



📁 Se quiere comparar las puntuaciones en estadística teórica de dos Universidades que siguen, respectivamente, una distribución normal, con varianzas desconocidas y se supone que estadísticamente iguales. Se toman al azar dos muestras iguales de estudiantes de ambas Universidades. Los resultados obtenidos en SPSS son:

SPSS: [Analizar/Comparar medias/ Prueba T para dos muestras independientes](#)

Prueba de muestras independientes

	Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
	F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ de la diferencia	95% Intervalo de confianza para la diferencia	
								Inferior	Superior
Calificación Se han asumido varianzas iguales	,152	,699	,443	28	,661	2,200	4,962	-7,964	12,364

Con un nivel de significación de 0,05, se pide:

- ¿Existe diferencia significativa sobre las puntuaciones de estadística teórica en las dos Universidades?
- ¿Cuál es la media ponderada de las cuasivarianzas muestrales?
- ¿Cuál sería el p-valor si se hubieran contrastado las puntuaciones

$$H_0 : \mu_{\text{Autónoma}} \leq \mu_{\text{Complutense}} \quad H_1 : \mu_{\text{Autónoma}} > \mu_{\text{Complutense}}$$

Solución:

- En los resultados obtenidos en SPSS la Prueba de Levene permite decidir si las varianzas poblacionales son iguales. Como la probabilidad asociada al estadístico de Levene (0,699) es mayor que 0,05 se acepta la hipótesis nula de que las varianzas poblacionales son iguales.

$$I_{1-\alpha}(\mu_1 - \mu_2) = \left[(\bar{x} - \bar{y}) \pm t_{\alpha/2, (n_1 + n_2 - 2)} \cdot \underbrace{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}_{\text{error típico diferencia}} \right]$$

Las dos Universidades tienen la misma puntuación en estadística teórica a un nivel de significación del 0,05 ya que el intervalo de confianza cubre el cero.

Un intervalo de confianza sirve para tomar una decisión sobre la hipótesis nula que permite contrastar el estadístico T (t de Student).

Por otra parte,

p-valor = Sig.(bilateral) = 0,661 > 0,05 \mapsto **Se acepta la hipótesis nula que establece $H_0: \mu_{\text{Autónoma}} = \mu_{\text{Complutense}}$**

b) $s_p^2 \equiv$ Media ponderada de las cuasivarianzas muestrales:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\text{Error típico de la diferencia} = s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \mapsto 4,962 = s_p \cdot \sqrt{\frac{1}{15} + \frac{1}{15}}$$

$$s_p = \frac{4,962}{\sqrt{\frac{1}{15} + \frac{1}{15}}} = \frac{4,962}{0,3651} = 13,588 \mapsto s_p^2 = 184,633$$

t \equiv valor experimental del estadístico de contraste:

$$t = \frac{\text{Diferencia de medias}}{\text{Error típico diferencia}} = \frac{|\bar{x} - \bar{y}|}{13,588 \cdot \sqrt{\frac{1}{15} + \frac{1}{15}}} = \frac{2,200}{4,962} = 0,443$$

t = 0,443 \leq $t_{0,025, 28} = 2,048$ se acepta H_0

p-valor = Sig.(bilateral) = 0,661. La Significación bilateral muestra el grado de compatibilidad entre el valor poblacional propuesto y la información muestral disponible.


Si el nivel crítico es pequeño (generalmente menor que 0,05), se concluye que la información recogida en la muestra es incompatible con la hipótesis nula de que las medias poblacionales son iguales.

La Prueba T se basa en el supuesto de normalidad, el cumplimiento de este supuesto sólo es exigible con muestras pequeñas. El supuesto de normalidad carece de relevancia en muestras grandes.

c) SPSS siempre calcula el p-valor para un contraste bilateral o de dos colas, si se desea realizar un contraste unilateral o de una cola hay que dividir entre 2 el contraste bilateral (p-valor = 0,699 / 2 = 0,3495)

Contraste unilateral (cola a la derecha: $H_1 : \mu_{\text{Autónoma}} > \mu_{\text{Complutense}}$)

p-valor = Sig.(unilateral derecha) = 0,3495 > 0,05 \mapsto Se acepta H_0

 Las puntuaciones de estadística teórica se distribuyen según una ley normal, con μ y σ desconocidas. Se desea contrastar $H_0 : \mu = 74$ frente a $H_1 : \mu \neq 74$. Para ello, se introducen en SPSS las puntuaciones de quince alumnos escogidos aleatoriamente. Los resultados han sido:

SPSS: Analizar/Comparar medias/ Prueba T para una muestra

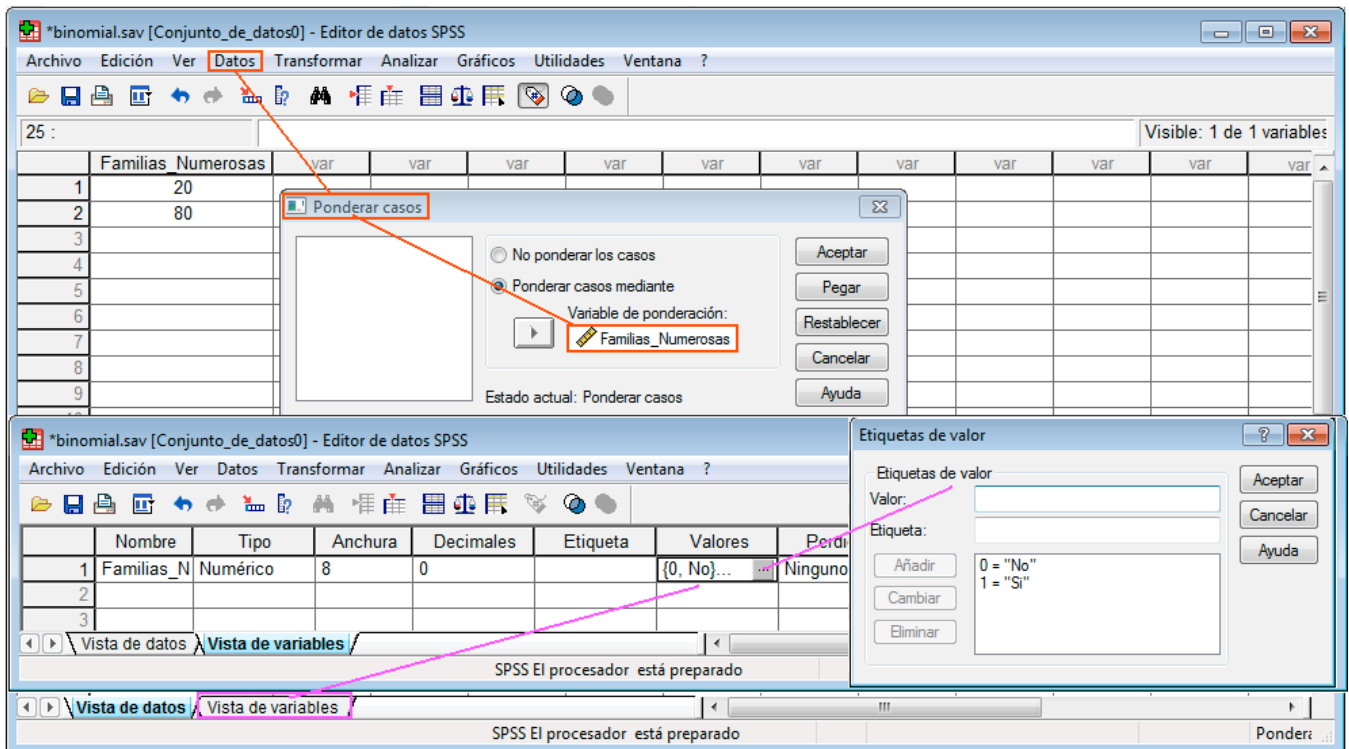
Prueba para una muestra

	Valor de prueba = 74					
	t	gl	Sig. (bilateral)	Diferencia de medias	95% Intervalo de confianza para la diferencia	
					Inferior	Superior
Autónoma	1,923	14	,075	7,200	-,83	15,23

Con un nivel de significación de 0,05, se pide:

- ¿Qué se puede decir del contraste?
- Media y cuasidesviación típica muestral
- Define el p-valor y como calcular la significación bilateral
- Calcula el p-valor al contrastar $H_0 : \mu \leq 74$ frente a $H_1 : \mu > 74$
- Calcula el p-valor al contrastar $H_0 : \mu \geq 74$ frente a $H_1 : \mu < 74$

Solución:



a) El intervalo de confianza $[-0,83, 15,23]$ cubre el cero, se puede afirmar que la muestra procede de una población $N(74, \sigma)$, con un nivel de significación de 0,05

Por otro lado,

p - valor = Sig.(bilateral) = 0,075 > 0,05 y se acepta la hipótesis nula $H_0: \mu = 74$

$$b) I_{1-\alpha}(\mu) = \left[\bar{x} \pm t_{\alpha/2, (n-1)} \frac{s_x}{\sqrt{n}} \right]$$

$$\bar{x} - 74 = 7,2 \rightarrow \bar{x} = 81,2$$

t ≡ valor experimental del estadístico de contraste:

$$t = \frac{|\bar{x} - 74|}{s / \sqrt{15}} \mapsto \frac{s_x}{\sqrt{15}} = \frac{|\bar{x} - 74|}{t} = \frac{7,2}{1,923} = 3,743$$

$$\text{Error típico de la media: } \frac{s_x}{\sqrt{15}} = 3,743 \mapsto s_x = 14,496$$

c) La Significación bilateral (p-valor) muestra el grado de compatibilidad entre el valor poblacional propuesto y la información muestral disponible. Si el nivel crítico es pequeño (generalmente menor que 0,05), se concluye que la información recogida en la muestra es incompatible con la hipótesis nula de que la muestra procede de la población.

$$\alpha_p = \text{p-valor} = P[\text{Rechazar } \bar{x} \mid H_0 \text{ es cierta}]$$

$$\alpha_p = \text{p-valor} = P[|t_{14}| > 1,923] = 2 \cdot P[t_{14} > 1,923] = 0,075$$

d) SPSS siempre calcula el p-valor para un contraste bilateral o de dos colas, si se desea realizar un contraste unilateral o de una cola hay que dividir entre 2 el contraste bilateral (p-valor = 0,075 / 2 = 0,0375)

$$\alpha_p = \text{p-valor} = P[t_{14} > 1,923] = 0,0375$$


Se trata de un contraste unilateral (cola a la derecha: $H_1 : \mu > 74$)

$$\text{p-valor} = \text{Sig. (unilateral derecha)} = 0,0375 < 0,05 \quad \mapsto \quad \text{Se rechaza } H_0$$

e) Se trata de un contraste unilateral (cola a la izquierda: $H_1 : \mu < 74$)

$$\text{p-valor} = \text{Sig. (unilateral izquierda)} = 1 - 0,0375 = 0,9625 > 0,05$$

