Y) = -\sum_{j=1}^{3}\left(\frac{O_{\bullet j}}{N}\right)Ln\left(\frac{O_{\bullet j}}{N}\right) = 1,0771$$

(O_{ij} / N)			$Ln\,(O_{ij} / N)$			$(O_{ij} / N)\ Ln\,(O_{ij} / N)$		
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
						$\sum_{i=1}^{4}\sum_{j=1}^{3}(O_{ij} / N)\,Ln(O_{ij} / N) = -2,4475$		

$$I(XY) = -\sum_{i=1}^{4}\sum_{j=1}^{3}\left(\frac{O_{ij}}{N}\right)Ln\left(\frac{O_{ij}}{N}\right) = 2,4475$$

$$I_{Y/X} = \frac{I(X)+I(Y)-I(XY)}{I(Y)} = \frac{1,3755+1,0771-2,4475}{1,0771} = 0,004734$$

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

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

♦ <u>d</u> de Somers (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: $d_X = \frac{(C-D)}{C+D+E_X} = \frac{(2020-1906)}{2020+1906+2078} = 0,019$

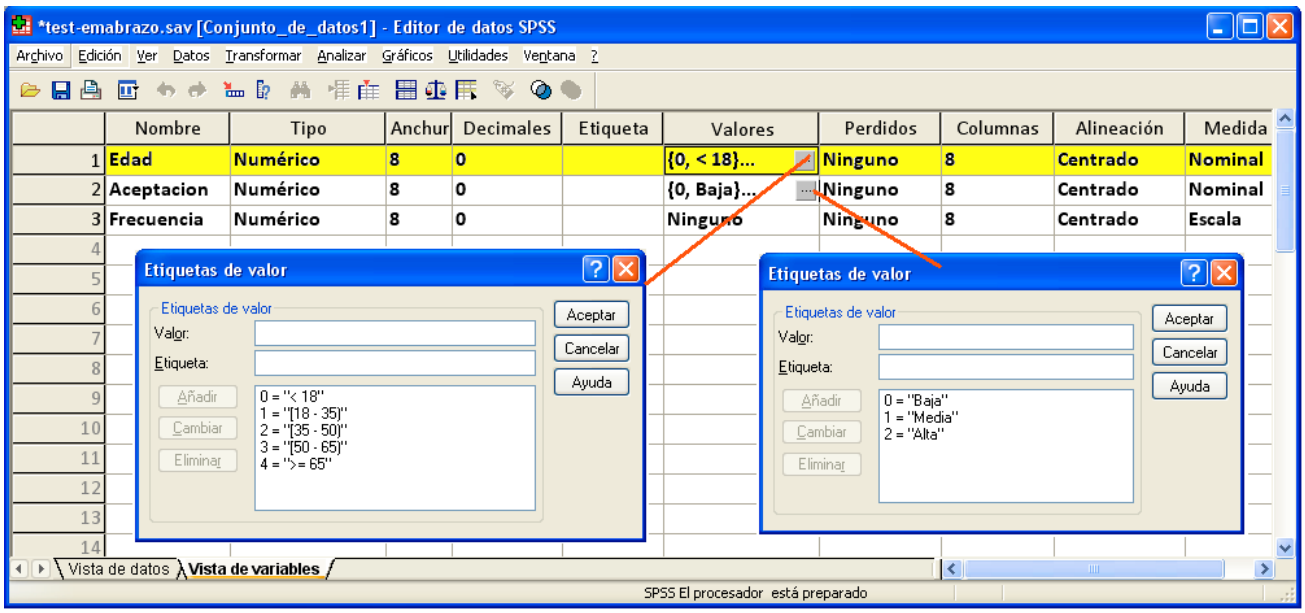
Variable X como independiente: $d_Y = \frac{(C-D)}{C+D+E_Y} = \frac{(2020-1906)}{2020+1906+1338} = 0,022$

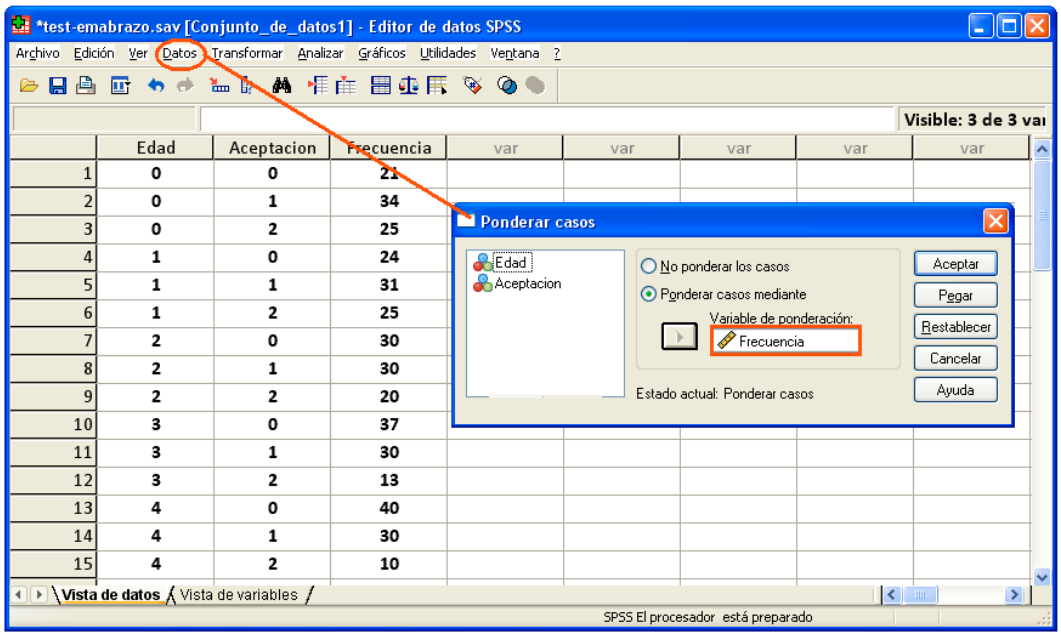
Se quiere estudiar la relación entre la edad de las mujeres y su aceptación de una ley sobre interrupción del embarazo.

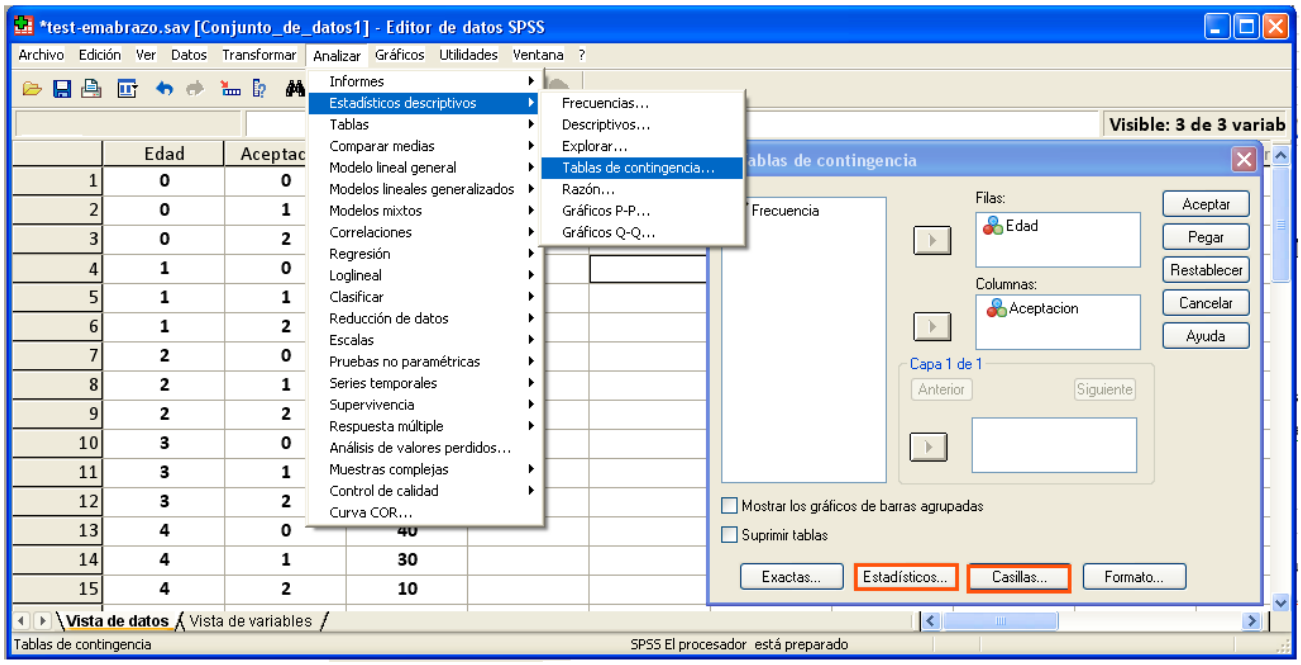
Para ello se ha llevado a cabo una encuesta sobre 400 mujeres cuyos resultados se adjuntan en la tabla. Con un nivel de significación de 0,05, ¿tienen los tres métodos la misma eficacia?.

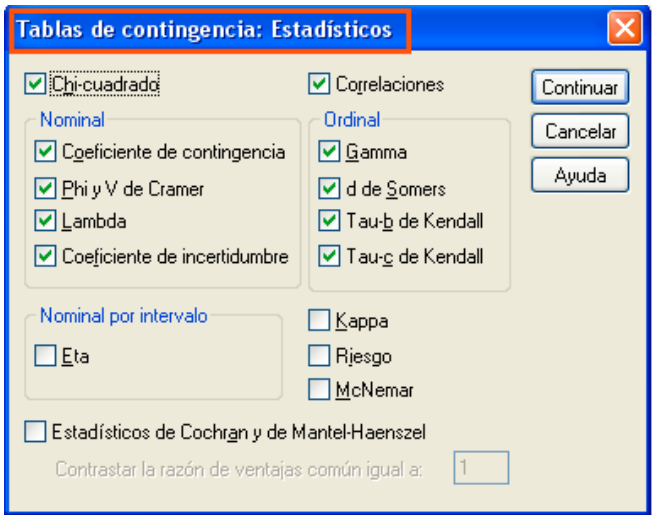
	Aceptación		
Edad	**Baja**	**Media**	**Alta**
0 - 18	**21**	**34**	**25**
18 - 35	**24**	**31**	**25**
35 - 50	**30**	**30**	**20**
50 - 65	**37**	**30**	**13**
> 65	**40**	**30**	**10**

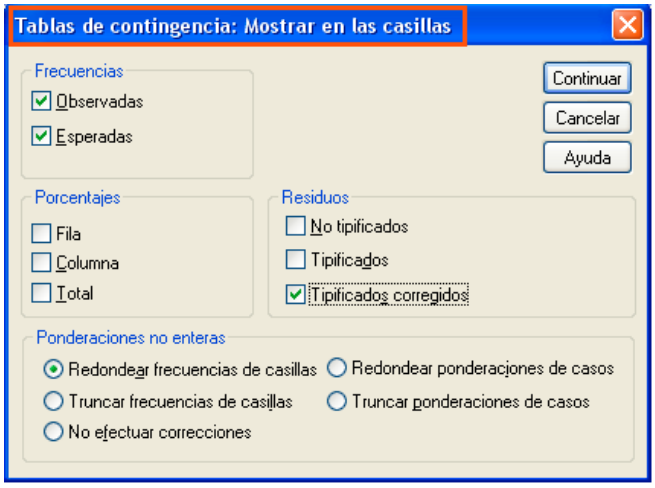
En *Vista de variables* y *Vista de datos* de SPSS:











Los residuos son especialmente útiles para interpretar las pautas de asociación presentes.

Los residuos *No tipificados* son las diferencias existentes entre las diferencias observadas y esperadas de cada casilla: $O_{ij} - e_{ij}$

Los *residuos Tipificados* es el cociente entre el residuo No tipificado y la raíz cuadrada de su correspondiente frecuencia esperada, tienen un valor esperado de 0 y una desviación típica menor que 1 por lo que no pueden interpretarse como puntuaciones normales. Sirven como indicadores del grado en que cada casilla contribuye al valor del estadístico Chi-cuadrado.
Sumando los cuadrados de los residuos tipificados se obtiene el valor del estadístico Chi-cuadrado:

$$r_{ij} = \frac{O_{ij} - e_{ij}}{\sqrt{e_{ij}}} \qquad \chi^2_{(k-1)(m-1)} = \sum_{i=1}^{k}\sum_{j=1}^{m} r_{ij}^2$$

Los *residuos Tipificados corregidos* de Haberman se distribuyen según una distribución teórica N(0,1) , se obtienen dividiendo el residuo de cada casilla por su *error típico*:

$$d_{ij} = \frac{r_{ij}}{\sqrt{V(r_{ij})}} = \frac{(O_{ij} - e_{ij}) / \sqrt{e_{ij}}}{\sqrt{\left(1 - \frac{N_{i\bullet}}{N}\right)\left(1 - \frac{N_{\bullet j}}{N}\right)}} \approx N(0,1)$$

En el *Visor de resultados* de SPSS:

Tabla de contingencia Edad * Aceptacion

			Aceptacion			Total
			Baja	Media	Alta	
Edad	< 18	Recuento	21	34	25	80
		Frecuencia esperada	30,4	31,0	18,6	80
		Residuo	-9,4	3,0	6,4	
		Residuos tipificados	-1,705	,539	1,484	
		Residuos corregidos	-2,421	,770	1,894	
	[18 - 35)	Recuento	24	31	25	80
		Frecuencia esperada	30,400	31,000	18,600	80
		Residuo	-6,400	,000	6,400	
		Residuos tipificados	-1,161	,000	1,484	
		Residuos corregidos	-1,648	,000	1,894	
	[35 - 50)	Recuento	30	30	20	80
		Frecuencia esperada	30,400	31,000	18,600	80
		Residuo	-,400	-1,000	1,400	
		Residuos tipificados	-,073	-,180	,325	
		Residuos corregidos	-,103	-,257	,414	
	[50 - 65)	Recuento	37	30	13	80
		Frecuencia esperada	30,400	31,000	18,600	80
		Residuo	6,600	-1,000	-5,600	
		Residuos tipificados	1,197	-,180	-1,298	
		Residuos corregidos	1,700	-,257	-1,657	
	>= 65	Recuento	40	30	10	80
		Frecuencia esperada	30,400	31,000	18,600	80
		Residuo	9,600	-1,000	-8,600	
		Residuos tipificados	1,741	-,180	-1,994	
		Residuos corregidos	2,472	-,257	-2,545	
Total		Recuento	152	155	93	400
		Frecuencia esperada	152,0	155,0	93,0	400

Los distintos datos de las casillas pueden a ayudar a intuir pautas de asociación, aunque son los *residuos tipificados corregidos* los que permiten interpretar de forma precisa la relación existente entre las variables. Con un nivel de significación del 5%, basta fijarse en aquellos que son mayores de 1,96 o menores que – 1,96, observando 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 mujeres mayores de 65 años de clase alta tienen una opinión favorable sobre la interrupción del embarazo.
Por el contrario, la opinión de las mujeres mayores de 65 años de clase baja mantienen una percepción claramente negativa. De este modo, se evidencia que existe una relación y del tipo que es ésta.