\mapsto Se acepta H_0

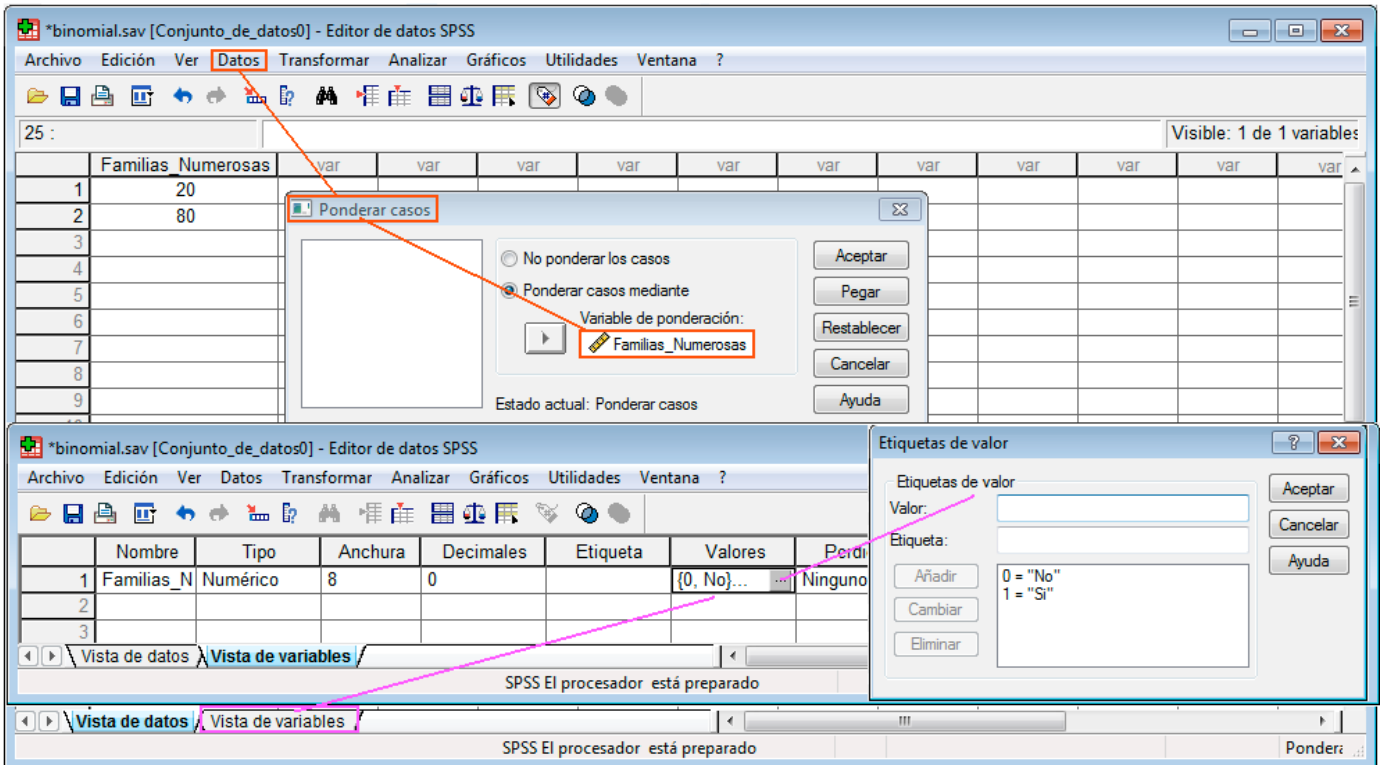
 Para determinar el número de familias numerosas de Fuentesrebollo se toma una muestra de 100 familias, resultando una proporción de 0,20. Con un nivel de significación 0,05, ¿se puede afirmar que la proporción de familias numerosas es 0,25?

Solución:

Se analiza el contraste: $H_0 : p = 25$ frente a $H_1 : p \neq 25$

Se procede a introducir los datos en SPSS:

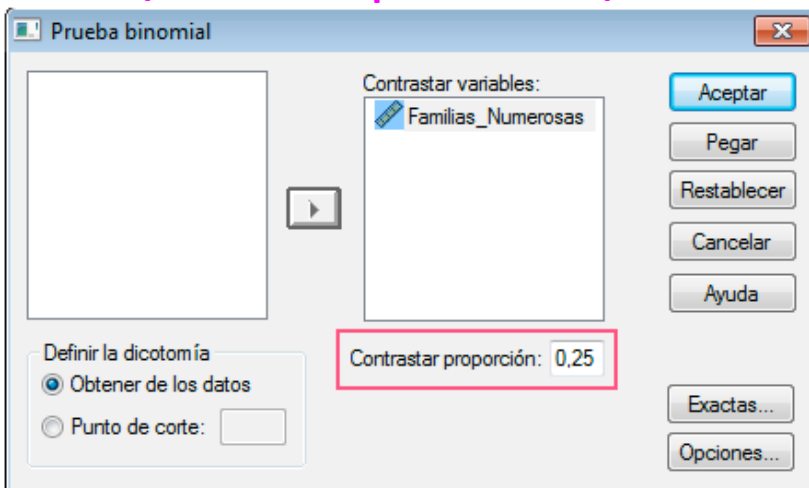
Numerosas	20
No Numerosas	80



Introducir las familias, en Vista de variables se clasifican con valores las familias numerosas: (0, No) y (1, Sí).

Ponderar los datos: Datos/Ponderar casos

Analizar/Pruebas no paramétricas/Binomial



Prueba binomial

		Categoría	N	Proporción observada	Prop. de prueba	Sig. asintót. (unilateral)
Famílias_Numerosas	Grupo 1	20	20	,20	,25	,149 ^{a, b}
	Grupo 2	80	80	,80		
	Total		100	1,00		


a. La hipótesis alternativa establece que la proporción de casos del primer grupo sea $< ,25$.

b. Basado en la aproximación Z.

El p-valor de la Prueba (Sig. Unilateral) es 0,149

p-valor = Sig.(unilateral) = 0,149 > 0,05 → Se acepta H_0

Con una fiabilidad del 95% se puede afirmar que la proporción de familias numerosas en Fuenterrebollo es 0,25.

 Un equipo de investigación biológica está interesado en ver si una nueva droga reduce el colesterol en la sangre. Con tal fin toma una muestra de diez pacientes y determina el contenido en colesterol en la sangre antes y después del tratamiento. Los datos muestrales expresados en miligramos por 100 mililitros son los siguientes:

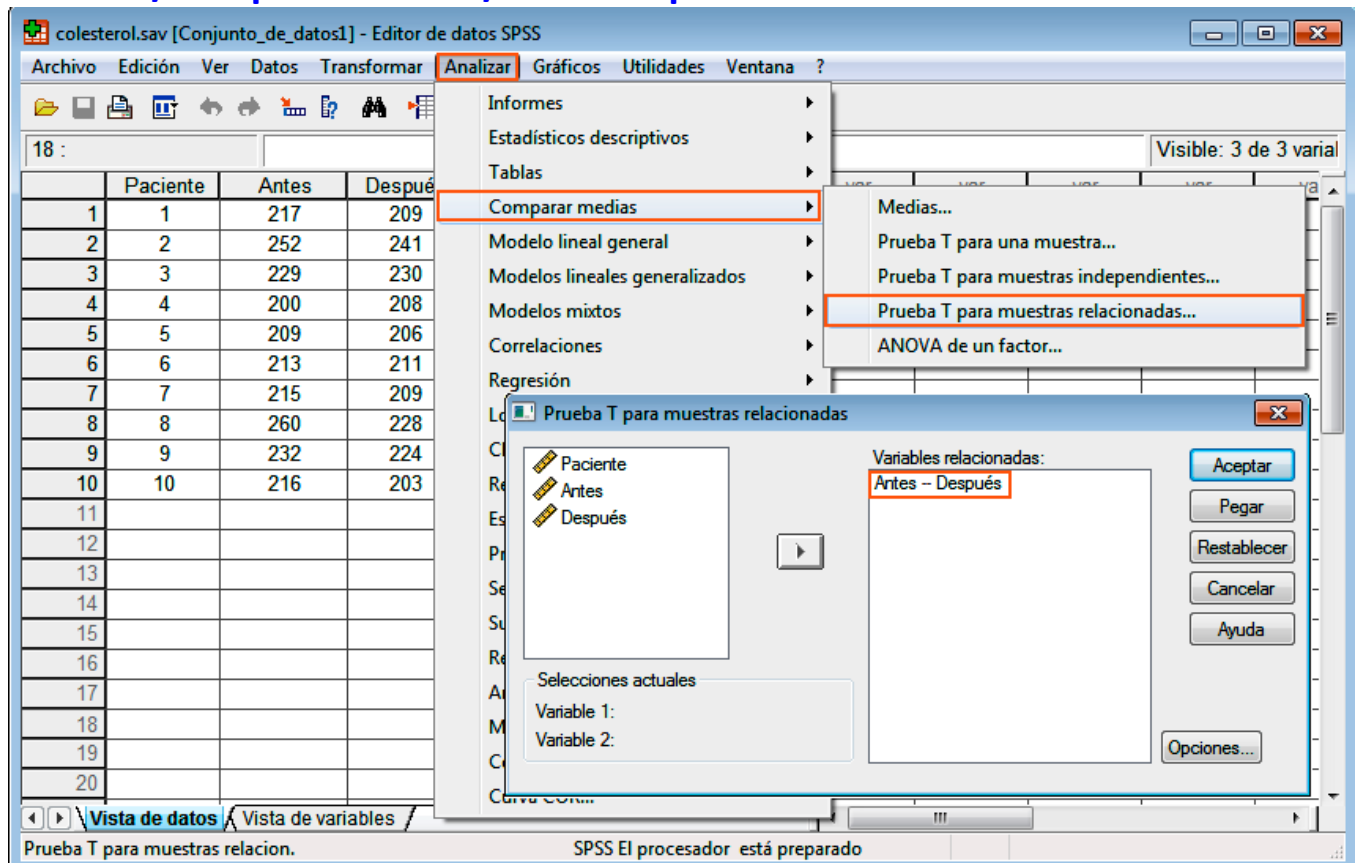
Antes	217	252	229	200	209	213	215	260	232	216
Después	209	241	230	208	206	211	209	228	224	203

¿La nueva droga reduce el colesterol con una confianza del 95%?

Solución:

Se trata de muestras pequeñas apareadas, en donde no existe independencia entre los datos muestrales.

Analizar/Comparar medias/Prueba T para muestras relacionadas



The screenshot shows the SPSS interface with the 'Analizar' menu open. The 'Prueba T para muestras relacionadas...' option is highlighted. The dialog box for 'Prueba T para muestras relacionadas' is open, showing the following details:

- Selecciones actuales:** Paciente, Antes, Después
- Variables relacionadas:** Antes - Después
- Botones:** Aceptar, Pegar, Restablecer, Cancelar, Ayuda, Opciones...

Se analiza el contraste: $H_0: \mu_1 - \mu_2 = 0$ frente a $H_1: \mu_1 - \mu_2 \neq 0$

O bien, $H_0: d = 0$ frente a $H_1: d \neq 0$

$$\text{donde, } \bar{d}_0 = \frac{1}{10} \sum_{i=1}^{10} d_i = \frac{1}{10} \sum_{i=1}^{10} (x_i - y_i) \quad s_{d_0}^2 = \frac{1}{9} \sum_{i=1}^{10} (d_i - \bar{d})^2$$

$$I_{0,95}(d) = \left[\bar{d}_0 \pm t_{0,025,9} \frac{s_{d_0}}{\sqrt{10}} \right]$$

En la salida del proceso:

Estadísticos de muestras relacionadas

		Media	N	Desviación típ.	Error típ. de la media
Par 1	Antes	224,30	10	19,102	6,041
	Después	216,90	10	12,810	4,051

Correlaciones de muestras relacionadas

		N	Correlación	Sig.
Par 1	Antes y Después	10	,852	,002

La muestra antes-después se encuentra muy correlacionada (0,852) y el contraste de correlación poblacional $H_0: \rho = 0$ frente a $H_1: \rho \neq 0$ presenta p – valor (Sig) = 0,02 < 0,05 rechazando la hipótesis nula y en consecuencia existe correlación entre las variables.