Señalar que este método supone un análisis celda a celda, mientras que 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)

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.

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 $0,013$.

Puesto que esta probabilidad (denominada *nivel crítico* o *nivel de significación asintótica*) o *p_valor* es pequeña (menor que 0,05) se rechaza 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, indica que hay evidencia de asociación entre el grado de aceptación del aborto y la edad de las mujeres.

$$RV = 2\sum_{i=1}^{5}\sum_{j=1}^{3} O_{ij}\log\left(\frac{O_{ij}}{e_{ij}}\right)$$

$$RV < \chi^2_{\alpha;\,(k-1)\,.\,(m-1)} \Rightarrow \text{X e Y son independientes al nivel } \alpha$$

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

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 $< 0,05$, por lo que se rechaza la hipótesis nula de independencia, llegando a la misma conclusión que con los estadísticos anteriores.

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
		Aceptacion dependiente	,069	,055	1,213	,225
	Tau de Goodman y Kruskal	Edad dependiente	,012	,005		,014[c]
		Aceptacion dependiente	,022	,010		,022[c]
	Coeficiente de incertidumbre	Simétrica	,019	,008	2,300	,011[d]
		Edad dependiente	,015	,007	2,300	,011[d]
		Aceptacion dependiente	,023	,010	2,300	,011[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.

Los valores obtenidos de Lambda, Tau de Goodman y Kruskal, 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 Kruskal 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.*

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	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.

ANÁLISIS DE REGRESIÓN LOGÍSTICA

La regresión logística (RL) se incluye dentro de las denominadas técnicas estadísticas del análisis de datos, es un tipo de análisis de regresión que se utiliza para predecir el resultado de una variable categórica (número limitado de categorías) en función de las variables independientes o predictoras. Su estudio se enmarca en el conjunto de Modelos Lineales Generalizados (GLM) utilizando como función de enlace la función LOGIT.

Su uso se hace imprescindible cuando se quiere relacionar una variable dependiente cualitativa con una o más variables independientes. En el análisis de datos sociales, antes que su capacidad para establecer relaciones funcionales y predecir sucesos, su utilidad deriva de la lectura de los coeficientes *Odd Ratio* para interpretar los efectos que tienen las categorías sobre la variable dependiente.

Uno de los problemas fundamentales cuando intervienen diversas variables en un fenómeno es determinar cuál es la contribución de cada una de ellas, suponiendo que el resto de variables no cambian.

La regresión logística puede considerarse un caso especial de la regresión lineal donde la variable dependiente es dicotómica, con la particularidad de que el dominio de salida de la función está acotado en el intervalo [0, 1] y que el proceso de estimación, en lugar de mínimos cuadrados, utiliza el procedimiento de estimación *máxima verosimilitud*.

Es por tanto, una técnica multivariante de dependencia que trata de estimar la probabilidad p de que ocurra un suceso en función de la dependencia de otras variables independientes $X_1, X_2, \cdots, X_k$. Se asigna un sujeto a una categoría cuando tiene una probabilidad $p > 0,5$ de estar encuadrado en esa categoría, en otro caso se le asignara la otra categoría.

HERAMIENTAS: Odd Ratio y Logit

La tabla de contingencia, en los entrevistados, indica la participación en una huelga según el sexo:

	Hombre	Mujer	Total
No Participan	640	920	1560
Participan	310	2080	520
Total	950	1130	

La tabla muestra una gran mayoría que han no han participado en la huelga, es decir, han participado en la huelga 520 sujetos entre los 2080 entrevistados. En porcentajes, $\frac{520}{2080} = 0,25$ (25%) participan frente a un 75% que no participan.

Se pueden expresar los datos mediante una razón, pudiendo indicar que hay 0,25 sujetos que participan en la huelga por cada uno que no participa, o lo que es equivalente: 25 participan en la huelga por cada 100 que no participan.

También se puede leer a la inversa, $\frac{1560}{520}=3$, hay 3 personas que no participan por cada una que participa, o bien, hay 300 personas que no participan por cada 100 que participan.

La *razón* o *ratio* es el cociente entre dos cantidades e indica cuantas veces una cantidad es mayor o menor respecto a la otra.

Los datos indican que los hombres participan más en la huelga que las mujeres.

✓ **Para saber cuánto más se puede calcular la diferencia de porcentajes:**

Participación en la huelga por sexo (porcentajes verticales)

	Hombre	Mujer	Total
No Participan	640 / 950 = 67,368422%	920 / 1130 = 81,415929%	1560 / 2080 = 75%
Participan	310 / 950 = 32,631578%	210 / 1130 = 18,584071%	520 / 2080 = 25%
Total	100%	100%	100%

$32,631578\% - 18,584071\% = 14,047507\%$ → Los hombres participan un 14,047507% más que las mujeres.

✓ **Describiendo la tabla en términos de razón: El 32,631578% de los hombres declaran haber hecho huelga mientras que el 67,368422% restante declara no haber participado, con lo que la relación entre hombres que hacen huelga y no la hacen es $\frac{32,631578}{67,368422}=0,484375$.**

La ratio se interpreta: Hay 0,484375 hombres que hacen huelga por cada uno que no la hace.

Esta razón o ratio se denomina con el término *Odd* refiriéndose a la razón que se establece entre la ocurrencia (o su probabilidad) de un suceso respecto a su no-ocurrencia, interpretándose como ventaja comparativa: $\text{Odd}=\frac{p}{q}=\frac{p}{1-p}$

De forma análoga, la razón entre las mujeres que declaran haber hecho huelga frente a las que no han participado en la huelga es $\frac{18,584071}{81,415929}=0,228261$, concluyendo que hay 0,228261 mujeres que hacen huelga por cada una que no la hace.

	Hombre	Mujer
$\text{Odd}_{\text{Participa/No participa}}$ =	0,484375	0,228261

Se observa que los hombres participan más en la huelga que las mujeres.

✓ **Para responder cuánto más participan los hombres en la huelga que las mujeres se puede utilizar otro ratio:**

$$\text{Mujeres/Hombres} = \frac{0,228261}{0,484375} = 0,471249$$

Este nuevo ratio, ratio de ratios, se interpreta: La probabilidad de encontrar una mujer que hace huelga sobre una mujer que no la hace es de 0,471249 respecto al caso de los hombres.

Este término, que es una razón de Odds, se denomina *Odd Ratio (OR)* y se interpreta como ventaja comparativa o como razón de probabilidades. Cuando el Odd Ratio alcanza el valor 1 indica que no hay diferencias.

$$\text{Odd Ratio} = \frac{\text{Odd}_H}{\text{Odd}_M} = \frac{\left(\frac{p_H}{q_H}\right)}{\left(\frac{p_M}{q_M}\right)} = \frac{\left(\frac{p_H}{1-p_H}\right)}{\left(\frac{p_M}{1-p_M}\right)}$$

La relación que existe entre Odd y la proporción *p* entendida como la probabilidad de que ocurra el suceso en estudio:

$$\text{Odd} = \frac{p}{1-p} \rightarrow (1-p).\text{Odd} = p \rightarrow (1+\text{Odd}).p = \text{Odd} \rightarrow p = \frac{\text{Odd}}{1+\text{Odd}}$$

MODELO DE REGRESIÓN LOGÍSTICA

El modelo de regresión logística parte de la hipótesis de que los datos siguen el modelo:

$$\text{Ln}(\text{Odd}) = \text{Ln}\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 . x_1 + \beta_2 . x_2 + \cdots + \beta_k . x_k + \in_I$$

donde $\in_i \equiv$ **término error o residuo**, con la hipótesis $E(\in_i) = 0$

$\beta_0 \equiv$ **intersección o término constante**, $(\beta_1, \beta_2, \cdots, \beta_k)$ denotan la magnitud del efecto que pueden adoptar las variables aleatorias independientes.

Para simplificar la notación se define $z = \beta_0 + \beta_1 . x_1 + \beta_2 . x_2 + \cdots + \beta_k . x_k$

Operando algebraicamente sobre el modelo:

$$\text{Ln}\left(\frac{p}{1-p}\right) = z \rightarrow \frac{p}{1-p} = e^z \rightarrow p = (1-p).e^z \rightarrow p + p.e^z = e^z$$

$$p = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

Siendo la función de distribución logística $F(x) = \frac{e^x}{1+e^x}$ se puede expresar el modelo de una forma más compacta:

$$p = \frac{e^z}{1+e^z} = F(z)$$

En consecuencia, las dos formas más importantes del modelo de regresión logística:

$$\text{Logit} = \text{Ln}(\text{Odd}) = \text{Ln}\left(\frac{p}{1-p}\right) \qquad \text{Odds Ratio} = \frac{p}{1-p} = e^{\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k}$$

Participación en la huelga por sexo (porcentajes verticales)

	Hombre	Mujer	Total
No Participan	0,67368422	0,81415929	0,75
Participan	0,32631578	0,18584071	0,25
Total	1	1	1

Se puede emplear una variable de categoría como independiente, en este caso será el sexo. Dicha variable se codifica de la siguiente forma: 0 cuando es hombre y 1 cuando es mujer. Esta codificación de *ceros* y *unos* se denomina *dummy*.

🕮 A partir de la participación por sexo, expresada en proporciones, se pueden calcular los Logit.

✍ Logit para hombre ($X = 0$):

$$\text{Logit}_H = \text{Ln}(\text{Odd}_H) = \text{Ln}\left(\frac{p_H}{1-p_H}\right) = \text{Ln}\left(\frac{0,32631578}{0,67368422}\right) = -0,724896$$

$$\text{Odds}\,(X=0) = \frac{p_H}{1-p_H} = e^{-0,724896} = 0,484375$$

✍ Logit para mujer ($X = 1$):

$$\text{Logit}_M = \text{Ln}(\text{Odd}_M) = \text{Ln}\left(\frac{p_M}{1-p_M}\right) = \text{Ln}\left(\frac{0,18584071}{0,81415929}\right) = -1,477266$$

$$\text{Odds}\,(X=1) = \frac{p_M}{1-p_M} = e^{-1,477266} = 0,228261$$

En el caso de una variable independiente dicotómica el ajuste de la función a partir de los Logit resulta sencillo: $\text{Logit} = \text{Ln}\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 . x_1$

$\text{Logit}(X=0) = -0,724896 = \beta_0 + \beta_1 . 0 \quad \rightarrow \quad \beta_0 = -0,724896$

$\text{Logit}(X=1) = -1,477266 = \beta_0 + \beta_1 . x \quad \rightarrow \quad -1,477266 = -0,724896 + \beta_1 . 1$

$\beta_1 = -0,75237$

Se observa que una vez conocidos $\beta_0 = -0,724896$ y $\beta_1 = -0,75237$ se pueden calcular las probabilidades $p = \frac{e^{\beta_o + \beta_1 . x}}{1 + e^{\beta_o + \beta_1 . x}} = \frac{1}{1 + e^{-(\beta_o + \beta_1 . x)}}$

Cuando $X = 0$ (hombres): $p = \frac{e^{-0,724896}}{1 + e^{-0,724896}} = 0,32631578$ (32,631578% de los hombres que hacen huelga)

Cuando $X = 1$ (mujeres): $p = \frac{e^{-0,724896-0,75237}}{1 + e^{-0,724896-0,75237}} = 0,18584071$ (18,584071% de las mujeres que hacen huelga)

La interpretación de los coeficientes es la siguiente:

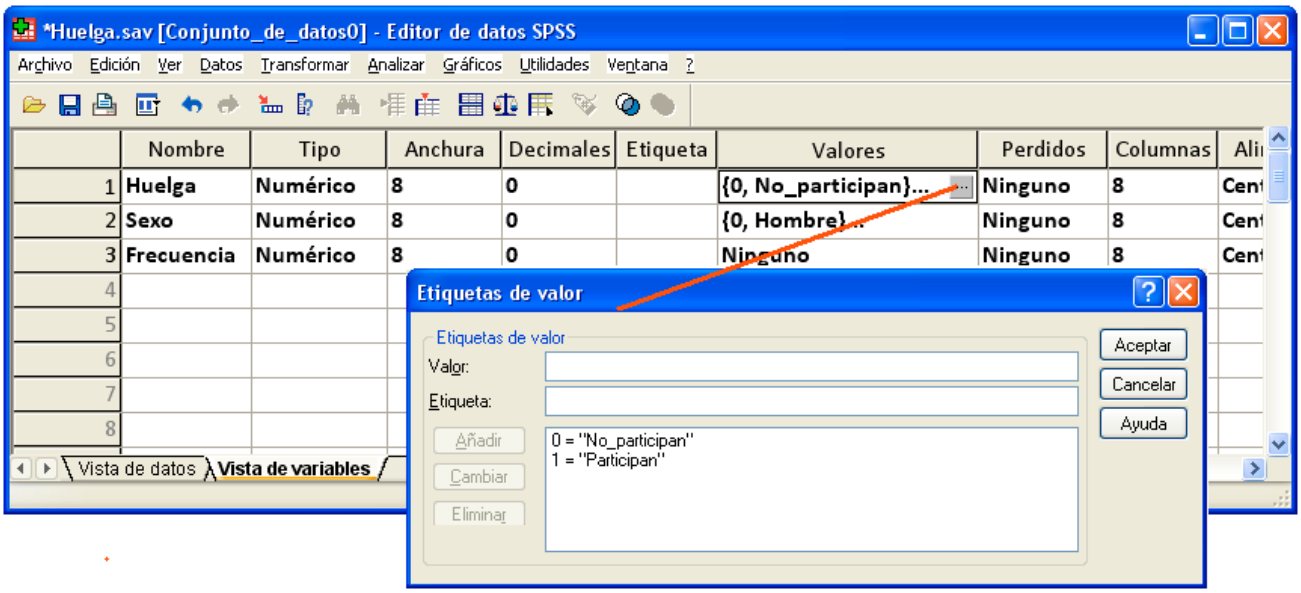
Constante $\beta_0 = -0,724896 \Rightarrow e^{-0,724896} = 0,484375 = \text{Odd}_H$

Pendiente $\beta_1 = -0,75237 \Rightarrow e^{-0,75237} = 0,471248 = \text{OR}$

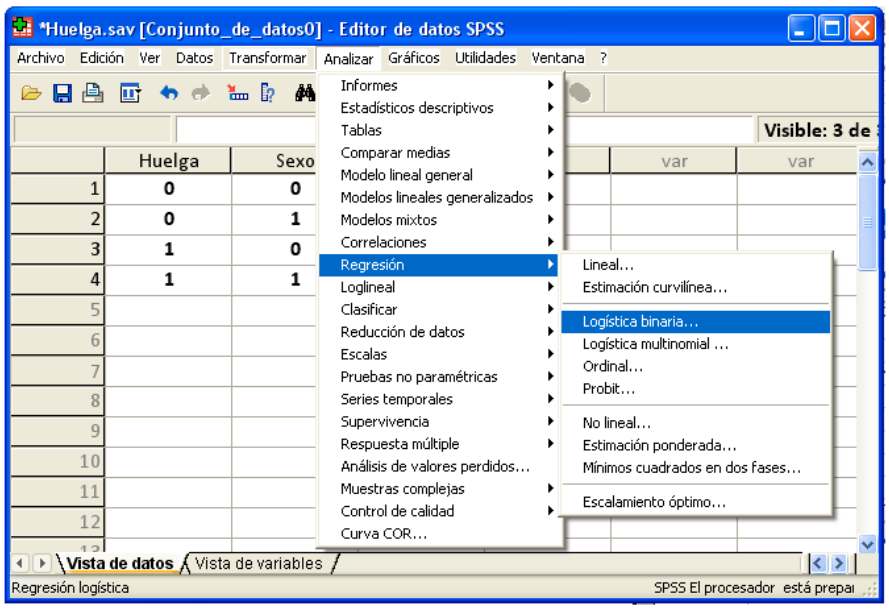
Los coeficientes β son Odds Ratio, es decir, relacionan una categoría respecto a otra. En este caso, donde se analiza mujeres por cada hombre, los hombres son la categoría que sirve de comparación.

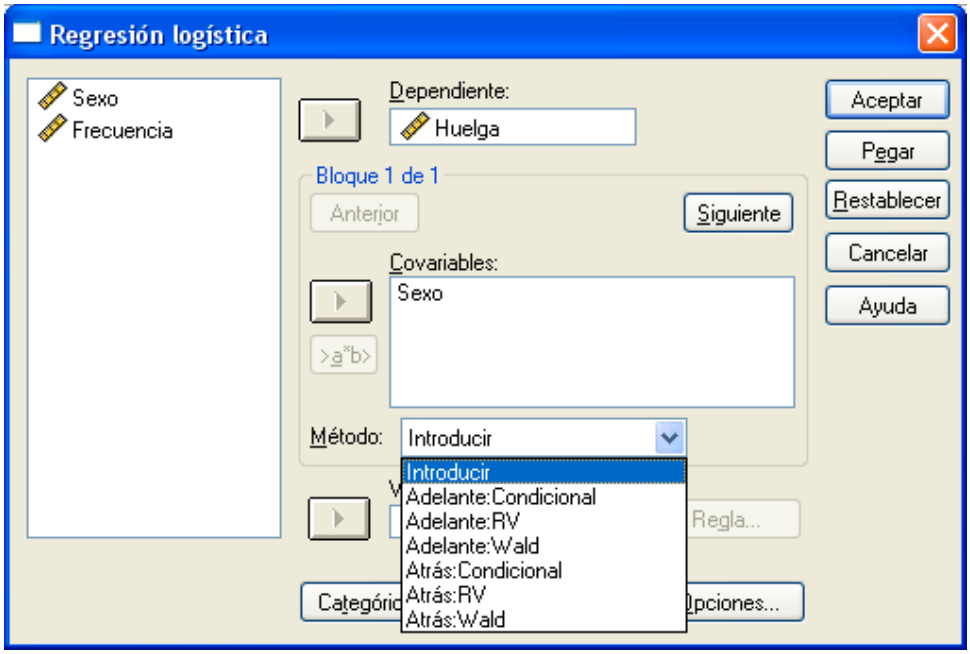
Cuando $\text{OR} < 1$ hay menos mujeres que hombres que hacen huelga, más concretamente la relación es de 0,471249, expresando que hay 47 mujeres que hacen huelga por cada 100 hombres que también hacen huelga.

PROCESO EN SPSS



La regresión logística binaria se utiliza cuando la variable dependiente es una variable binaria, es decir, de solo dos categorías, también conocidas como *dummy* o *dicotómica*.





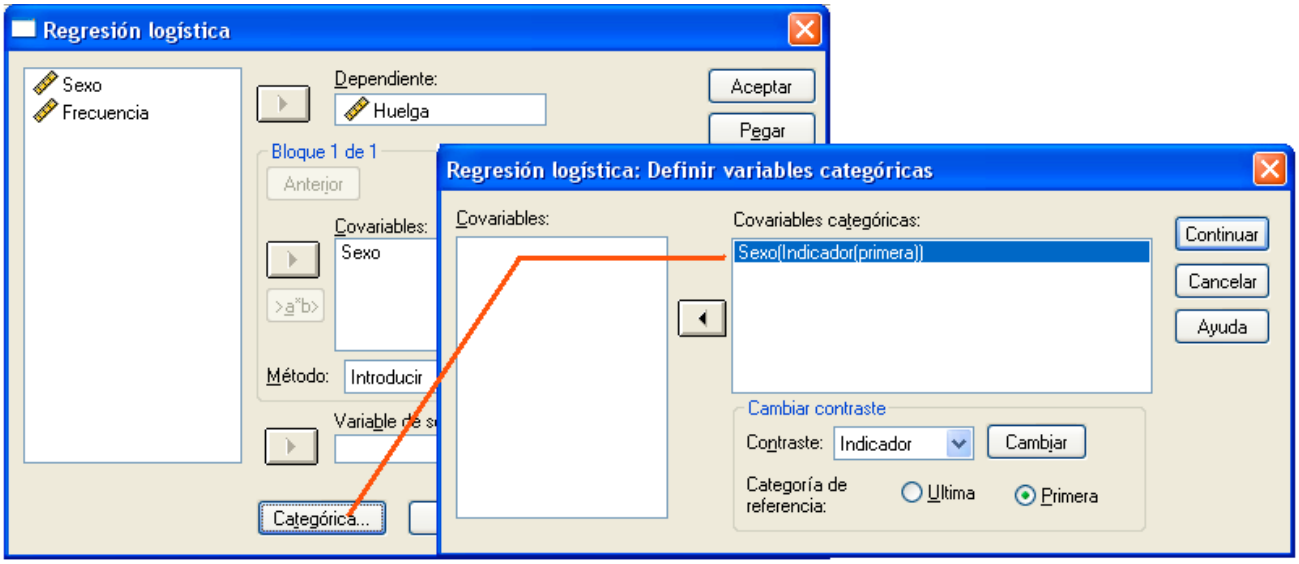
✓ **La *variable dependiente* (o resultado) que se desea modelizar o predecir, será categórica dicotómica, codificada con valores 0 y 1 (si no estuviera así codificada el programa le asigna ese código interno).**

✓ ***La covariable o covariables* ya sean *predictoras*, *confundentes y/o modificadoras* del efecto, y que parece deben incluirse en el modelo**.

✓ **En *Caja Método* se despliega una ventana con el fin de ajustar el modelo con todas las variables que se introdujeron en el cajón de Covariables. El *Método Introducir* es un procedimiento en el que todas las variables de un bloque se introducen en un solo paso. El *Método Adelante RV* (método automático por pasos, hacia adelante, que utiliza la prueba de la Razón de Verosimilitud para comprobar las covariables a incluir o excluir).**

Una variable categórica es una variable que puede tomar uno de un número limitado, y por lo general fijo, de posibles valores, asignando a cada unidad individual u otro tipo observación a un grupo en particular o categoría nominal sobre la base de alguna característica cualitativa.
Una variable categórica que puede tomar dos valores se denomina una *variable binaria* o una *variable dicotómica*. Un caso especial importante es la variable de Bernoulli.
Las variables categóricas con más de dos valores posibles se denominan variables *politómicas*. Las variables categóricas a menudo se supone que son politómicas menos que se especifique lo contrario.

✓ **El botón *Categórica* (que se ilumina sólo cuando hay alguna variable en la caja de las Covariables) permite indicar qué covariables son categóricas (SPSS trata todas las variables como numéricas (discretas o continuas) a menos que se indique que son categóricas). Al pulsar aparece una ventana que se rellena.**



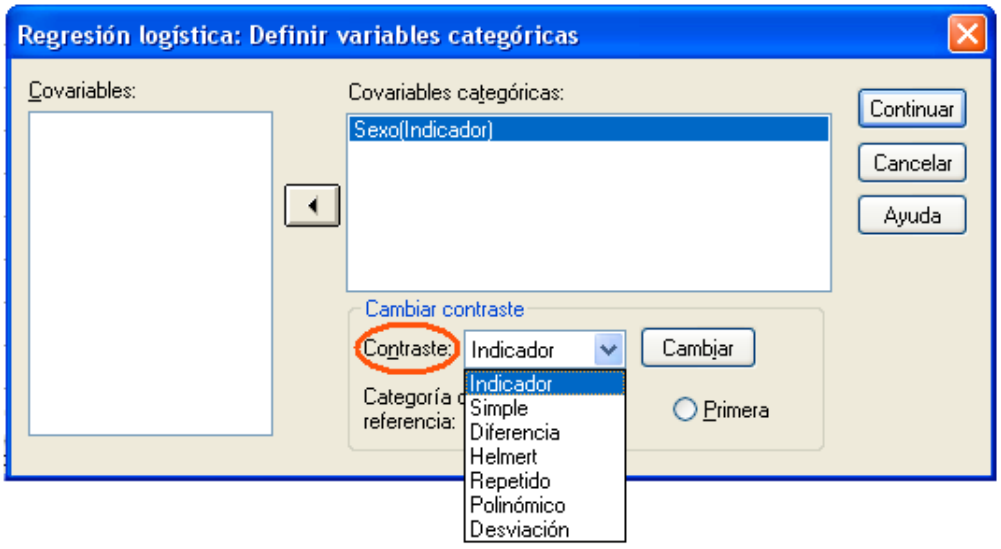
Hay que considerar que al pasar de la caja de Covariables a la Caja de Covariables categóricas:

→ Cada covariable categórica es sustituida por una o más covariables.

→ Cuando la covariable es binaria (como en este caso) se creará una variable que tendrá un 0 en la categoría de referencia y un 1 en la categoría de riesgo.

→ Cuando la covariable tiene k categorías, se crearán $(k-1)$ covariables codificadas.

Queda por definir qué categoría es la de referencia (la de no riesgo) en cada de una de las variables categóricas incluidas en el problema (en este caso, solo Sexo).

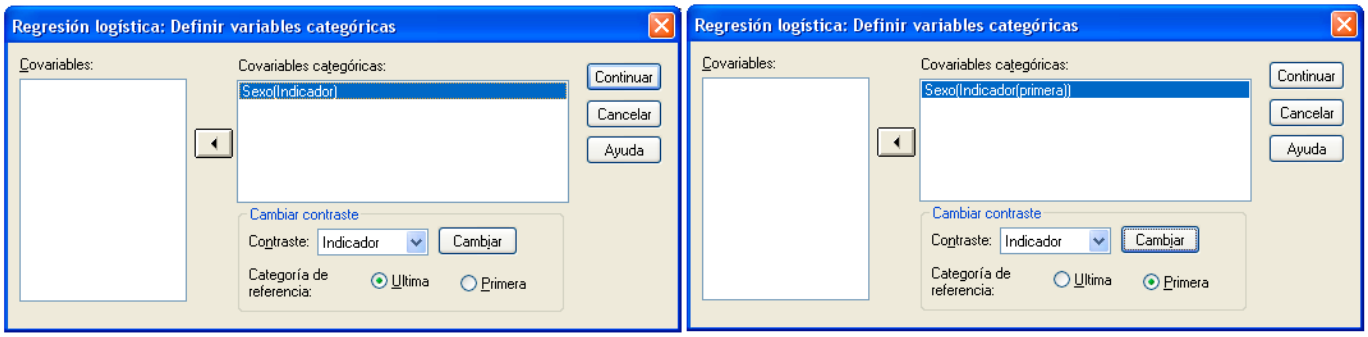


En un caso general, se selecciona una a una las covariables, utilizando las opciones en Cambiar de contraste en cada una de ellas.

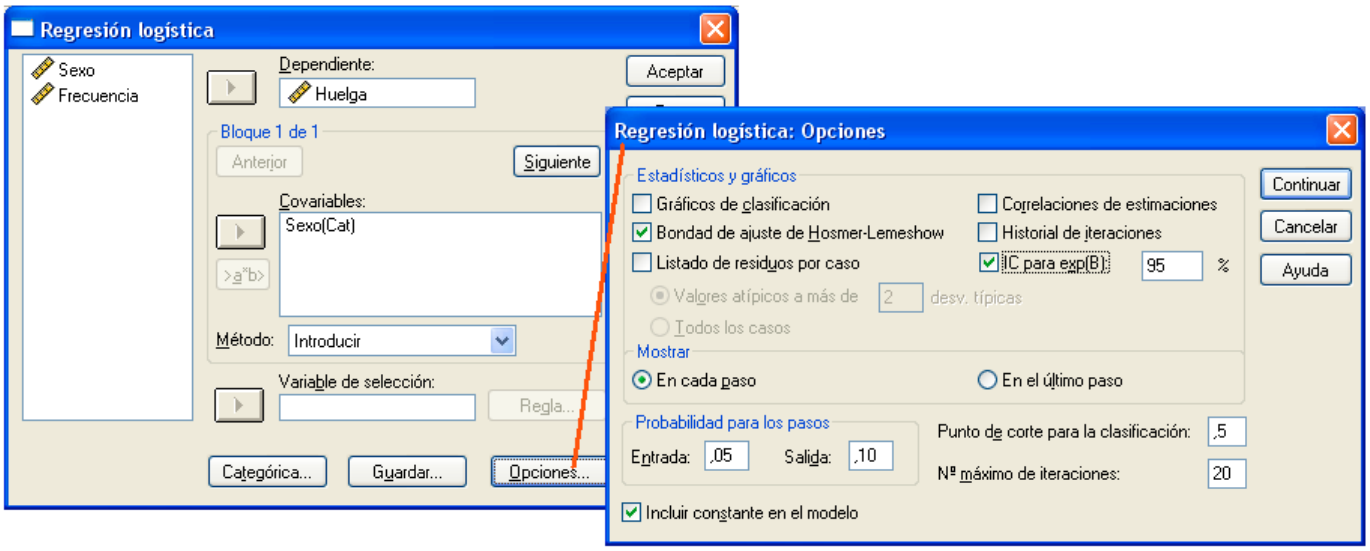
→ En Contraste se elige siempre Indicador.