Prueba de muestras relacionadas

		Diferencias relacionadas				t	gl	Sig. (bilateral)	
		Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
					Inferior				Superior
Par 1	Antes - Después	7,400	10,585	3,347	-,172	14,972	2,211	9	,054

$$\text{Error típico de la media} \equiv \frac{s_{d_0}}{\sqrt{n}} = \frac{10,585}{\sqrt{10}} = 3,347$$

$$\text{El estadístico de contraste} \equiv \frac{\bar{d}_0}{\frac{s_{d_0}}{\sqrt{n}}} = \frac{7,400}{3,347} = 2,211$$

- ◆ El intervalo de confianza $[-0,172, 14,972]$ cubre el cero, por lo que no se modifica el contenido de colesterol en sangre antes y después del tratamiento.
- ◆ p – valor (Sig bilateral) = $0,054 > 0,05$ por lo que se acepta la hipótesis nula $H_0: \mu_1 - \mu_2 = 0 \cong H_0: d = 0$ de igualdad de medias.

📁 Se quiere comprobar la efectividad de una determinada vacuna contra una alergia. Para ello se suministró la vacuna a cien pacientes y se les comparó con un grupo placebo de cien pacientes afectados también por la alergia en épocas pasadas. Entre los vacunados, ocho sufrieron alergia y entre los no vacunados veinticinco volvieron a sufrir alergia. ¿Se puede concluir que la vacuna es eficaz en disminuir la alergia?. Utilizar un nivel de significación de $0,05$.

Solución:

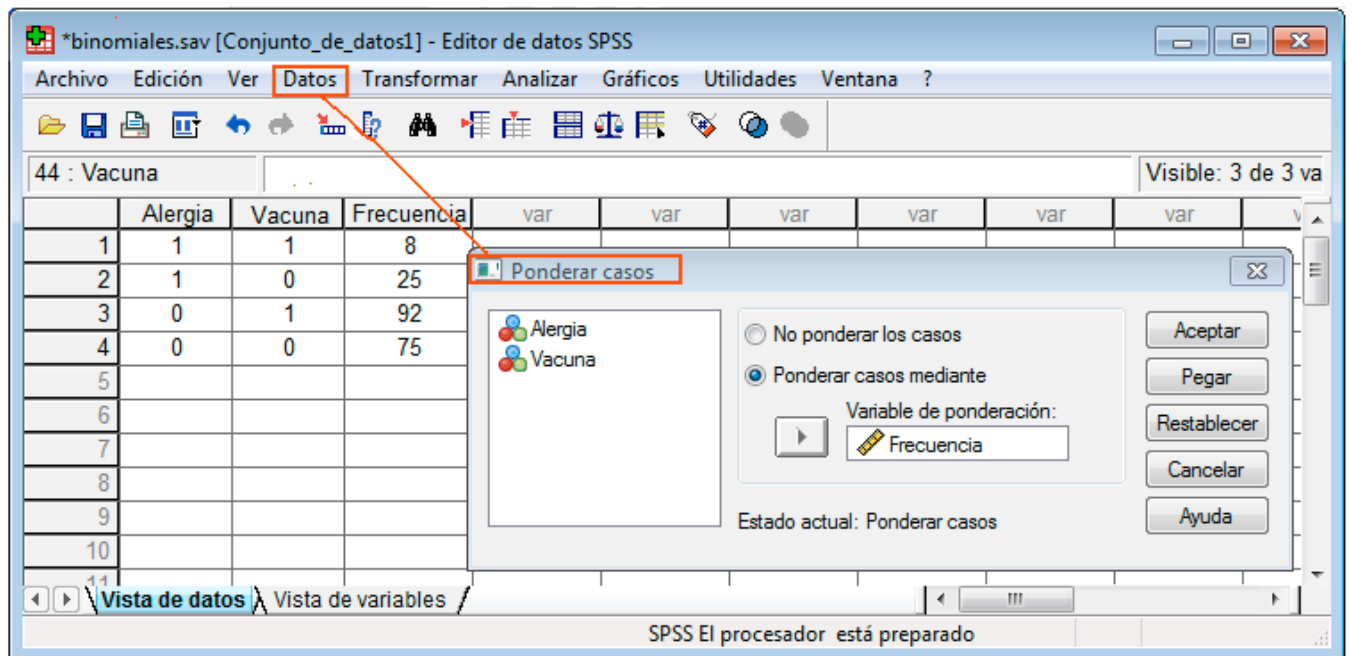
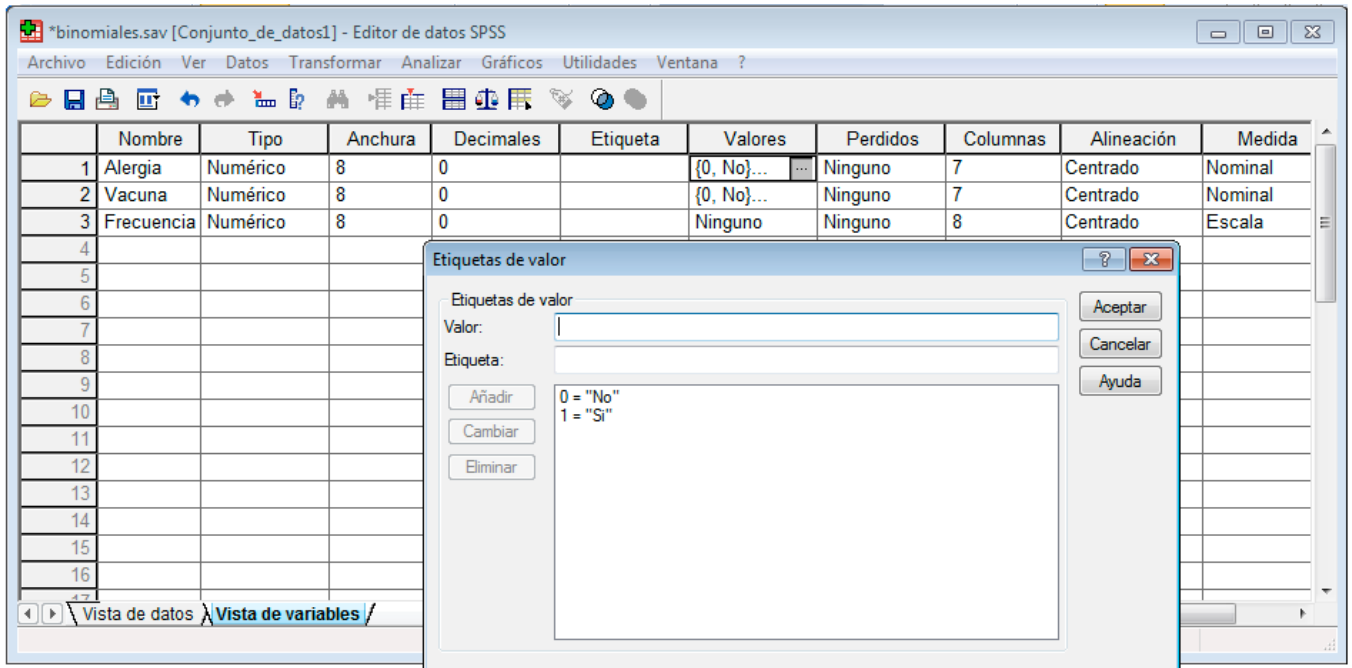
Modelo: Se considera que el número de pacientes vacunados afectados por la alergia es una variable binomial $B(p_1, n_1)$ y el número de pacientes no vacunados es otra binomial $B(p_2, n_2)$, en donde $n_1 = n_2 = 100$ muestras grandes. Para analizar si la vacuna es no eficaz se debe, pues, contrastar $p_2 \leq p_1$. Se establecen las hipótesis:

$H_0: p_2 \leq p_1$ frente a $H_1: p_2 > p_1$

Datos a introducir en SPSS:

	Vacuna	Sí	No
Alergia			
Sí		8	25
No		92	75

donde se tienen que ponderar los casos según la frecuencia



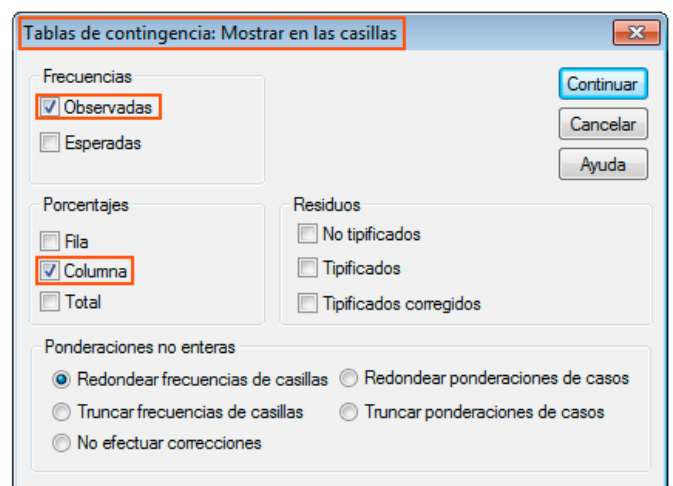
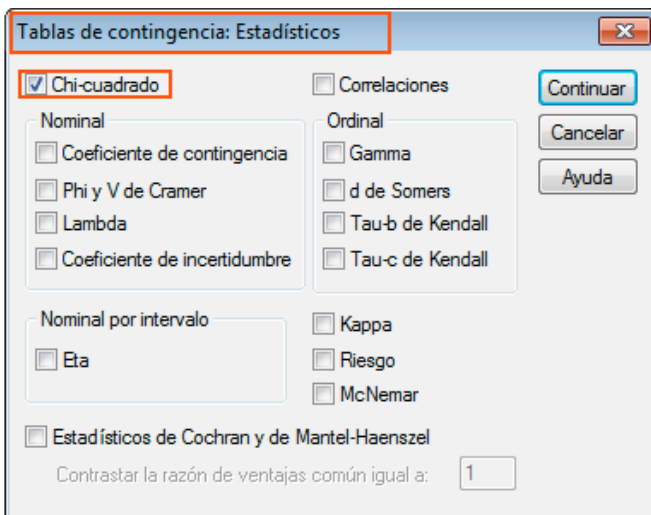
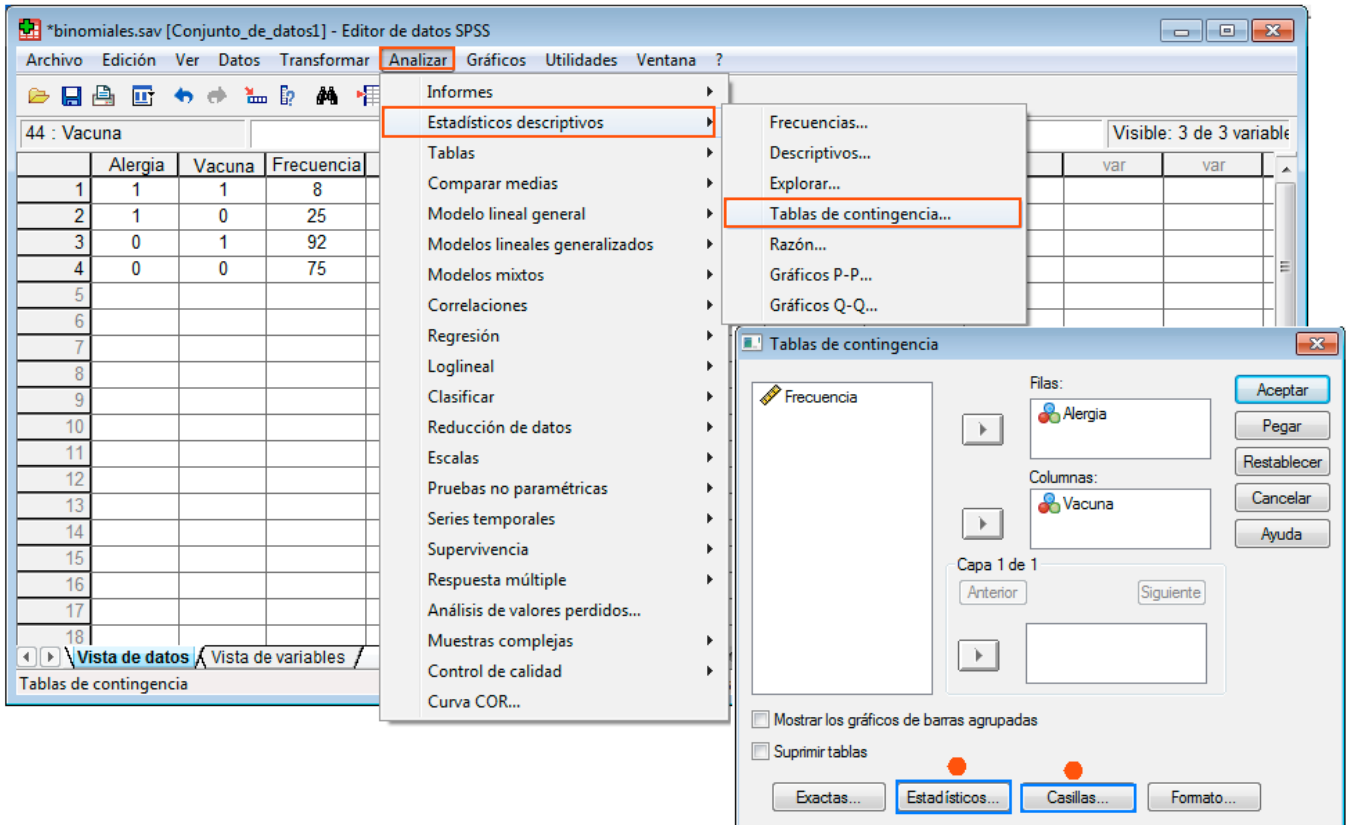