→ **En Categoría de referencia hay que indicar cuál es la categoría de no riesgo para la variable seleccionada - SPSS incorpora por defecto *la Última* (se deja por defecto cuando sea apropiado. En este caso es *la Primera* (pues el 0 se refiere siempre al *no riesgo*).**

→ **Para que el cambio sea efectivo hay que pulsar el botón de Cambiar.**
En la ventana Covariables categóricas el texto *Sexo(indicador)* cambia a *Sexo(indicador (primera))* cuando se seleccionó Primera. Si se hubiera dejado *Última* aparecería *Sexo(indicador)*



✓ **El botón *Opciones* permite obtener estadísticos y gráficos, o cambiar el criterio de construcción del modelo. Tras pulsar aparece una ventana.**



→ ***Bondad de ajuste de Hosmer-Lemeshow:*** **Verifica si el modelo de regresión logística ajusta bien o no a los datos:**
H_0: El modelo ajusta bien , H_1: El modelo no ajusta bien Si el test de significación $(p\_valor) \leq 0{,}10$ (en este caso) se rechaza la hipótesis nula, nada de lo que se calcule es válido.

→ ***IC para exp(B):*** **Calcula el intervalo de confianza para las razones del producto cruzado de todas las variables presentes en el modelo. Permite fijar la confianza deseada (el 95% en este caso).**

→ ***Incluir constante en la ecuación:*** **Al marcar ajusta un modelo con término independiente β_0 (viene marcado por defecto).**

✓ **Al pulsar *Continuar* aparecen los siguientes resultados en el Visor de SPSS:**

Resumen del procesamiento de los casos

Casos no ponderados[a]	N	Porcentaje
Casos seleccionado: Incluidos en el análisis	4	100,0
Casos perdidos	0	,0
Total	4	100,0
Casos no seleccionados	0	,0
Total	4	100,0

a. Si está activada la ponderación, consulte la tabla de clasificació para ver el número total de casos.

Indica la codificación empleada para la variable dependiente, tanto real (No/Si) como interna 0 (No participa) y 1 (Sí participa)

Codificación de la variable dependiente

Valor original	Valor interno
No_participan	0
Participan	1

Analiza el modelo solo con el término independiente (sin interés).

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-1,099	,051	470,710	1	,000	,333

Variables que no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	Sexo(1)	54,317	1	,000
	Estadísticos globales		54,317	1	,000

Es un análisis *univariante* que permite saber si las diferentes variables (en este caso, Sexo) del modelo consideradas individualmente están o no asociadas con la variable dependiente. Cuando cada Signatura o $p\_valor \leq 0,05$ (en este caso, una sola) se rechaza la hipótesis nula concluyendo que, considerando cada covariable individualmente (sin tener en cuenta las otras), están asociadas con la variable dependiente.
Tiene un interés pequeño.

Presenta diferentes tests globales dependiendo del método de construcción empleado. Solo interesa la fila *Modelo* que es la que alude al método *Introducir* que se ha empleado.

Pruebas omnibus sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	54,284	1	,000
	Bloque	54,284	1	,000
	Modelo	54,284	1	,000

La hipótesis nula del test global: $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ (Independencia global) frente a la hipótesis alternativa H_1: Alguna igualdad no es cierta (Dependencia global).

Como $\chi^2_{exp} = 54,284$ con $g.l. = 1 = n^o$ de covariables tiene un $p\_valor \leq 0,05$ se rechaza la hipótesis nula, con lo que al menos una de las variables en el modelo (en este caso, sexo) está asociada a la variable dependiente.

Es la primera tabla relevante, en caso de no dar significativa el problema finalizaría (la variable o variables no influirían en la variable dependiente).

La prueba de Hosmer y Lemeshow se utiliza para evaluar la bondad del ajuste de un modelo de regresión logística.

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	,000	0	.

Parte de la idea que si el ajuste es bueno, un valor alto de la probabilidad predicha (p) se asociará con el resultado 1 de la

Tabla de contingencias para la prueba de Hosmer y Lemeshow

		Huelga = No_participan		Huelga = Participan		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	920	920,000	210	210,000	1130
	2	640	640,000	310	310,000	950

variable binomial dependiente. Mientras que un valor bajo de (p) próximo a 0 corresponde (en la mayoría de las ocasiones) con el resultado $Y = 0$. Esta prueba de bondad aquí tiene algunos inconvenientes, el estadígrafo de Hosmer y Lemeshow no se computa cuando, para algunos grupos, $e_i \equiv$ valores esperados ó $(O_i - e_i)$ son nulos o muy pequeños (valores menores que 5).

Señalar que lo que se desea en esta prueba es que *no haya significación* (lo contrario a los que es habitual), motivo por el que muchos estudios proponen simplemente cotejar valores observados y esperados mediante simple observación y evaluar el grado de concordancia entre unos y otros a partir del sentido común.

En este sentido, una forma de evaluar la ecuación de regresión y el modelo obtenido es construir una tabla de 2x2 clasificando a todos los individuos de la muestra según la concordancia de los valores observados con los predichos o estimados por el modelo, de forma similar a como se evalúan las pruebas diagnósticas.

Una ecuación sin poder de clasificación alguno tendría una especificidad, sensibilidad y total de clasificación correctas igual al 50% (por el simple azar). Un modelo puede considerarse aceptable si tanto la especificidad como la sensibilidad tienen un nivel alto, de al menos el 75%.

Tabla de clasificación[a]

			Pronosticado		
			Huelga		Porcentaje
	Observado		No_participan	Participan	correcto
Paso 1	Huelga	No_participan	1560	0	100,0
		Participan	520	0	,0
	Porcentaje global				75,0

a. El valor de corte es ,500

En la tabla de clasificación se puede observar que el modelo tiene una especificidad alta (100%) y una sensibilidad nula (0%), con lo que la constante y única variable predictora (sexo) clasifica mal a los individuos que participaron en la huelga (Huelga = 1) cuando el punto de corte de la probabilidad de Y se establece por defecto en 50% (0,5).

Resumen de los modelos

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	2285,030[a]	,026	,038

a. La estimación ha finalizado en el número de iteración 4 porque estimaciones de los parámetros han cambiado en menos de ,00

Como el test global dio significativo conviene ver las medidas de resumen.

El estadístico (–2LL), menos dos veces el logaritmo neperiano de la verosimilitud, mide hasta qué punto un modelo se ajusta bien a los datos. El resultado de esta medición se denomina *desviación*. Cuanto más pequeño sea el valor, mejor será el ajuste.

R^2 de Cox-Snell es un valor muy discreto, indica que solo el 2,6% de la variación de la variable dependiente (hacer o no huelga) es explicada por el sexo.

El coeficiente de Nagelkerke, versión corregida del coeficiente de Cox-Snell, indica que solo el 3,8% de la variación de la variable dependiente es explicada por la variable incluida en el modelo.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	Sexo(1)	-,752370	,103137	53,215388	1	,000	,471248	,384999	,576819
	Constante	-,724896	,069198	109,741104	1	,000	,484375		

a. Variable(s) introducida(s) en el paso 1: Sexo.

SPSS ofrece las variables que dejará en la ecuación, coeficientes de regresión con sus correspondientes errores estándar, el valor del estadístico Wald para evaluar la hipótesis nula $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$, la significación o p_valor asociado, y el valor $OR \equiv \exp(B)$ con el intervalo de confianza.

→ Pendiente : Estimador $\hat{\beta}_1 = -0,75237 \Rightarrow e^{-0,75237} = 0,471248 = OR$

→ Constante: Estimador $\hat{\beta}_0 = -0,724896 \Rightarrow e^{-0,724896} = 0,484375 = Odd_H$

→ El estadístico de Wald $W^2 = \left(\frac{\hat{\beta}_i}{\sigma(\hat{\beta}_i)}\right)^2 = \left(\frac{\hat{\beta}_i}{E.T.(\hat{\beta}_i)}\right)^2$ bajo la hipótesis nula sigue una Chi-cuadrado con 1 grado de libertad.

Cuando $p\_valor \leq 0,05$ se rechaza la hipótesis nula $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$, concluyendo que la variable sexo es significativa.

$$W^2_{sexo} = \left(\frac{\hat{\beta}_1}{\sigma(\hat{\beta}_1)}\right)^2 = \left(\frac{-0,752370}{0,103137}\right)^2 = 53,215388$$

$$W^2_{constante} = \left(\frac{\hat{\beta}_0}{\sigma(\hat{\beta}_0)}\right)^2 = \left(\frac{-0,724896}{0,069198}\right)^2 = 109,741104$$

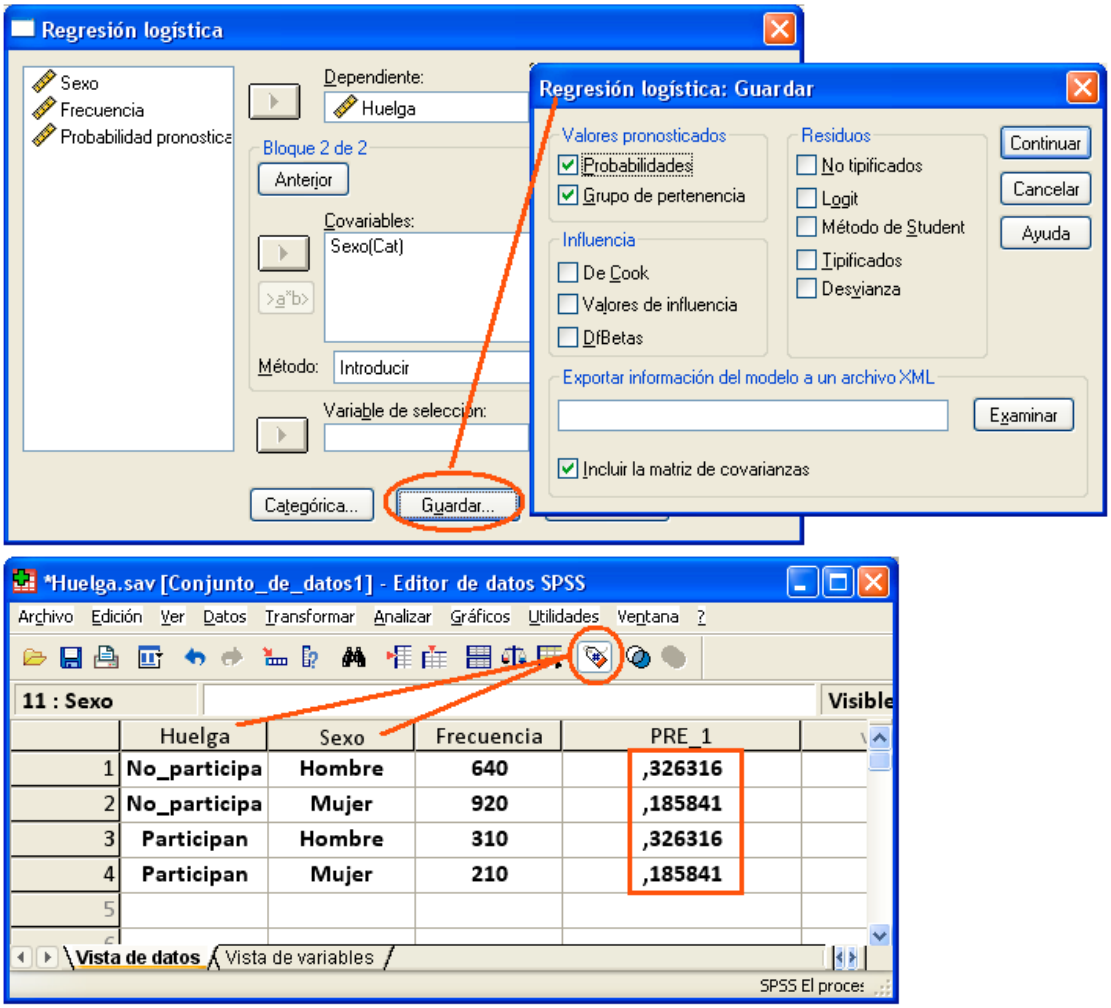
→ El intervalo de confianza para β_1 :

$$I.C(\hat{\beta}_1) = \left[\hat{\beta}_1 \pm z_{\alpha/2} \,.\, \sigma(\hat{\beta}_1)\right] = \left[-0,752370 \pm 1,96 \,.\, 0,103137\right] = \left[-0,954519\ ,\ -0,550221\right]$$

$$I.C(e^{\hat{\beta}_1}) = \left[e^{-0,752370 - 1,96\,.\,0,103137}\ ,\ e^{-0,752370 + 1,96\,.\,0,103137}\right] = \left[e^{-0,954519}\ ,\ e^{-0,550221}\right] = \left[0,3484999\ ,\ 0,576819\right]$$

→ Ecuación de regresión logística: $p = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(\beta_o + \beta_1 . x)}}$

En la ventana de Regresión logística, con el botón Guardar, se selecciona Probabilidades y Grupo de Pertenencia, se pueden consultar los valores pronosticados en Vista de datos.



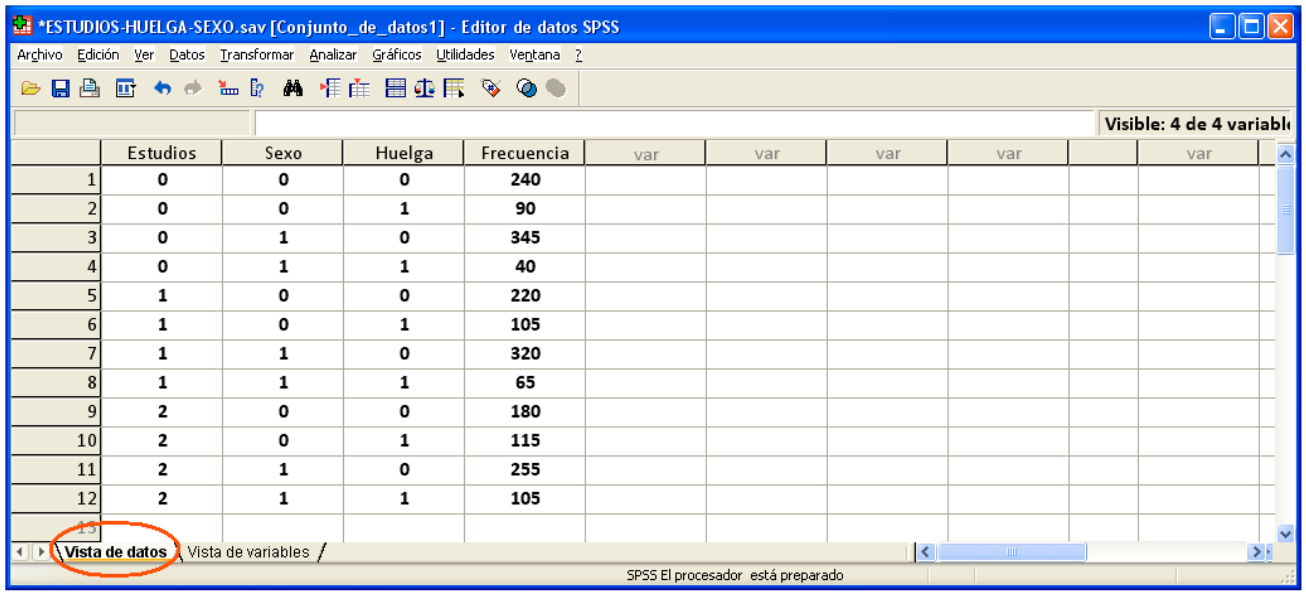
$$P(\text{participan en la huelga mujeres}) = P(X=1) = \frac{1}{1+e^{-(\beta_o + \beta_1 .1)}} = \frac{1}{1+e^{0,724896 + 0,75237}} = 0,185841$$

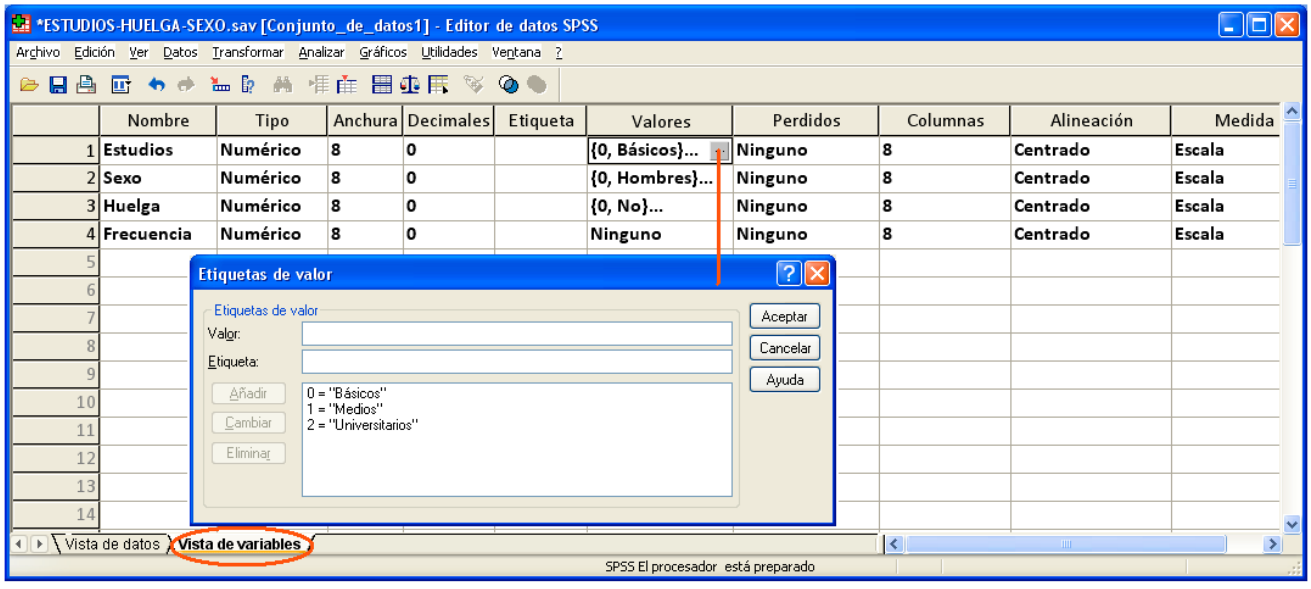
$$P(\text{participan en la huelga hombres}) = P(X=0) = \frac{1}{1+e^{-(\beta_o + \beta_1 .0)}} = \frac{1}{1+e^{0,724896 + 0}} = 0,326316$$

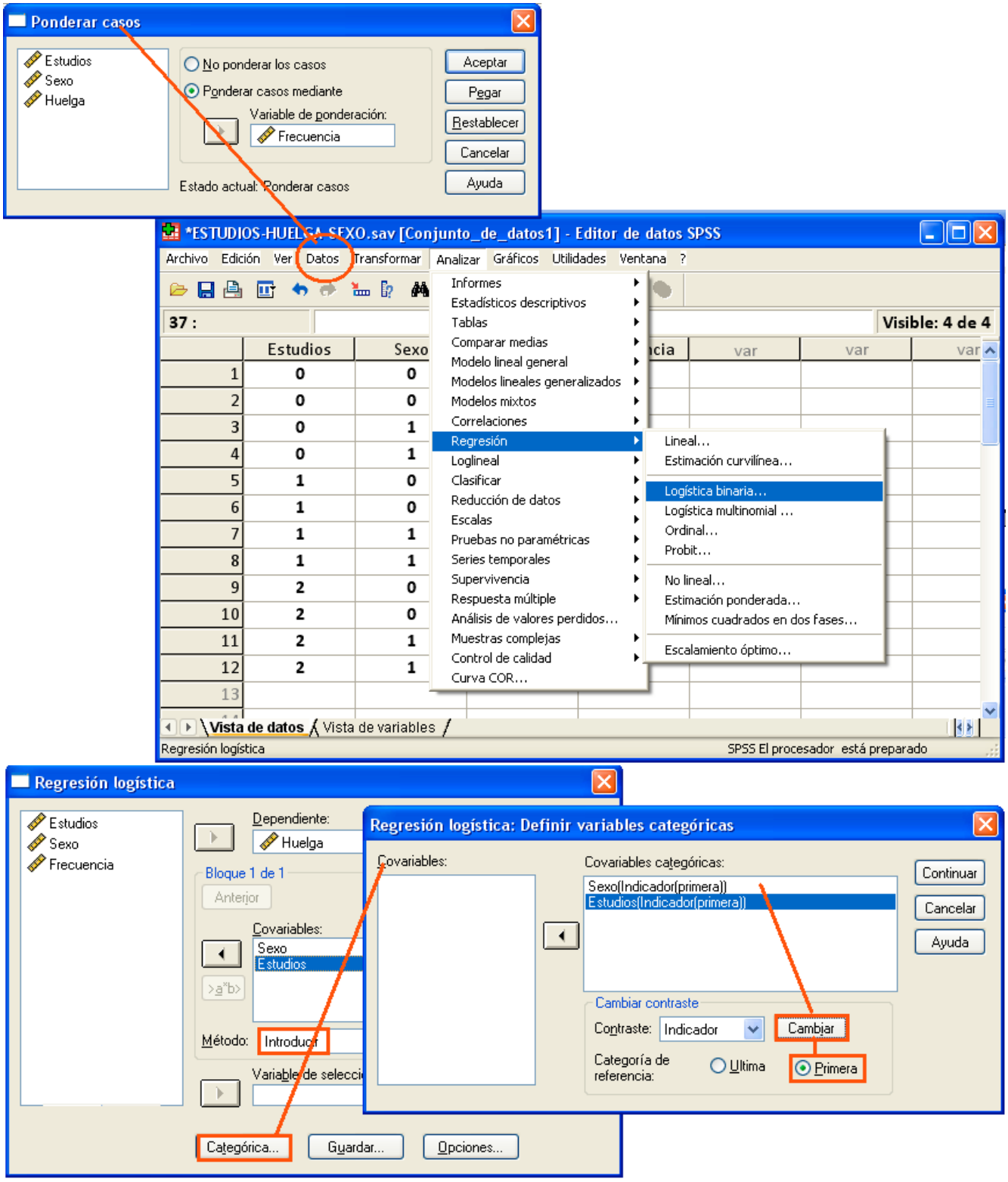
GENERALIZACIÓN DEL MODELO

Los modelos Logit permiten el análisis conjunto de un grupo de variables independientes. En la tabla adjunta se puede observar de forma conjunta el efecto del nivel de estudios y el sexo sobre la participación en la huelga.

		Huelga		
Estudios	Sexo	No Participan	Participan	Total
Básicos	Hombre	240	90	330
	Mujer	345	40	385
Medios	Hombre	220	105	325
	Mujer	320	65	385
Universitarios	Hombre	180	115	295
	Mujer	255	105	360
Total		1560	520	2080

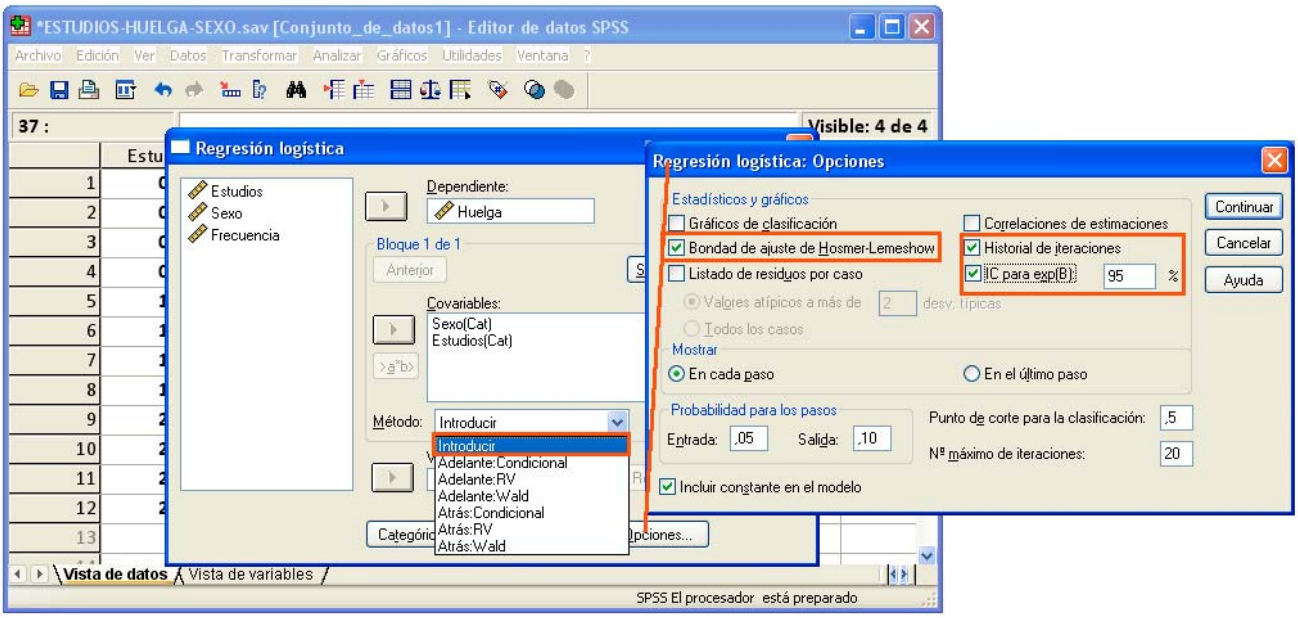






El método Introducir permite al investigador conducir el análisis en función de los resultados que va obteniendo. Cuando se realiza la regresión logística se debe especificar las variable o variables de control (covariables) que son categóricas.

Una vez que son seleccionadas hay que indicar cuál es el método de Contraste y cuál es la Categoría de referencia (la Última, por defecto). Si se desea cambiar algunas de ellas se debe aplicar la pestaña Cambiar y se observa cómo se modifican en la ventana Covariables categóricas.

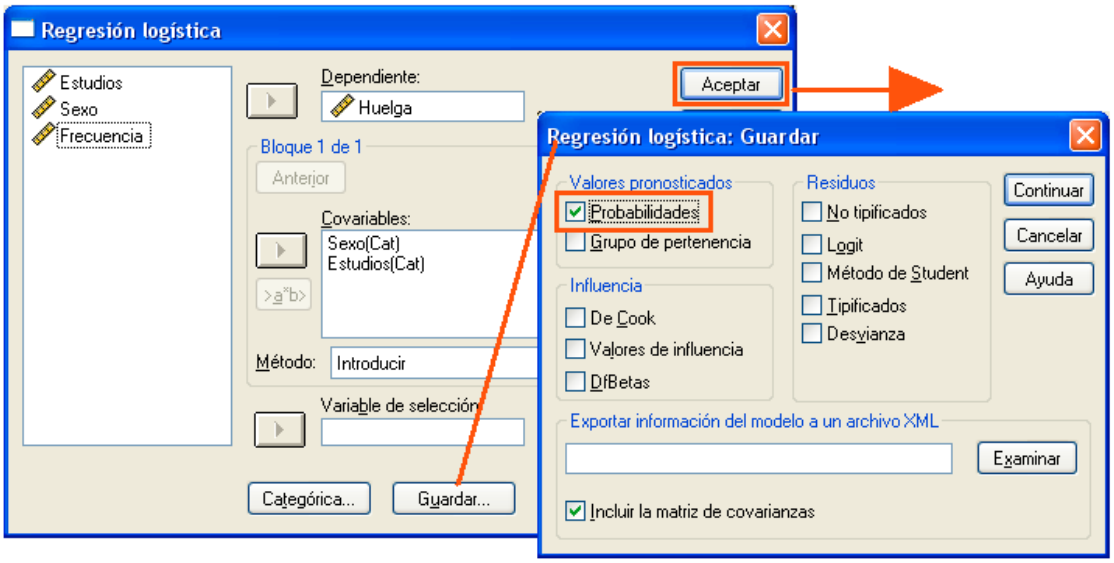


El estadístico de *Hosmer-Lemeshow* de bondad de ajuste es un método para evaluar el ajuste global del modelo, más robusto que el estadístico de bondad de ajuste tradicionalmente utilizado en la regresión logística, especialmente para los modelos con covariables continuas y los estudios con tamaños de muestra pequeños. Se basa en agrupar los casos en deciles de riesgo y comparar la probabilidad observada con la probabilidad esperada dentro de cada decil.