Tabla de contingencia Alergia * Vacuna

		Vacuna		Total	
		No	Si		
Alergia	No	Recuento	75	92	167
		% de Vacuna	75,0%	92,0%	83,5%
	Si	Recuento	25	8	33
		% de Vacuna	25,0%	8,0%	16,5%
Total		Recuento	100	100	200
		% de Vacuna	100,0%	100,0%	100,0%

Cada casilla de la Tabla de contingencia muestra la frecuencia observada y el porcentaje que ésta representa sobre el total de la columna Alergia*Vacuna.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	10,488 ^b	1	,001		
Corrección por continuidad	9,291	1	,002		
Razón de verosimilitudes	10,927	1	,001		
Estadístico exacto de Fisher				,002	,001
Asociación lineal por lineal	10,436	1	,001		
N de casos válidos	200				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 16,50.

La Prueba de Chi-cuadrado muestra los resultados de cinco estadísticos para la comparación de las dos proporciones. En general, para muestras grandes se elige el estadístico Chi-cuadrado con la corrección de continuidad de Yates.

Chi-cuadrado corrección por continuidad es 9,291 con un p-valor asociado de 0,002. Como p-valor (Sig. asintótica bilateral) = 0,002 < 0,05 se rechaza la hipótesis nula, aceptando en consecuencia la hipótesis alternativa $H_1: p_2 > p_1$, pudiendo afirmar que la vacuna es eficaz con un nivel de significación 0,05

📁 En un examen final de estadística teórica los estudiantes recibieron las siguientes calificaciones:

80	70	90	75	55	80	80	65	100	75	60	60
75	95	80	80	90	85	70	95	75	70	85	80
80	65	65	50	75	75	85	85	90	70		

Comprobar si las calificaciones fueron o no distribuidas según una ley normal a un nivel de significación 0,05

Solución:

El método de aplicación de la Prueba de ajuste para la normalidad de la distribución de frecuencias es:

$$\text{Número de intervalos} = \sqrt{34} \approx 6$$

$$\text{Amplitud del intervalo} = \frac{X_{\text{máx}} - X_{\text{mín}}}{n} = \frac{100 - 50}{6} \approx 10$$

Utilizando intervalos de clase convenientes, se clasifican los datos en una distribución de frecuencias:

Intervalos	x_i	n_i	p_i	$e_i = p_i n$	n_i^2	$\frac{n_i^2}{e_i}$
45 - 55	50	1	0,0121	0,41	1	2,44
55 - 65	60	2	0,08	2,72	4	1,47
65 - 75	70	7	0,2366	8,04	49	6,09
75 - 85	80	13	0,34	11,56	169	14,62
85 - 95	90	8	0,2366	8,04	64	7,96
95 - 105	100	3	0,08	2,72	9	3,31
		n = 34				35,87

Las condiciones necesarias para aplicar el test de la Chi-cuadrado exigen que al menos el 80% de los valores esperados de las celdas sean mayores que 5. Cuando esto no ocurre hay que agrupar modalidades contiguas en una sola hasta lograr que la nueva frecuencia sea mayor que cinco.

Se calculan la media y la desviación típica: $\mu = 80$ y $\sigma = 11,4$

Mediante la tabla normal se hallan las probabilidades de cada uno de los intervalos:

- $$P[45 < x < 55] = P\left[\frac{45 - 80}{11,4} < \frac{x - 80}{11,4} < \frac{55 - 80}{11,4}\right] = P[-3,07 < z < -2,19] =$$

$$= P[2,19 < z < 3,07] = P[z > 2,19] - P[z > 3,07] = 0,0143 - 0,00135 = 0,0129$$
- $$P[55 < x < 65] = P[-2,19 < z < -1,32] = P[1,32 < z < 2,19] =$$

$$= P[z > 1,32] - P[z > 2,19] = 0,0934 - 0,0143 = 0,0791$$
- $$P[65 < x < 75] = P[-1,32 < z < -0,44] = P[0,44 < z < 1,32] =$$

$$= P[z > 0,44] - P[z > 1,32] = 0,33 - 0,0934 = 0,2366$$
- $$P[75 < x < 85] = P[-0,44 < z < 0,44] = 1 - 2P[z > 0,44] = 1 - 2 \cdot 0,33 = 0,34$$
- $$P[85 < x < 95] = P[0,44 < z < 1,32] = 0,2366$$
- $$P[95 < x < 105] = P[1,32 < z < 2,19] = 0,08$$

Intervalos	x_i	n_i	p_i	$e_i = p_i n$	n_i^2	$\frac{n_i^2}{e_i}$
45 - 55	50	1	0,0121	0,41	1	2,44
55 - 65	60	2	0,08	2,72	4	1,47
65 - 75	70	7	0,2366	8,04	49	6,09
75 - 85	80	13	0,34	11,56	169	14,62
85 - 105	95	11	0,3157	10,73	121	11,28
		$n = 34$				35,9

- $$P[85 < x < 105] = P[0,44 < z < 2,19] = 0,33 - 0,0143 = 0,3157$$

Se establece la hipótesis nula

H_0 : Las calificaciones se distribuyen según una ley normal

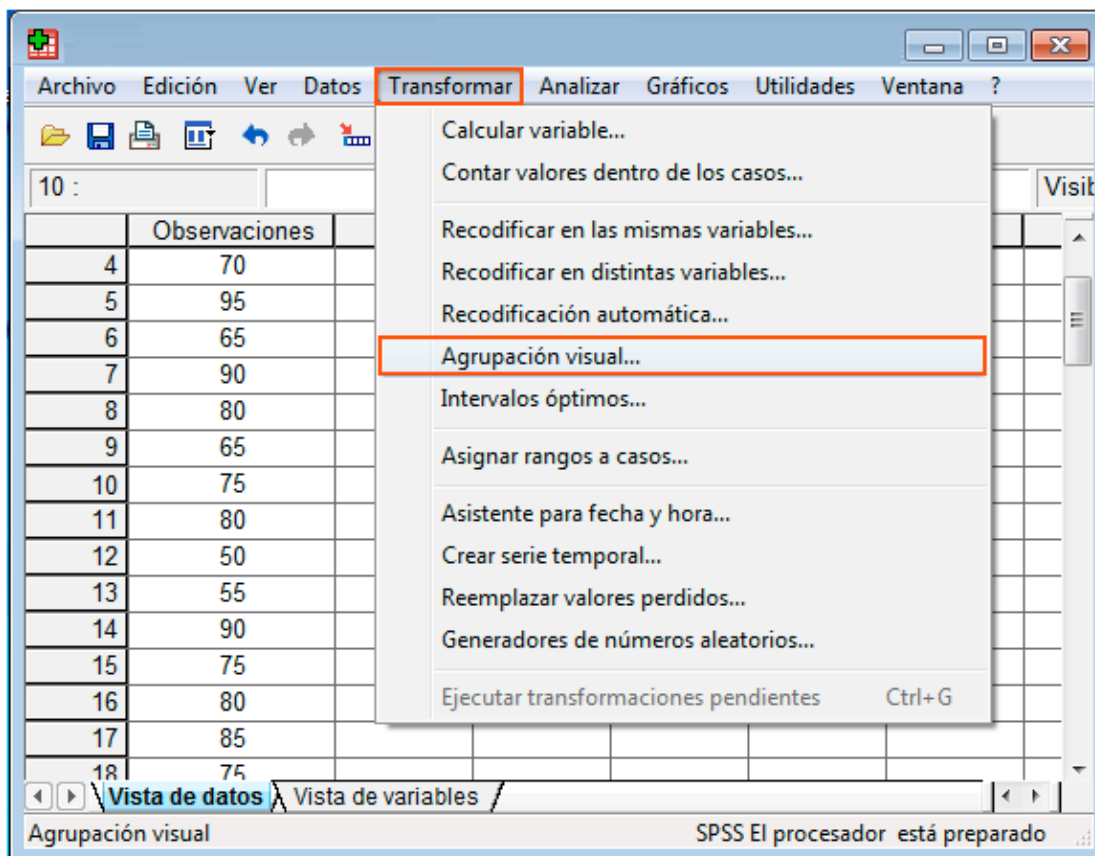
El número de grados de libertad: $k - p = 5 - 3 = 2$, se han perdido tres grados de libertad, ya que se han calculado tres parámetros:

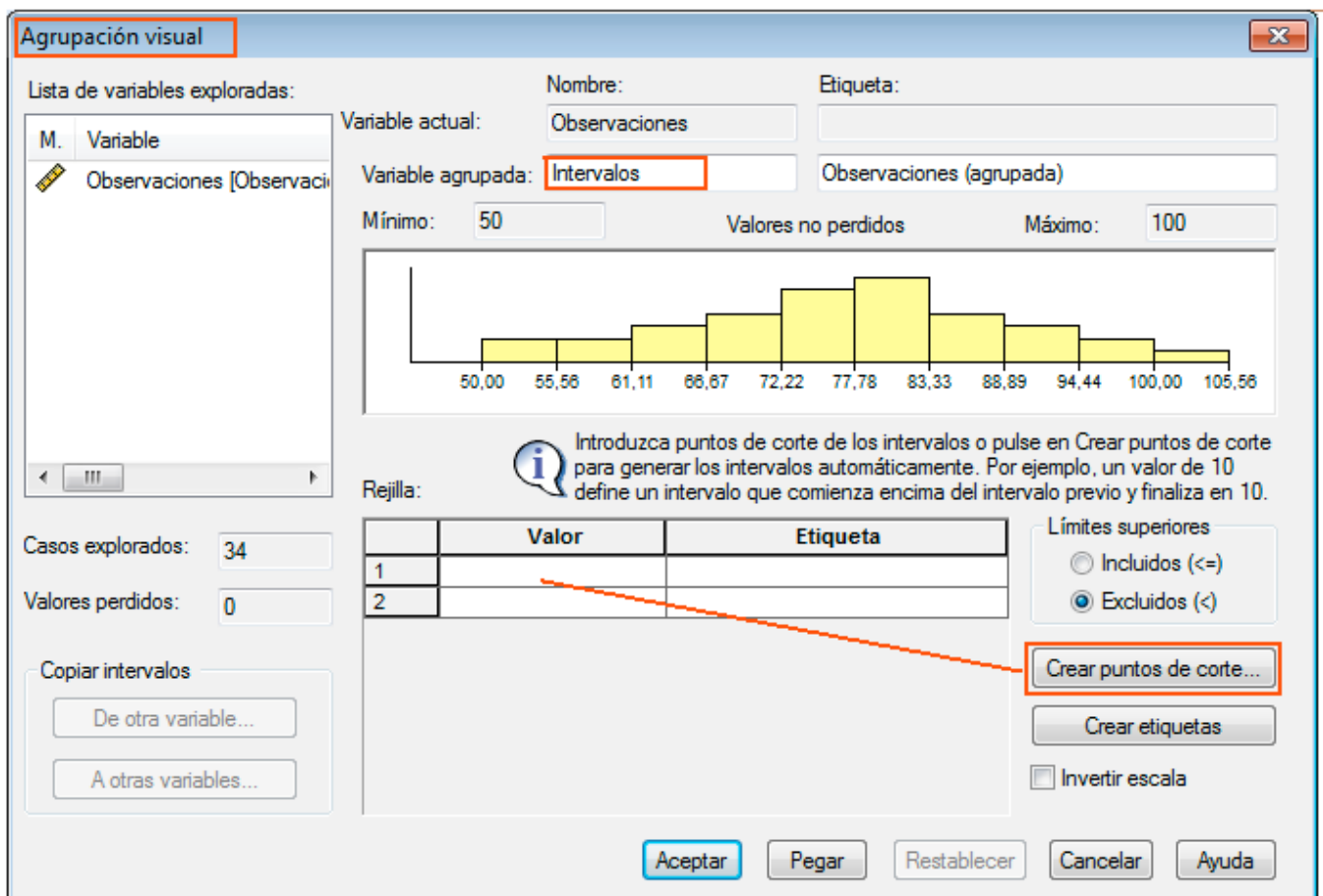
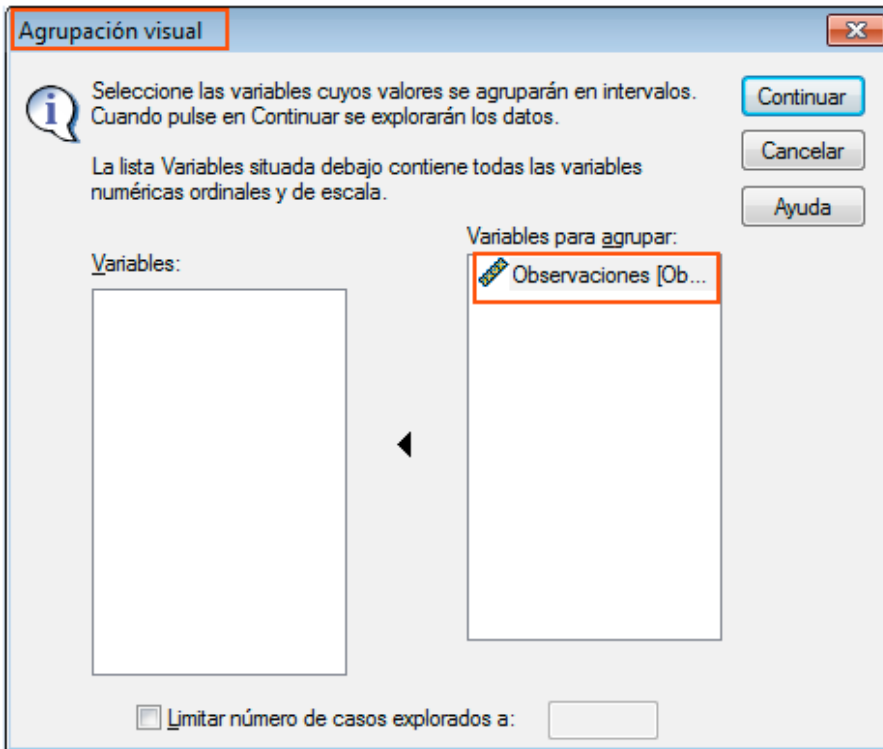
$$\sum_{i=1}^5 n_i = n \quad \mu \quad \gamma \quad \sigma$$

Estadístico de contraste $\chi^2 = \sum_{i=1}^5 \frac{(n_i - e_i)^2}{e_i} = \sum_{i=1}^5 \frac{n_i^2}{e_i} - n = 35,9 - 34 = 1,9$

Como el valor observado $\chi^2 = 1,9 < \chi_{0,05, 2}^2 = 5,991$ se acepta la hipótesis nula afirmando que las calificaciones se pueden considerar que se distribuyen normalmente a un nivel $\alpha = 0,05$

Se introducen los datos en SPSS:





Crear puntos de corte

Intervalos de igual amplitud

Intervalos: rellene al menos dos campos

Posición del primer punto de corte: 55

Número de puntos de corte: 6

Amplitud: 10

Posición del último punto de corte: 105

Percentiles iguales basados en los casos explorados

Intervalos - rellene cualquiera de los dos campos

Número de puntos de corte:

% de casos:

Puntos de corte en media y desviaciones típicas seleccionadas, basadas en casos explorados

+/- 1 Desv. típica

+/- 2 Desv. típicas

+/- 3 Desv. típicas

Aplicar reemplazará las definiciones de los puntos de corte actuales con esta especificación.
Un intervalo final incluirá todos los valores restantes: N puntos de corte generan N+1 intervalos.

Aplicar Cancelar Ayuda

Agrupación visual

Lista de variables exploradas:

M. Variable

Observaciones [Observaciones]

Nombre: Observaciones

Etiqueta: Observaciones (agrupada)

Variable actual: Observaciones

Variable agrupada: Intervalos

Mínimo: 50 Valores no perdidos Máximo: 100

Introduzca puntos de corte de los intervalos o pulse en Crear puntos de corte para generar los intervalos automáticamente. Por ejemplo, un valor de 10 define un intervalo que comienza encima del intervalo previo y finaliza en 10.

Rejilla:

	Valor	Etiqueta
1	55	
2	65	
3	75	
4	85	
5	95	
6	105	
7		
8		

Límites superiores

Incluidos (<=)

Excluidos (<)

Crear puntos de corte...

Crear etiquetas

Invertir escala

Aceptar Pegar Restablecer Cancelar Ayuda

Agrupación visual

Lista de variables exploradas:

M.	Variable
1	Observaciones [Observaciones]

Variable actual: Observaciones

Variable agrupada: Intervalos

Mínimo: 50 Valores no perdidos Máximo: 100

Introduzca puntos de corte de los intervalos o pulse en Crear puntos de corte para generar los intervalos automáticamente. Por ejemplo, un valor de 10 define un intervalo que comienza encima del intervalo previo y finaliza en 10.

Rejilla:

	Valor	Etiqueta
1	55	<55
2	65	55 - 64
3	75	65 - 74
4	85	75 - 84
5	95	85 - 94
6	105	95 - 104
7		
8		

Límites superiores

Incluidos (<=)

Excluidos (<)

Crear puntos de corte...

Crear etiquetas

Invertir escala

Aceptar Pegar Restablecer Cancelar Ayuda

normal.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar **Analizar** Gráficos Utilidades Ventana ?

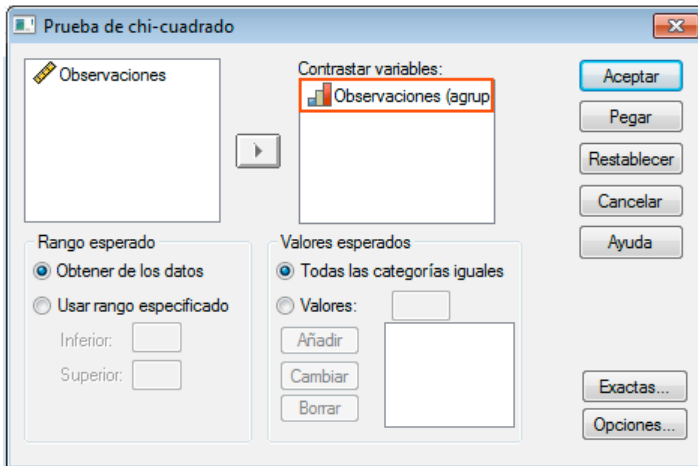
1 : Observaciones 80

	Observaciones	Intervalos
1	80	75 - 84
2	75	75 - 84
3	80	75 - 84
4	70	65 - 74
5	95	95 - 104
6	65	65 - 74
7	90	85 - 94
8	80	75 - 84
9	65	65 - 74
10	75	75 - 84
11	80	75 - 84
12	50	<55
13	55	55 - 64
14	90	85 - 94
15	75	75 - 84
16	80	75 - 84
17	85	85 - 94
18	75	75 - 84
19	80	75 - 84

Analizar

- 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**
 - Chi-cuadrado...**
 - Binomial...
 - Rachas...
 - K-S de 1 muestra...
 - 2 muestras independientes...
 - K muestras independientes...
 - 2 muestras relacionadas...
 - K muestras relacionadas...
- Series temporales
- Supervivencia
- Respuesta múltiple
- Análisis de valores perdidos...
- Muestras complejas
- Control de calidad
- Curva COR...

Chi-cuadrado



Observaciones (agrupada)			
	N observado	N esperado	Residual
<55	1	5,7	-4,7
55 - 64	3	5,7	-2,7
65 - 74	7	5,7	1,3
75 - 84	13	5,7	7,3
85 - 94	7	5,7	1,3
95 - 104	3	5,7	-2,7
Total	34		

El número esperado para cada fila (suma de frecuencias observadas dividida por el número de filas. Hay 34 calificaciones observadas divididas por 6 filas resulta un número esperado de 5,7.