Listado de residuos por caso **muestra los residuos no estandarizados, la probabilidad pronosticada y los grupos de pertenencia observado y pronosticado.**

Historial de iteraciones **muestra los coeficientes y el logaritmo de la verosimilitud en cada iteración del proceso de estimación de los parámetros.**

Punto de corte para la clasificación **permite determinar el punto de corte para la clasificación de los casos. Los casos con valores pronosticados que han sobrepasado el punto de corte para la clasificación se clasifican como positivos (tendrían el evento o resultado que se modeliza), mientras que aquellos con valores pronosticados menores que el punto de corte se clasifican como negativos (no tendrían el evento o resultado)**



Se consultan en Vista de datos los Valores Pronosticados.

Las salidas directas en el Visor de SPSS para el análisis conjunto de sexo y niveles de estudios:

Historial de iteraciones[a,b,c]

Iteración		-2 log de la verosimilitud	Coeficientes: Constant
Paso 0	1	2343,168620	-1,000000
	2	2339,316218	-1,096339
	3	2339,314202	-1,098611
	4	2339,314202	-1,098612

a. En el modelo se incluye una constante.

b. -2 log de la verosimilitud inicial: 2339,314

c. La estimación ha finalizado en el número de iteración 4 porque las estimaciones de los parámetros han cambiado en menos de ,001.

El programa estadístico estima la constante β_0, para ello utiliza algoritmos iterativos que finalizan cuando los nuevos valores que se obtienen no se diferencian de los anteriores.
En este caso, realiza cuatro cálculos seguidos para el valor de la constante. Cuando la diferencia entre los valores es menor de 1/1000 el programa finaliza. Se observa que se estima el valor de $\beta_0 = -1,098612$ con una medida de verosimilitud $-2LL = 2339,314202$

Historial de iteraciones[a,b,c,d]

Iteración		-2 log de la verosimilitud	Coeficientes: Constant	Sexo(1)	Estudios(1)	Estudios(2)
Paso 1	1	2253,234325	-,967639	-,566592	,232622	,622560
	2	2239,407825	-1,118605	-,757672	,344520	,832539
	3	2239,312119	-1,133556	-,776275	,359861	,854482
	4	2239,312112	-1,133703	-,776424	,360026	,854683

a. Método: Introducir

b. En el modelo se incluye una constante.

c. -2 log de la verosimilitud inicial: 2339,314

d. La estimación ha finalizado en el número de iteración 4 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Después de varias iteraciones el programa ajusta el modelo con los coeficientes:
$\beta_0 = -1,133703$, $\beta_1 = -0,776424$, $\beta_2 = 0,360026$, $\beta_3 = 0,854683$

Este modelo tiene una verosimilitud $-2LL = 2239,312112$

La razón de verosimilitud (diferencia de los logaritmos) sigue una distribución Chi-cuadrado. La diferencia en verosimilitud es:
$\text{Chi} - \text{cuadrado} = (-2LL_{\text{MODELO 0}}) - (-2LL_{\text{MODELO 1}}) = 2339,314202 - 2239,312112 = 100,002090$

Prueba Omnibus, SSPS ofrece tres entradas (Paso, Bloque y Modelo):

Pruebas omnibus sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	100,002090	3	,000
	Bloque	100,002090	3	,000
	Modelo	100,002090	3	,000

La fila primera (Paso) es la correspondiente al cambio de verosimilitud $(-2LL)$ entre pasos sucesivos en la construcción del modelo, contrastando la hipótesis nula H_0 de que los coeficientes de las variables añadidas en el último paso son cero

La segunda fila (Bloque) es el cambio en $(-2LL)$ entre bloques de entrada sucesivos durante la construcción del modelo. Si como es habitual en la práctica se introducen las variables en un solo bloque, la Chi-cuadrado del Bloque es el mismo que la Chi-cuadrado del Modelo.

La tercera fila (Modelo) es la diferencia entre el valor de $(-2LL)$ para el modelo sólo con la constante y el valor de $(-2LL)$ para el modelo actual.

En el resumen de los modelos tres medidas *permiten* evaluar de forma global su validez.

Resumen de los modelos

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	2239,312112[a]	,046940	,069517

a. La estimación ha finalizado en el número de iteración 4 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Los coeficientes de determinación tienen valores muy pequeños, indicando que sólo el 4,6940% o el 6,9517% de la variación de la variable dependiente es explicada por las variables incluidas en el modelo, y debe mejorar cuando se vayan incluyendo variables más explicativas del resultado o términos de interacción.

- **– 2 logaritmo de la verosimilitud $(-2LL)$ mide hasta qué punto un modelo se ajusta bien a los datos. El resultado de esta medición recibe también el nombre de *desviación*. Cuanto más pequeño sea el valor, mejor será el ajuste.**

- **La R cuadrado de *Cox y Snell* es un coeficiente de determinación generalizado que se utiliza para estimar la proporción de varianza de la variable dependiente explicada por las variables predictoras (independientes). Los valores oscilan entre 0 y 1.**

- **La R cuadrado de *Nagelkerke* es una versión corregida de la R cuadrado de *Cox y Snell*. La R cuadrado de *Cox y Snell* tiene un valor máximo inferior a 1, incluso para un modelo "perfecto".**

La R cuadrado de *Nagelkerke* corrige la escala del estadístico para cubrir el rango completo de 0 a 1. Denotando por $L_s \equiv$ valor de la verosimilitud del modelo con todas las variables (modelo saturado) , $L_0 \equiv$ valor de verosimilitud que solo tiene la constante (modelo base), y $L_m \equiv$ valor de verosimilitud que considera determinada/s variable/s.

El efecto de la variable en el modelo viene dado por la diferencia en verosimilitud. Al manejar logaritmos la diferencia se convierte en *ratio*.

En este caso, se calcula el ratio entre los logaritmos de la verosimilitud del modelo con la variable sexo respecto al modelo solo con la constante, se tiene:

$$R^2_{McFadden} = \frac{Ln(L_m)}{Ln(L_0)} = \frac{-2239,312112/2}{-2339,314202/2} = \frac{-1119,656056}{-1169,657101} = 0,957251$$

El valor de $Ln(L_m)$ y $Ln(L_0)$ se obtiene al dividir entre (-2) los valores que ofrece SPSS para $(-2LL)$ en ambos pasos. A diferencia de lo que ocurre con el coeficiente de determinación de Pearson, el coeficiente de McFadden no es comparable entre distintos estudios, siendo solo orientativo para introducir o eliminar variables en función de su capacidad de ajuste.

El coeficiente R^2 de Cox y Snell en notación matemática es la media geométrica de la relación entre verosimilitudes:

$$R^2_{Cox-Snell} = 1 - \left(\frac{L_s}{L_0}\right)^{\frac{2}{n}}$$ **El valor del coeficiente está limitado por** $1 - (L_0)^{\frac{2}{n}}$

El coeficiente R^2 de Nagelkerke se resuelve: $R^2_{Nagelkerke} = \dfrac{1 - \left(\frac{L_s}{L_0}\right)^{\frac{2}{n}}}{1 - (L_0)^{\frac{2}{n}}}$

Los coeficientes R^2 de Cox y Snell y R^2 de Nagelkerke se pueden calcular con expresiones equivalentes a parir de los logaritmos de verosimilitud:

$$R^2_{Cox-Snell} = 1 - e^{\frac{[-2Ln(L_s) - 2Ln(L_0)]}{n}}$$ **donde el valor máximo de** $R^2_{Cox-Snell} = 1 - e^{\frac{2Ln(L_0)}{n}}$

$$R^2_{Nagelkerke} = \frac{1 - e^{\frac{[-2Ln(L_s) - 2Ln(L_0)]}{n}}}{1 - e^{\frac{2Ln(L_0)}{n}}}$$

En este caso,

$$R^2_{Cox-Snell} = 1 - e^{\frac{[-2Ln(L_s) - 2Ln(L_0)]}{n}} = 1 - e^{\frac{[2239,312112 - 2339,314202]}{2080}} = 1 - e^{-100,002090/2080} = 0,046940$$

Valor máximo de $R^2_{Cox-Snell} = 1 - e^{\frac{2Ln(L_0)}{n}} = 1 - e^{-\frac{2339,314202}{2080}} = 0,675240$

$$R^2_{Nagelkerke} = \frac{1 - e^{\frac{[-2Ln(L_s) - 2Ln(L_0)]}{n}}}{1 - e^{\frac{2Ln(L_0)}{n}}} = \frac{1 - e^{\frac{[2239,312112 - 2339,314202]}{2080}}}{1 - e^{-\frac{2339,314202}{2080}}} = \frac{0,046940}{0,675240} = 0,069517$$

Prueba de Hosmer y Lemeshow

Pas	Chi-cuadrado	gl	Sig.
1	7,989	4	,092

Tabla de contingencias para la prueba de Hosmer y Lemeshow

		Huelga = No		Huelga = Si		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	345	335,348	40	49,652	385
	2	320	317,597	65	67,403	385
	3	240	249,652	90	80,348	330
	4	255	267,055	105	92,945	360
	5	220	222,403	105	102,597	325
	6	180	167,945	115	127,055	295

El modelo logístico se ajusta bien a los datos si $p\_valor = Sig. > \alpha$ aceptando la hipótesis nula de que no hay diferencia entre los valores observados y los valores pronosticados

La prueba de *Hosmer-Lemeshow* es otra prueba para evaluar la *bondad del ajuste* de un modelo de regresión logística (RL).

Parte de la idea de que si el ajuste es bueno, un valor alto de la probabilidad predicha (p) se asociará con el resultado 1 de la variable binomial dependiente, mientras que un valor bajo de p (próximo a cero) corresponderá (en la mayoría de las ocasiones) con el resultado Y = 0

Para cada observación del conjunto de datos, se trata de calcular las probabilidades de la variable dependiente que predice el modelo, ordenarlas, agruparlas y calcular, a partir de ellas, las frecuencias esperadas, y compararlas con las observadas mediante una prueba Chi-cuadrado. Señalar que esta prueba de *bondad de ajuste* tiene algunas 'inconvenientes': El estadígrafo de *Hosmer-Lemeshow* no se computa cuando, para algunos grupos, e_i (valores esperados) ó $e_i \times (O_i - e_i)$ son nulos o muy pequeños (menores que 5).

Por otra parte, lo que se desea en esta prueba es que no haya significación (lo contrario a lo que suele ser habitual). Por este motivo, muchos autores proponen simplemente cotejar valores observados y esperados mediante simple inspección y evaluar el grado de concordancia entre unos y otros a partir del sentido común.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	Sexo(1)	-,776424	,104486	55,2186	1	,000	,460048	,374857	,564601
	Estudios			45,0572	2	,000			
	Estudios(1)	,360026	,132545	7,37804	1	,007	1,43337	1,10544	1,85857
	Estudios(2)	,854683	,129412	43,6177	1	,000	2,35063	1,82402	3,02928
	Constante	-1,1337	,107233	111,773	1	,000	,321839		

a. Variable(s) introducida(s) en el paso 1: Sexo, Estudios.

SPSS ofrece las variables que dejará en la ecuación, coeficientes de regresión con sus correspondientes errores estándar, el valor del estadístico Wald para evaluar la hipótesis nula $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$, la significación o p_valor asociado, y el valor $OR \equiv \exp(B_i)$ con el intervalo de confianza.

→ Estimador $\hat{\beta}_1 = -0,776424 \Rightarrow e^{-0,776424} = 0,460048$

→ Estimador $\hat{\beta}_2 = 0,360026 \Rightarrow e^{0,360026} = 1,43337$

→ Estimador $\hat{\beta}_3 = 0,854683 \Rightarrow e^{0,854683} = 2,35063$

→ Estimador $\hat{\beta}_0 = -1,1337 \Rightarrow e^{-1,1337} = 0,321839$

→ El estadístico de Wald $W^2 = \left(\frac{\hat{\beta}_i}{\sigma(\hat{\beta}_i)} \right)^2 = \left(\frac{\hat{\beta}_i}{E.T.(\hat{\beta}_i)} \right)^2$ bajo la hipótesis nula sigue una Chi-cuadrado con 1 grado de libertad.

Cuando $p\_valor \leq 0,05$ se rechaza la hipótesis nula $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

$$W^2_{constante} = \left(\frac{\hat{\beta}_0}{\sigma(\hat{\beta}_0)}\right)^2 = \left(\frac{-1{,}1337}{0{,}107233}\right)^2 = 111{,}773 \quad W^2_{sexo} = \left(\frac{\hat{\beta}_1}{\sigma(\hat{\beta}_1)}\right)^2 = \left(\frac{0{,}360026}{0{,}132545}\right)^2 = 7{,}37804$$

$$W^2_{medio} = \left(\frac{\hat{\beta}_2}{\sigma(\hat{\beta}_2)}\right)^2 = \left(\frac{0{,}854683}{0{,}129412}\right)^2 = 43{,}6177 \quad W^2_{universitario} = \left(\frac{\hat{\beta}_3}{\sigma(\hat{\beta}_3)}\right)^2 = \left(\frac{0{,}854683}{0{,}129412}\right)^2 = 43{,}6177$$

→ **Intervalos de confianza:**

$$I.C(\hat{\beta}_1) = \left[e^{\hat{\beta}_1 \pm z_{\alpha/2}\cdot\sigma(\hat{\beta}_1)}\right] = \left[e^{-0{,}776424 - 1{,}96\,.\,0{,}104486},\ e^{-0{,}776424 + 1{,}96\,.\,0{,}104486}\right] = [0{,}374857\ ,\ 0{,}564601]$$

$$I.C(\hat{\beta}_2) = \left[e^{\hat{\beta}_2 \pm z_{\alpha/2}\cdot\sigma(\hat{\beta}_2)}\right] = \left[e^{0{,}360026 - 1{,}96\,.\,0{,}132545},\ e^{0{,}360026 + 1{,}96\,.\,0{,}132545}\right] = [1{,}10544\ ,\ 1{,}85857]$$

$$I.C(\hat{\beta}_3) = \left[e^{\hat{\beta}_3 \pm z_{\alpha/2}\cdot\sigma(\hat{\beta}_3)}\right] = \left[e^{0{,}854683 - 1{,}96\,.\,0{,}129412},\ e^{0{,}854683 + 1{,}96\,.\,0{,}129412}\right] = [1{,}82402\ ,\ 3{,}02928]$$

→ **Ecuación regresión logística:**

$$z = \beta_0 + \beta_1\,.\,sexo + \beta_2\,.\,estudios\_medios + \beta_3\,.\,estudios\_universitarios =$$
$$= -1{,}1337 - 0{,}776424\,.\,sexo + 0{,}360026\,.\,estudios\_medios + 0{,}854683\,.\,estudios\_universitarios$$

$$p = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(-1{,}1337\ -\ 0{,}776424\,.\,sexo\ +\ 0{,}360026\,.\,estudios\_medios\ +\ 0{,}854683\,.\,estudios\_universitarios)}}$$

$$p_{hombre}(Huelga/Estu\_básicos) = \frac{1}{1+e^{-(-1{,}1337\ -\ 0{,}776424\,.\,0\ +\ 0{,}360026\,.\,0\ +\ 0{,}854683\,.\,0)}} = \frac{1}{4{,}10713} = 0{,}24348$$

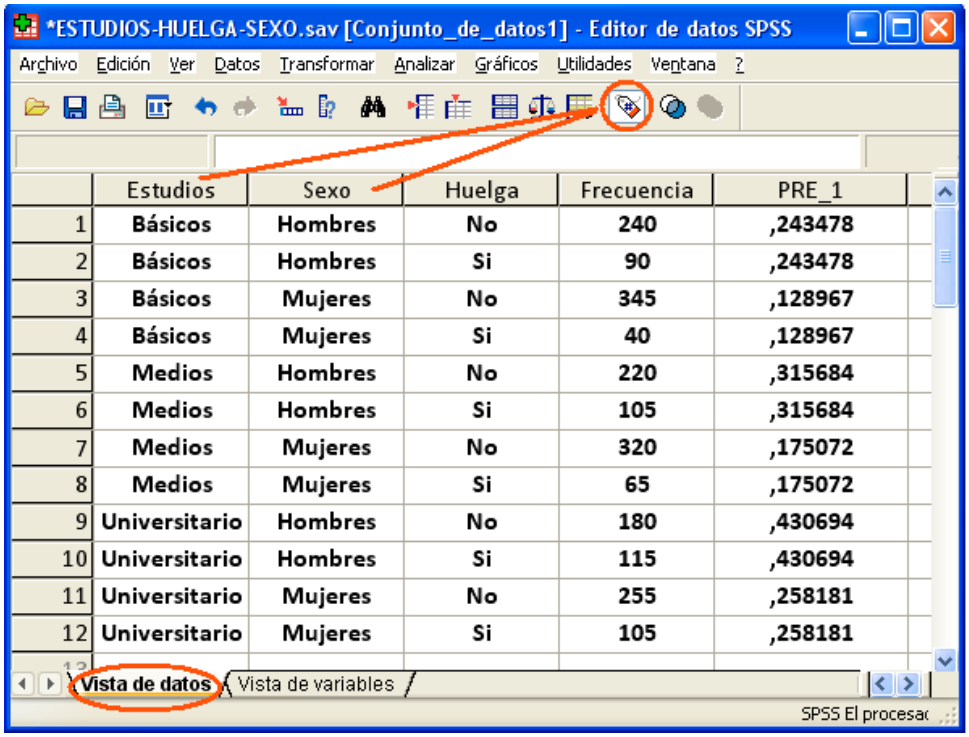
$$p_{hombre}(Huelga/Estu\_medios) = \frac{1}{1+e^{-(-1{,}1337\ -\ 0{,}776424\,.\,0\ +\ 0{,}360026\,.\,1\ +\ 0{,}854683\,.\,0)}} = \frac{1}{3{,}167716} = 0{,}315684$$

$$p_{hombre}(Huelga/Estu\_universitarios) = \frac{1}{1+e^{-(-1{,}1337\ -\ 0{,}776424\,.\,0\ +\ 0{,}360026\,.\,0\ +\ 0{,}854683\,.\,1)}} = \frac{1}{2{,}321829} = 0{,}430694$$

$$p_{mujer}(Huelga/Estu\_básicos) = \frac{1}{1+e^{-(-1{,}1337\ -\ 0{,}776424\,.\,1\ +\ 0{,}360026\,.\,0\ +\ 0{,}854683\,.\,0)}} = \frac{1}{7{,}753926} = 0{,}128967$$

$$p_{mujer}(Huelga/Estu\_medios) = \frac{1}{1+e^{-(-1{,}1337\ -\ 0{,}776424\,.\,1\ +\ 0{,}360026\,.\,1\ +\ 0{,}854683\,.\,0)}} = \frac{1}{5{,}711932} = 0{,}175072$$

$$p_{mujer}(Huelga/Estu\_universitarios) = \frac{1}{1+e^{-(-1{,}1337\ -\ 0{,}776424\,.\,1\ +\ 0{,}360026\,.\,0\ +\ 0{,}854683\,.\,1)}} = \frac{1}{3{,}873242} = 0{,}258181$$



	Estudios	Sexo	Huelga	Frecuencia	PRE_1
1	Básicos	Hombres	No	240	,243478
2	Básicos	Hombres	Si	90	,243478
3	Básicos	Mujeres	No	345	,128967
4	Básicos	Mujeres	Si	40	,128967
5	Medios	Hombres	No	220	,315684
6	Medios	Hombres	Si	105	,315684
7	Medios	Mujeres	No	320	,175072
8	Medios	Mujeres	Si	65	,175072
9	Universitario	Hombres	No	180	,430694
10	Universitario	Hombres	Si	115	,430694
11	Universitario	Mujeres	No	255	,258181
12	Universitario	Mujeres	Si	105	,258181

El caso más favorable para hacer huelga sería un hombre con estudios universitarios, con la probabilidad asignada de 0,430694.

Con el archivo *dispepsia.sav* con 497 controles se quiere hacer una regresión logística con la variable dependiente *Ardor* y las covariables (*Edad*, *Sexo*, *Tabaco* y *Estreñimiento*)

dispepsia.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

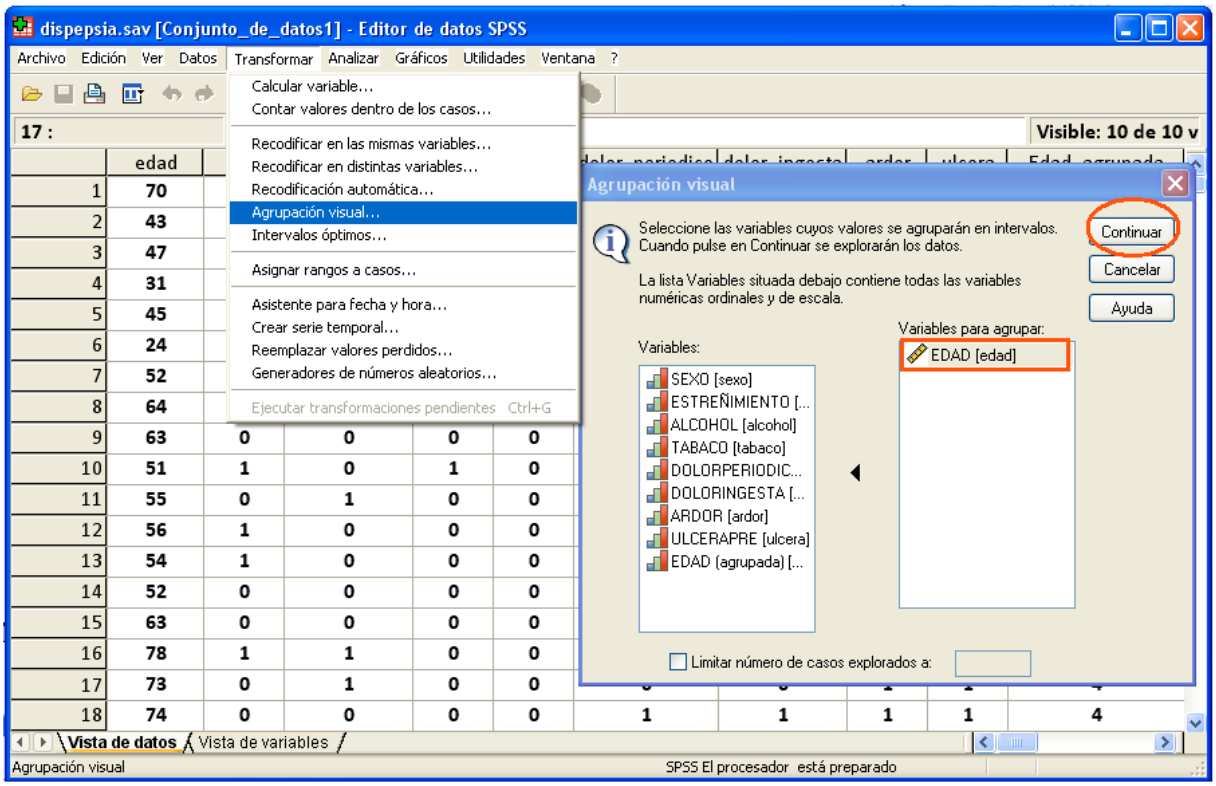
497 : sexo | 1 | Visible: 10 de 1

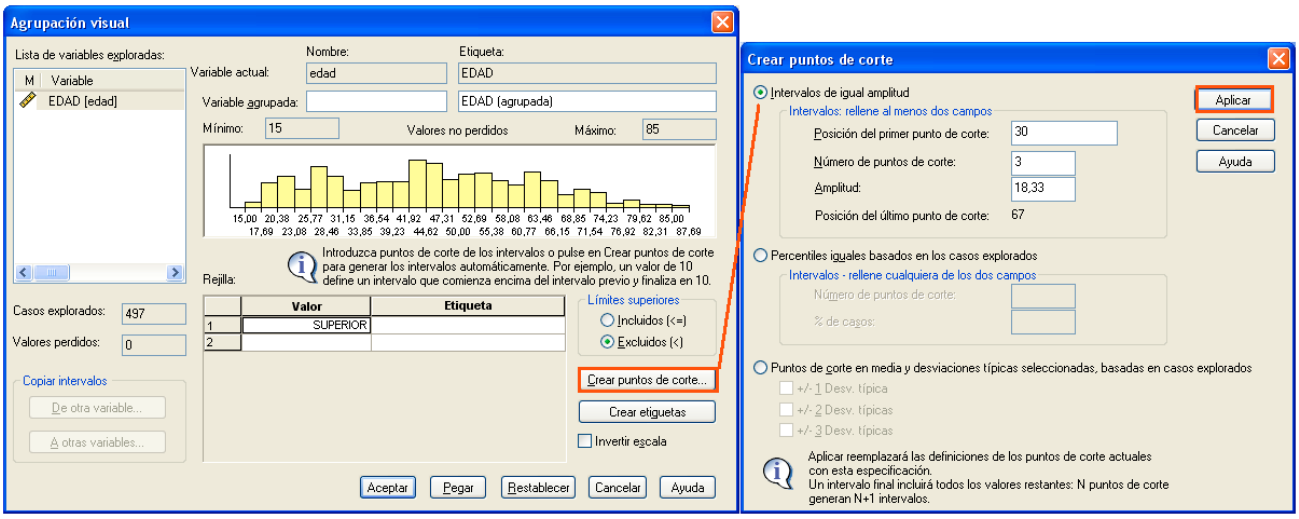
	edad	sexo	estreñimiento	alcohol	tabaco	dolor_periodico	dolor_ingesta	ardor	ulcera
1	70	0	1	0	0	0	0	1	0
2	43	0	0	0	0	0	0	0	0
3	47	0	0	0	0	0	0	1	0
4	31	0	0	0	0	0	1	1	0
5	45	0	1	0	1	0	0	1	0
6	24	1	1	0	0	0	0	1	0
7	52	0	1	0	0	0	0	1	0
8	64	1	0	0	1	0	0	1	0
9	63	0	0	0	0	0	0	1	0
10	51	1	0	1	0	1	1	0	0
11	55	0	1	0	0	0	0	0	0
12	56	1	0	0	0	0	0	1	0
13	54	1	0	0	0	0	0	1	0
14	52	0	0	0	0	0	0	0	0
15	63	0	0	0	0	0	1	1	0
16	78	1	1	0	0	0	0	1	0
17	73	0	1	0	0	0	0	1	1
18	74	0	0	0	0	1	1	1	1
19	42	0	1	0	0	0	0	1	0
20	64	0	1	0	0	0	0	0	0