El valor residual muestra el residuo (frecuencia observada menos valor esperado).

$$\chi^2 = \sum_{i=1}^5 \frac{\text{Residual}}{\text{esperado}} = 16,471$$


Estadísticos de contraste

	Observaciones (agrupada)
Chi-cuadrado ^a	16,471
gl	5
Sig. asintót.	,006

a. 0 casillas (,0%) tienen frecuencias esperadas menores que 5. La frecuencia de casilla esperada mínima es 5,7.

p-valor (Sig. asintótica) = 0,006 < 0,05 → Se rechaza H₀

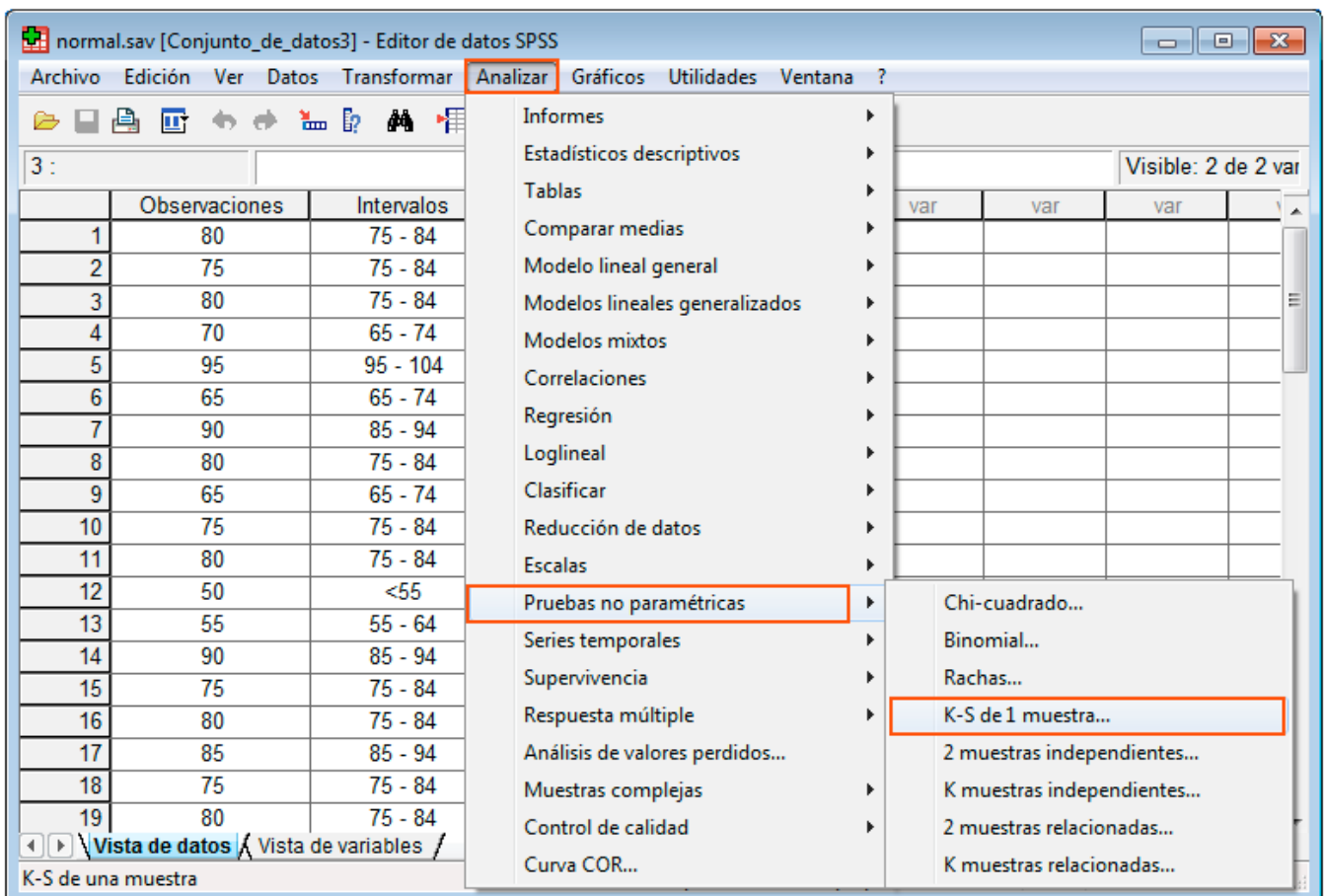
En consecuencia, se rechaza la hipótesis nula, las calificaciones no se pueden considerar que se distribuyen normalmente a un nivel $\alpha = 0,05$

 La prueba de Kolmogorov-Smirnov para una muestra es un proceso de "bondad de ajuste", que permite medir el grado de concordancia existente entre la distribución de un conjunto de datos y una distribución teórica determinada.

Es decir, contrasta si las observaciones pueden proceder de la distribución determinada. La prueba de Kolmogorov-Smirnov para una muestra se puede utilizar para comprobar si una variable se distribuye normalmente.

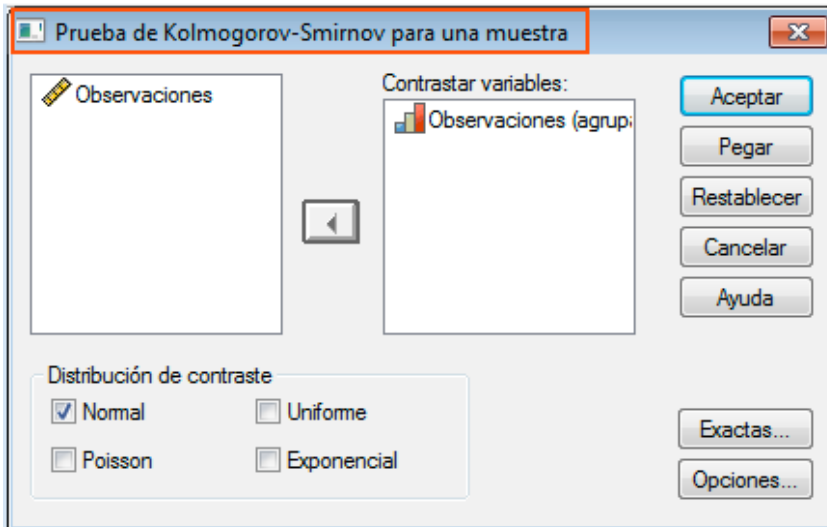
Para verificar la normalidad de una distribución, la prueba de Lilliefors conlleva algunas mejoras respecto a la de Kolmogorov-Smirnov.

En general, alternativas más potentes son el test de Shapiro–Wilk o la prueba de Anderson-Darling.



The screenshot shows the SPSS 'Analyze' menu with the following items: 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, and Curva COR... The 'Pruebas no paramétricas' submenu is open, showing: Chi-cuadrado..., Binomial..., Rachas..., K-S de 1 muestra..., 2 muestras independientes..., K muestras independientes..., 2 muestras relacionadas..., and K muestras relacionadas... The 'K-S de 1 muestra...' option is highlighted.

	Observaciones	Intervalos
1	80	75 - 84
2	75	75 - 84
3	80	75 - 84
4	70	65 - 74
5	95	95 - 104
6	65	65 - 74
7	90	85 - 94
8	80	75 - 84
9	65	65 - 74
10	75	75 - 84
11	80	75 - 84
12	50	<55
13	55	55 - 64
14	90	85 - 94
15	75	75 - 84
16	80	75 - 84
17	85	85 - 94
18	75	75 - 84
19	80	75 - 84



Prueba de Kolmogorov-Smirnov para una muestra

		Observaciones (agrupada)
N		34
Parámetros normales ^{a,b}	Media	3,91
	Desviación típica	1,190
Diferencias más extremas	Absoluta	,206
	Positiva	,176
	Negativa	-,206
Z de Kolmogorov-Smirnov		1,201
Sig. asintót. (bilateral)		,112

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

p-valor (Sig. asintótica bilateral) = 0,112 > 0,05 → Se acepta H_0

Las calificaciones se distribuyen normalmente a un nivel $\alpha = 0,05$

📁 En un laboratorio se observó el número de partículas α que llegan a una determinada zona procedente de una sustancia radiactiva en un corto espacio de tiempo siempre igual, obteniéndose los siguientes resultados:

Número partículas	0	1	2	3	4	5
Número períodos de tiempo	120	200	140	20	10	2

¿Se pueden ajustar los datos obtenidos a una distribución de Poisson, con un nivel de significación del 5%?

Solución:

Hipótesis nula

H_0 : La distribución empírica se ajusta a una distribución de Poisson

La hipótesis nula se acepta, a un nivel de significación α sí

$$\chi_{k-p-1}^2 = \underbrace{\sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i}}_{\text{estadístico contraste}} = \sum_{i=1}^k \frac{n_i^2}{e_i} - n < \underbrace{\chi_{\alpha; k-p-1}^2}_{\text{estadístico teórico}}$$

$k \equiv$ Número intervalos $p \equiv$ Número parámetros a estimar

$$\text{Región de rechazo de la hipótesis nula: } R = \left\{ \sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i} \geq \chi_{\alpha; k-p-1}^2 \right\}$$

La distribución de Poisson se caracteriza porque sólo depende del parámetro λ que coincide con la media.

Sea la variable aleatoria $X =$ Número de partículas y $n_i \equiv$ Número de períodos de tiempo

x_i	n_i	$x_i \cdot n_i$	$P(x_i = k) = p_i$
0	120	0	0,3012
1	200	200	0,3614
2	140	280	0,2169
3	20	60	0,0867
4	10	40	0,0260
5	2	10	0,0062
n = 492		590	

$$\bar{x} = \lambda = \frac{\sum x_i n_i}{n} = \frac{590}{492} = 1,2$$

$$P(x_i = k) = \frac{1,2^k}{k!} e^{-1,2} \quad k = 0, \dots, 5$$

Las probabilidades con que llegan las partículas $k = 0, \dots, 5$ se obtienen

sustituyendo los valores de k en $P(x_i = k) = \frac{1,2^k}{k!} e^{-1,2}$ o en las tablas con

$$\lambda = 1,2$$

Para verificar si el ajuste de los datos a una distribución de Poisson se acepta o no, mediante una χ^2 , hay que calcular las frecuencias esperadas ($e_i = n \cdot p_i$)

x_i	0	1	2	3	4	5
Fr	120 $e_1 = 148,2$	200 $e_2 = 177,8$	140 $e_3 = 106,7$	20 $e_4 = 42,7$	10 $e_5 = 12,8$	2 $e_6 = 3,05$

$$e_1 = 492 \cdot 0,3012 = 148,2 \quad e_2 = 492 \cdot 0,3614 = 177,8 \quad e_3 = 492 \cdot 0,2169 = 106,7$$

$$e_4 = 492 \cdot 0,0867 = 42,7 \quad e_5 = 492 \cdot 0,0260 = 12,8 \quad e_6 = 492 \cdot 0,0062 = 3,05$$

Dando lugar a una tabla de contingencia 1 x 6, en donde hay que agrupar las dos últimas columnas por tener la última columna frecuencias esperadas menores que cinco.