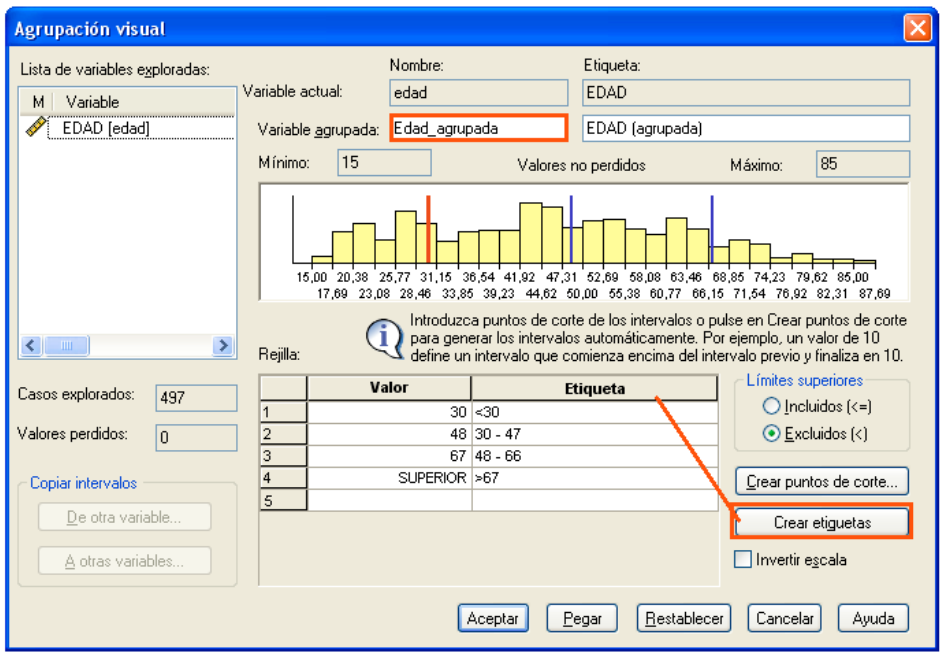
Vista de datos | Vista de variables

SPSS El procesador está preparado

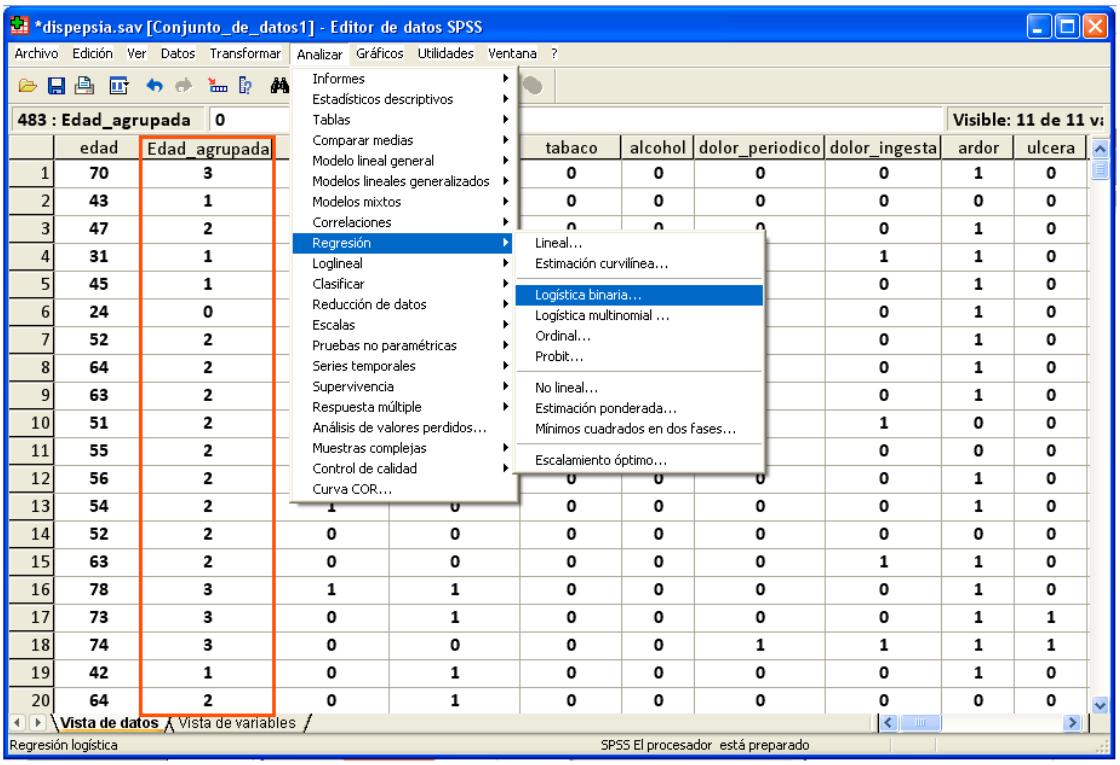
La mecánica comienza en agrupar en intervalos la variable Edad.

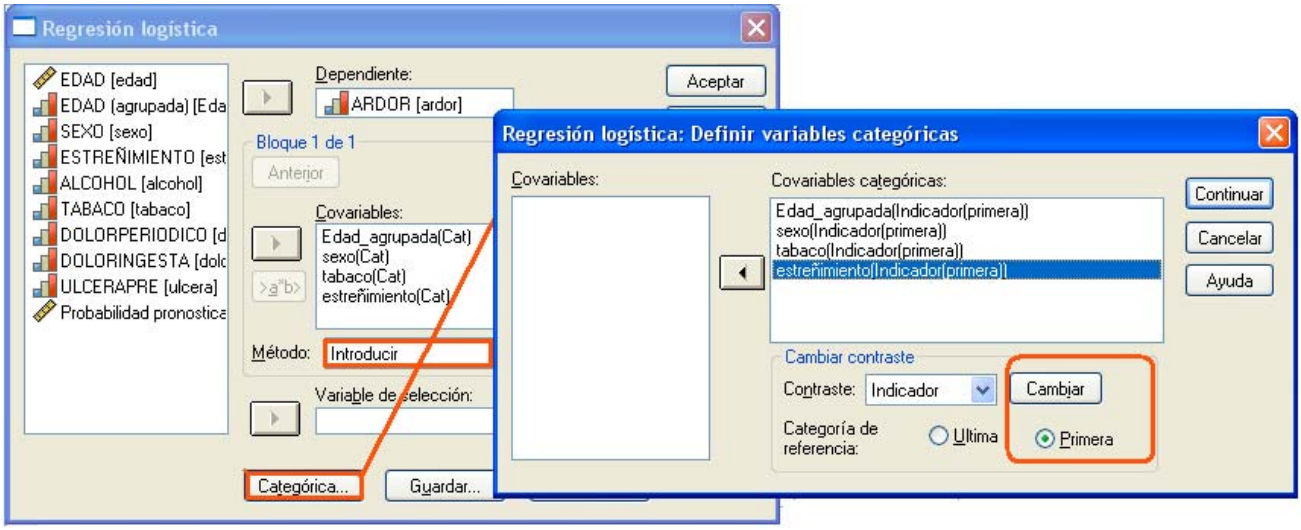






En el Visor de Datos se crea la variable Edad_agrupada





Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	Edad_agrupada			1,55134	3	,670473			
	Edad_agrupada(1)	,245618	,262378	,876321	1	,349211	1,27841	,764423	2,13800
	Edad_agrupada(2)	,034941	,269330	,016831	1	,896777	1,03556	,610830	1,75561
	Edad_agrupada(3)	-,091600	,357871	,065514	1	,797983	,912470	,452479	1,84009
	sexo(1)	,331585	,212419	2,43671	1	,118524	1,39317	,918744	2,11260
	tabaco(1)	-,154466	,269421	,328702	1	,566424	,856873	,505342	1,45294
	estreñimiento(1)	-,059000	,210552	,078521	1	,779312	,942707	,623956	1,42429
	Constante	,562613	,244817	5,28124	1	,021556	1,75525		

a. Variable(s) introducida(s) en el paso 1: Edad_agrupada, sexo, tabaco, estreñimiento.

Estimadores: $\beta_0 = 0{,}562613$ $\beta_1 = 0{,}245618$ $\beta_2 = 0{,}034941$ $\beta_3 = -0{,}091600$
$\beta_4 = 0{,}331585$ $\beta_5 = -0{,}154466$ $\beta_6 = -0{,}059000$

Valores de las variables:

$$\text{Edad\_agrupada} \begin{cases} 0 \equiv (\text{personas} < 30 \text{ años}) & 1 \equiv (30 \text{ años} \le \text{personas} < 47 \text{ años}) \\ 2 \equiv (47 \text{ años} \le \text{personas} < 67 \text{ años}) & 3 \equiv (\text{personas} \ge 67 \text{ años}) \end{cases}$$

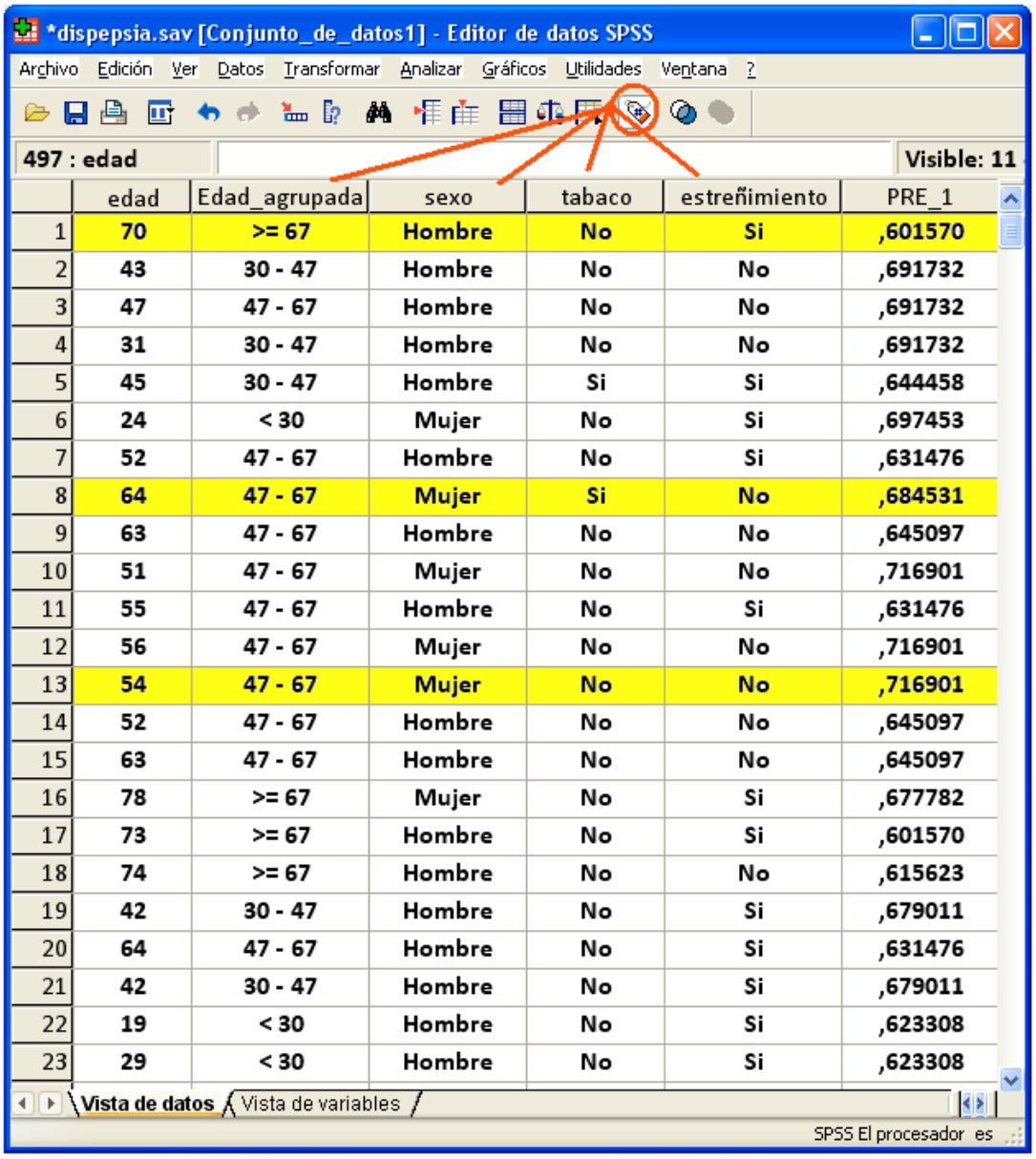
$$\text{Sexo} \begin{cases} 0 \equiv \text{Hombre} \\ 1 \equiv \text{Mujer} \end{cases} \quad \text{Tabaco} \begin{cases} 0 \equiv \text{Fuma} \\ 1 \equiv \text{No Fuma} \end{cases} \quad \text{Estreñimiento} \begin{cases} 0 \equiv \text{No} \\ 1 \equiv \text{Si} \end{cases}$$

→ **regresión logística:**

$$z = 0{,}562613 + 0{,}245618 \cdot edad_{[30,47)} + 0{,}034941 \cdot edad_{[47,67)} - 0{,}0916 \cdot edad_{\ge 67} + 0{,}331585 \cdot sexo - 0{,}154466 \cdot tabaco - 0{,}059 \cdot estreñimiento$$

Ecuación regresión logística:

$$p = \frac{1}{1 + e^{-(0{,}562613 + 0{,}245618 \cdot edad_{[30,47)} + 0{,}034941 \cdot edad_{[47,67)} - 0{,}0916 \cdot edad_{\ge 67} + 0{,}331585 \cdot sexo - 0{,}154466 \cdot tabaco - 0{,}059 \cdot estreñimiento)}}$$



	edad	Edad_agrupada	sexo	tabaco	estreñimiento	PRE_1
1	70	>= 67	Hombre	No	Si	,601570
2	43	30 - 47	Hombre	No	No	,691732
3	47	47 - 67	Hombre	No	No	,691732
4	31	30 - 47	Hombre	No	No	,691732
5	45	30 - 47	Hombre	Si	Si	,644458
6	24	< 30	Mujer	No	Si	,697453
7	52	47 - 67	Hombre	No	Si	,631476
8	64	47 - 67	Mujer	Si	No	,684531
9	63	47 - 67	Hombre	No	No	,645097
10	51	47 - 67	Mujer	No	No	,716901
11	55	47 - 67	Hombre	No	Si	,631476
12	56	47 - 67	Mujer	No	No	,716901
13	54	47 - 67	Mujer	No	No	,716901
14	52	47 - 67	Hombre	No	No	,645097
15	63	47 - 67	Hombre	No	No	,645097
16	78	>= 67	Mujer	No	Si	,677782
17	73	>= 67	Hombre	No	Si	,601570
18	74	>= 67	Hombre	No	No	,615623
19	42	30 - 47	Hombre	No	Si	,679011
20	64	47 - 67	Hombre	No	Si	,631476
21	42	30 - 47	Hombre	No	Si	,679011
22	19	< 30	Hombre	No	Si	,623308
23	29	< 30	Hombre	No	Si	,623308

$$p = \frac{1}{1+e^{-(0{,}562613 + 0{,}245618 \cdot edad_{[30,47)} + 0{,}034941 \cdot edad_{[47,67)} - 0{,}0916 \cdot edad_{\geq 67} + 0{,}331585 \cdot sexo - 0{,}154466 \cdot tabaco - 0{,}059 \cdot estreñimiento)}}$$

✓ Probabilidad de que un hombre con 70 años tenga ardor si no fuma y tiene estreñimiento:

$$p_{Hombre}(Ardor) = \frac{1}{1+e^{-(0{,}562613 + 0{,}245618 \cdot 0 + 0{,}034941 \cdot 0 - 0{,}0916 \cdot 1 + 0{,}331585 \cdot 0 - 0{,}154466 \cdot 0 - 0{,}059 \cdot 1)}} = \frac{1}{1{,}662315} = 0{,}601570$$

✓ Probabilidad de una mujer de 64 años tenga ardor si fuma y no tiene estreñimiento:

$$p = \frac{1}{1+e^{-(0{,}562613 + 0{,}245618 \cdot 0 + 0{,}034941 \cdot 1 - 0{,}0916 \cdot 0 + 0{,}331585 \cdot 1 - 0{,}154466 \cdot 1 - 0{,}059 \cdot 0)}} = \frac{1}{1{,}460854} = 0{,}684531$$

✓ Probabilidad de que una mujer con 54 años tenga ardor si no fuma y no tiene estreñimiento:

$$p_{Mujer}(Ardor) = \frac{1}{1+e^{-(0{,}562613 + 0{,}245618 \cdot 0 + 0{,}034941 \cdot 1 - 0{,}0916 \cdot 0 + 0{,}331585 \cdot 1 - 0{,}154466 \cdot 0 - 0{,}059 \cdot 0)}} = \frac{1}{1{,}394893} = 0{,}716901$$


Facultad de Ciencias
Económicas y Empresariales
ACCESO NORTE