Se tiene la tabla de contingencia 1 x 5:

x_i	0	1	2	3	4 y 5
Frecuencias	120 $e_1 = 148,2$	200 $e_2 = 177,8$	140 $e_3 = 106,7$	20 $e_4 = 42,7$	12 $e_5 = 15,8$

Así, los grados de libertad son tres: $k - p - 1 = 5 - 1 - 1 = 3$

◆ Estadístico de contraste:

$$\chi^2_3 = \sum_{i=1}^5 \frac{(n_i - e_i)^2}{e_i} = \sum_{i=1}^5 \frac{n_i^2}{e_i} - n =$$

$$= \frac{120^2}{148,2} + \frac{200^2}{177,8} + \frac{140^2}{106,27} + \frac{20^2}{42,7} + \frac{12^2}{15,8} - 492 = 32,31$$

◆ Estadístico teórico: $\chi^2_{0,05; 3} = 7,815$

El estadístico de contraste (bondad de ajuste) es mayor que el estadístico teórico, rechazándose la hipótesis nula.

Es decir, la distribución NO se puede ajustar a una distribución de Poisson a un nivel de significación del 5%.

Se verifica la región de rechazo:

$$R = \left\{ \sum_{i=1}^5 \frac{(n_i - e_i)^2}{e_i} \geq \chi^2_{0,05; 3} \right\} \equiv \{ 32,31 > 7,815 \}$$

The screenshot shows the SPSS 'Medias' dialog box. The 'Dependientes:' field contains 'Numero_particulas' and the 'Independientes:' field contains 'Periodos_tiempo'. The 'Aceptar' button is highlighted.

Numero_particulas

Periodos tiempo	Media	N	Desv. típ.
2	5,00	2	,000
10	4,00	10	,000
20	3,00	20	,000
120	,00	120	,000
140	2,00	140	,000
200	1,00	200	,000
Total	1,20	492	,949

*poisson2.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos **Transformar** Analizar Gráficos Utilidades Ventana ?

Calcular variable

Variable de destino: Probabilidad_Poisson = Expresión numérica: PDF.POISSON(Numero_particulas,1.2)

Tipo y etiqueta...

Numero_particulas
Periodos_tiempo
Probabilidad_Poisson

Grupo de funciones:
FDP y FDP no centrada
Fecha/hora actual
GL inversos
Números aleatorios
Otras
Significación
Todas
Valores perdidos

Funciones y variables especiales:
Pdf.Halfnm
Pdf.Hyper
Pdf.Igauss
Pdf.Laplace
Pdf.Lnormal
Pdf.Logistic
Pdf.Negbin
Pdf.Normal
Pdf.Pareto
Pdf.Poisson
Pdf.T
Pdf.Uniform

PDF.POISSON(cart,media) Numérico.
Devuelve la probabilidad de que un valor de la distribución de Poisson, con el parámetro de media o tasa especificado, sea igual a cart.

Si... (condición de selección de casos opcional)

Aceptar Pegar Restablecer Cancelar Ayuda

	Numero_particulas	Periodos_tiempo
1	0	120
2	1	200
3	2	140
4	3	20
5	4	10
6	5	2
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		

Vista de datos / Vista de variables

*poisson2.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

	Numero_particulas	Periodos_tiempo	Probabilidad Poisson	var
1	0	120	,3012	
2	1	200	,3614	
3	2	140	,2169	
4	3	20	,0867	
5	4	10	,0260	
6	5	2	,0062	
7				

Vista de datos / Vista de variables

SPSS El procesador está preparado

*poisson2.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver **Datos** Transformar Analizar Gráficos Utilidades Ventana ?

Visible: 3 de 3 variables

Ponderar casos

Numero_particulas
Probabilidad_Poisson

No ponderar los casos
 Ponderar casos mediante

Variable de ponderación: Periodos_tiempo

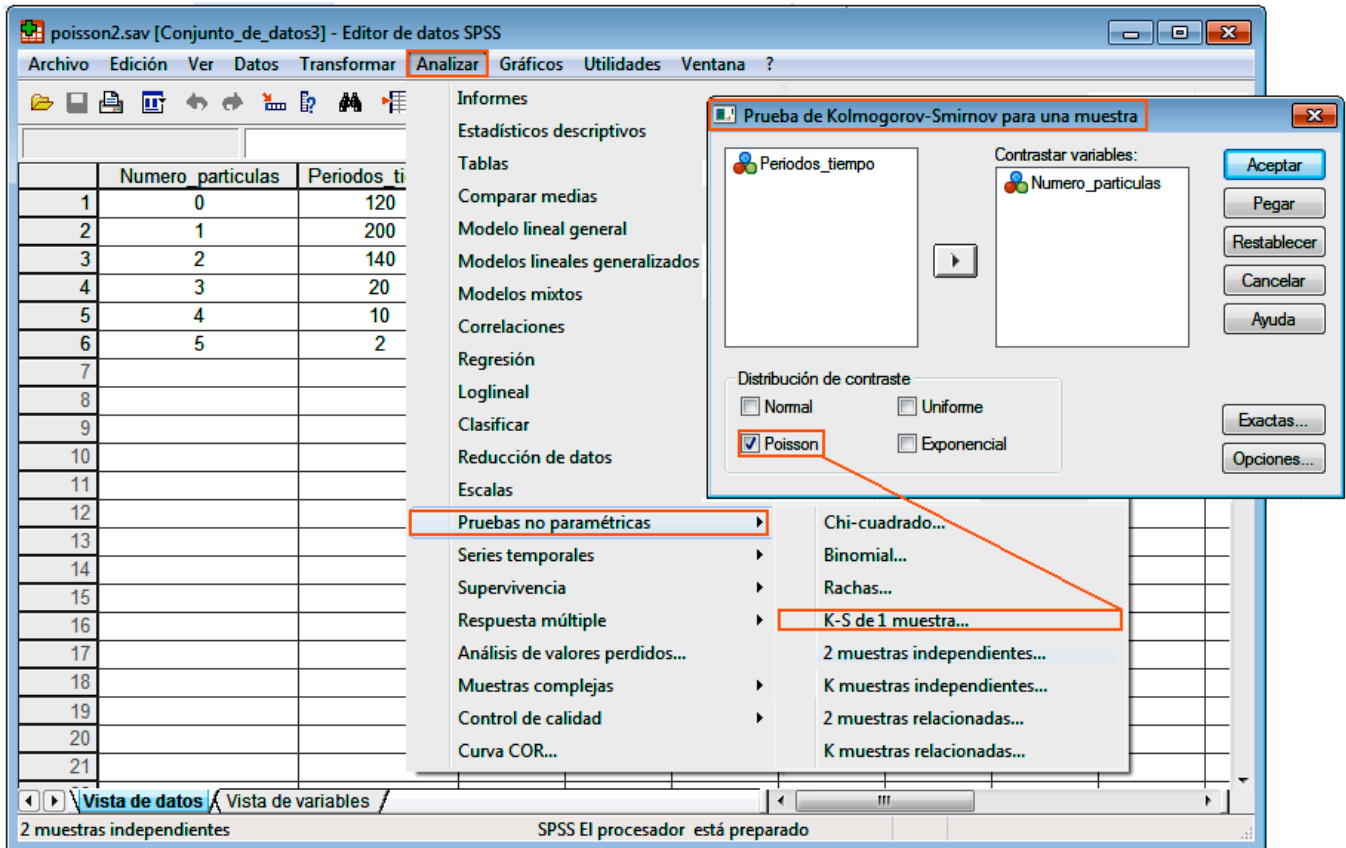
Estado actual: Ponderar casos

Aceptar Pegar Restablecer Cancelar Ayuda

	Numero_particulas	Periodos tiempo	Probabilidad Poisson	var	var	var	var
1	0	120					
2	1	200					
3	2	140					
4	3	20					
5	4	10					
6	5	2					
7							
8							
9							
10							
11							

Vista de datos / Vista de variables

Área de información SPSS El procesador está preparado



Prueba de Kolmogorov-Smirnov para una muestra

		Numero_particulas
N		492
Parámetro de Poisson ^{a,b}	Media	1,20
Diferencias más extremas	Absoluta	,058
	Positiva	,055
	Negativa	-,058
Z de Kolmogorov-Smirnov		1,276
Sig. asintót. (bilateral)		,077

- a. La distribución de contraste es la de Poisson.
 b. Se han calculado a partir de los datos.

El p-valor (Signatura asintótica bilateral) es 0,077 mayor que 0,05, indicando que no debe rechazarse la hipótesis nula, de modo que se admite que la distribución del número de partículas se ajusta a una distribución de Poisson.

📄 La tabla refleja el número de accidentes mortales de tráfico que se producen en una carretera a lo largo de un período de tiempo.

Accidentes mortales por día	0	1	2	3	4	5
Número de días	132	195	120	60	24	9

¿Se ajustan los datos a una distribución de Poisson?. Utilizar un nivel de significación 0,05

Solución:

Hipótesis nula H_0 : La distribución empírica se ajusta a la Poisson

La hipótesis nula se acepta, a un nivel de significación α si

$$\chi_{k-p-1}^2 = \underbrace{\sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i}}_{\text{estadístico contraste}} = \sum_{i=1}^k \frac{n_i^2}{e_i} - n < \underbrace{\chi_{\alpha; k-p-1}^2}_{\text{estadístico teórico}}$$

$k \equiv$ Número intervalos $p \equiv$ Número parámetros a estimar

La distribución de Poisson se caracteriza porque sólo depende del parámetro λ que coincide con la media.

Sea la variable aleatoria $X =$ Número de accidentes mortales por día
y $n_i \equiv$ Número de días

x_i	n_i	$x_i n_i$	$P(x_i = k) = p_i$
0	132	0	0,2466
1	195	195	0,3452
2	120	240	0,2417
3	60	180	0,1128
4	24	96	0,0395
5	9	45	0,0111
$n = 540$		756	

$$\bar{x} = \lambda = \frac{\sum x_i n_i}{n} = \frac{756}{540} = 1,4$$

$$P(x_i = k) = \frac{1,4^k}{k!} e^{-1,4} \quad k = 0, \dots, 5$$

Las probabilidades con que llegan las partículas $k = 0, \dots, 5$ se obtienen

sustituyendo los valores de k en $P(x_i = k) = \frac{1,4^k}{k!} e^{-1,4}$ o en las tablas con

$\lambda = 1,4$

Para verificar si el ajuste de los datos a una distribución de Poisson se acepta o no, mediante una χ^2 , hay que calcular las frecuencias esperadas ($e_i = n \cdot p_i$)

x_i	0	1	2	3	4	5
Fr	132 133,16	195 186,43	120 130,50	60 60,90	24 21,31	9 5,97

$$e_1 = 540 \cdot 0,2466 = 133,16 \quad e_2 = 540 \cdot 0,3452 = 186,43 \quad e_3 = 540 \cdot 0,2417 = 130,5$$

$$e_4 = 540 \cdot 0,1128 = 60,90 \quad e_5 = 540 \cdot 0,0395 = 21,31 \quad e_6 = 540 \cdot 0,0111 = 5,97$$

Dando lugar a una tabla de contingencia 1 x 6, no teniendo que agrupar columnas contiguas al no aparecer frecuencias esperadas menor que cinco.

Los grados de libertad son cuatro: $k - p - 1 = 6 - 1 - 1 = 4$

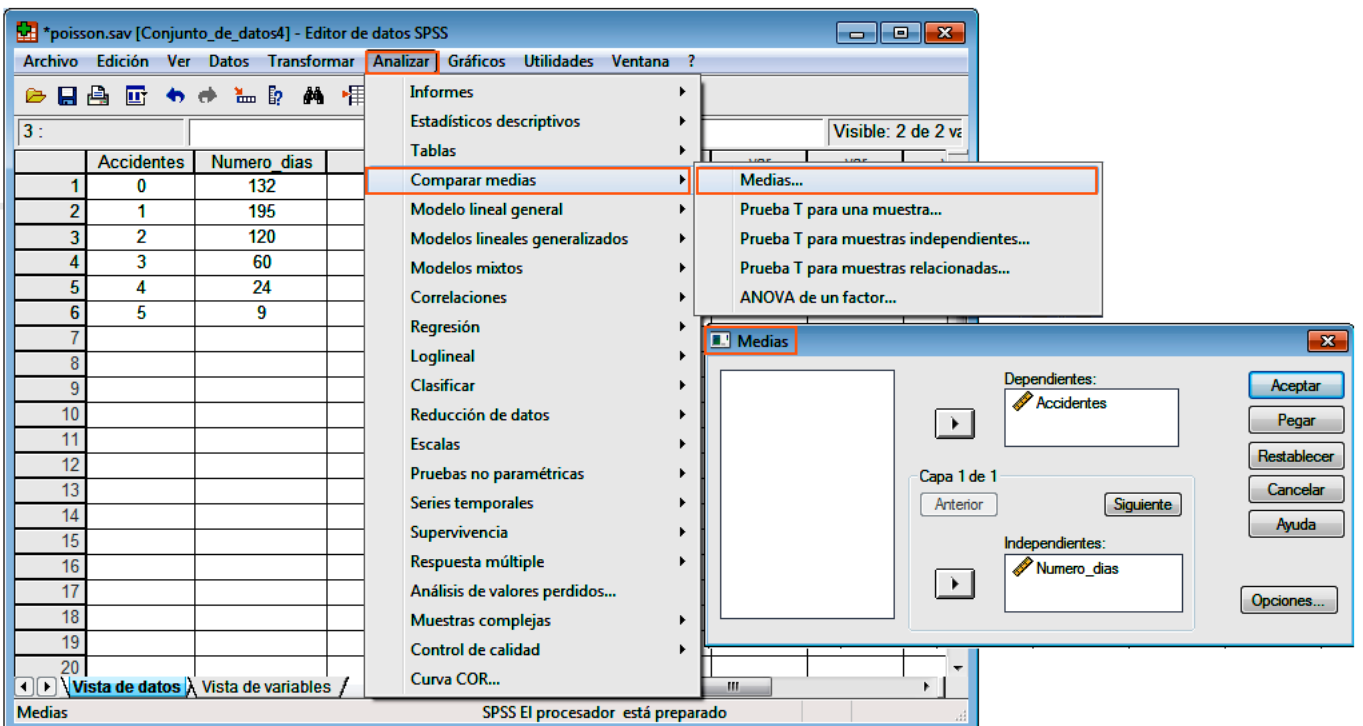
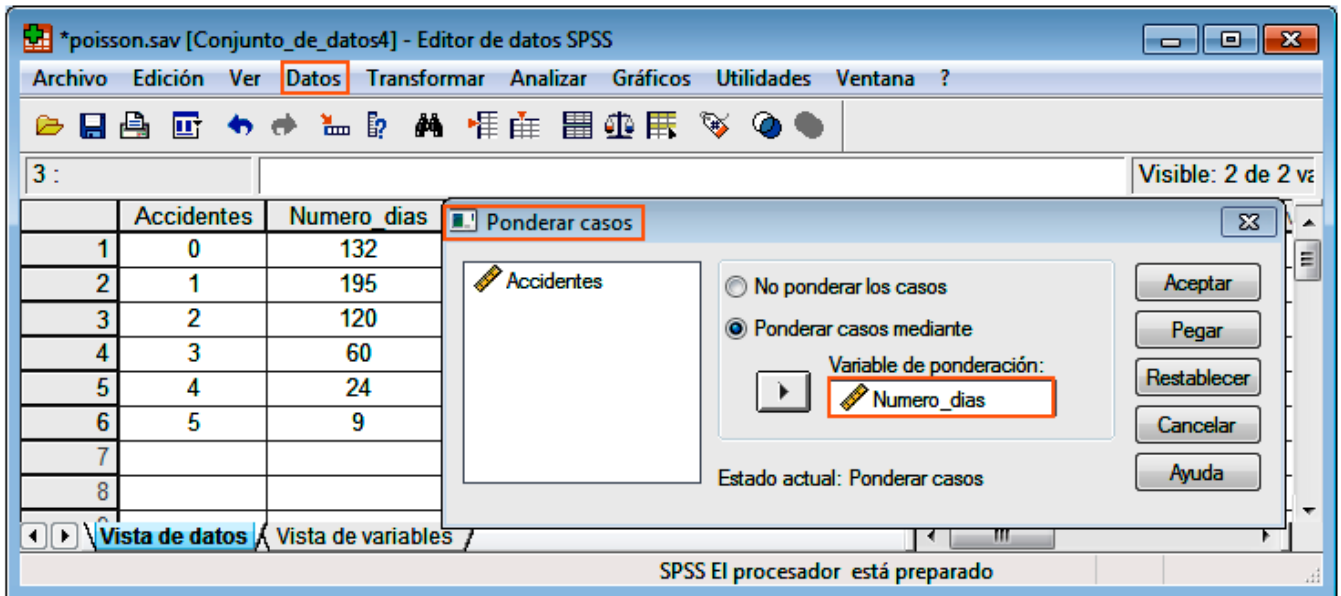
◆ El estadístico de contraste:

$$\chi_3^2 = \sum_{i=1}^6 \frac{(n_i - e_i)^2}{e_i} = \sum_{i=1}^6 \frac{n_i^2}{e_i} - n =$$

$$= \frac{132^2}{133,16} + \frac{195^2}{186,43} + \frac{120^2}{130,5} + \frac{60^2}{60,9} + \frac{24^2}{21,31} + \frac{9^2}{5,97} - 540 = 4,87$$

◆ El estadístico teórico: $\chi_{0,05; 4}^2 = 9,488$

El estadístico de contraste (bondad de ajuste) es menor que el estadístico teórico (9,488), por lo que se acepta la hipótesis nula, es decir, con un nivel de significación 0,05, los accidentes mortales de tráfico en la carretera se ajustan a una distribución de Poisson.



Accidentes

Numero_dias	Media	N	Desv. típ.
9	5,00	9	,000
24	4,00	24	,000
60	3,00	60	,000
120	2,00	120	,000
132	,00	132	,000
195	1,00	195	,000
Total	1,40	540	1,192

*poisson.sav [Conjunto_de_datos4] - Editor de datos SPSS

Archivo Edición Ver Datos **Transformar** Analizar Gráficos Utilidades Ventana ?

Calcular variable

Variable de destino: Probabilidad_Poisson = Expresión numérica: PDF.POISSON(Accidentes,1.4)

Tipo y etiqueta...

Accidentes
Numero_dias
Probabilidad_Poisson

Grupo de funciones:
FDP y FDP no centrada
Fecha/hora actual
GL inversos
Números aleatorios
Otras
Significación
Todas
Valores perdidos

Funciones y variables especiales:
Pdf.Hyper
Pdf.lgauss
Pdf.Laplace
Pdf.Lnormal
Pdf.Logistic
Pdf.Negbin
Pdf.Normal
Pdf.Pareto
Pdf.Poisson
Pdf.T
Pdf.Uniform
Pdf.Weibull

PDF.POISSON(cant,media) Numérico.
Devuelve la probabilidad de que un valor de la distribución de Poisson, con el parámetro de media o tasa especificado, sea igual a cant.

Si... (condición de selección de casos opcional)

Aceptar Pegar Restablecer Cancelar Ayuda

	Accidentes	Numero_dias
1	0	132
2	1	195
3	2	120
4	3	60
5	4	24
6	5	9
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		

Vista de datos Vista de variables

SPSS El procesador está preparado

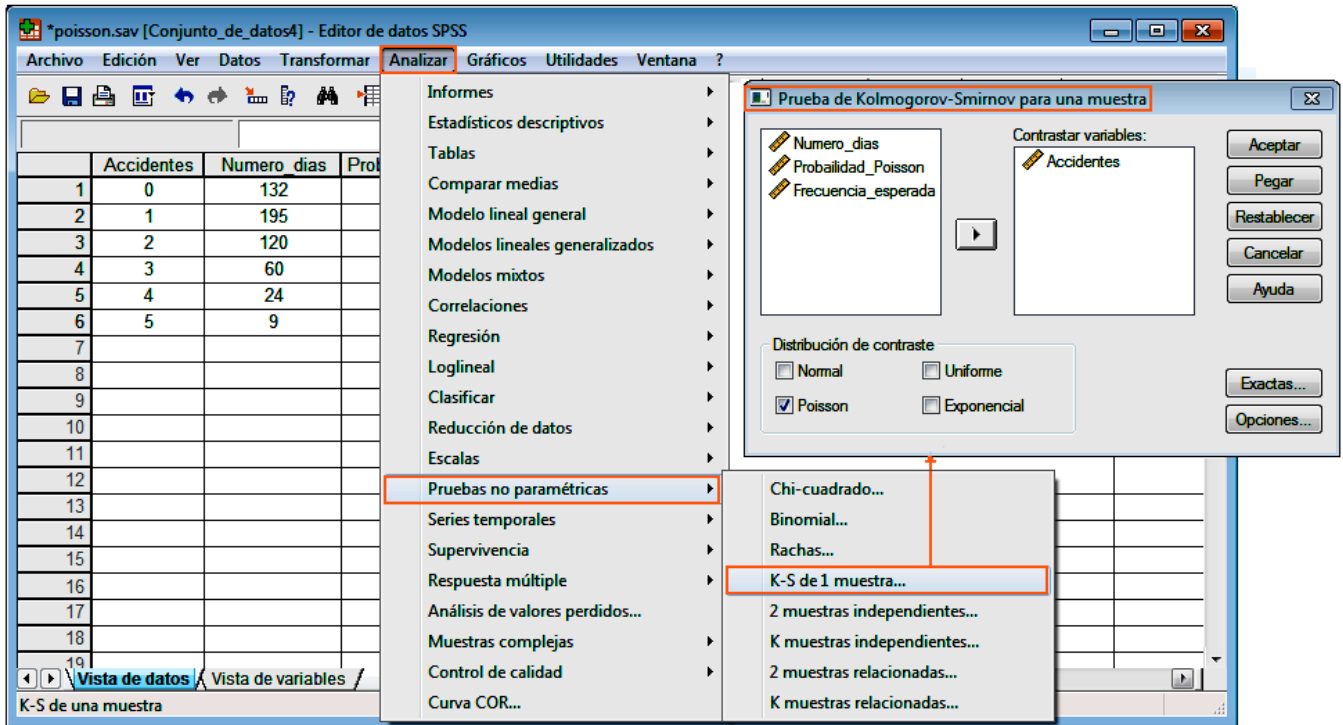
*poisson.sav [Conjunto_de_datos4] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

	Accidentes	Numero_dias	Probailidad Poisson	Frecuencia esperada
1	0	132	,2466	133,16
2	1	195	,3452	186,43
3	2	120	,2417	130,50
4	3	60	,1128	60,90
5	4	24	,0395	21,31
6	5	9	,0111	5,97
7				

Vista de datos Vista de variables

SPSS El procesador está preparado




Prueba de Kolmogorov-Smirnov para una muestra

		Accidentes
N		540
Parámetro de Poisson ^{a,b}	Media	1,40
Diferencias más extremas	Absoluta	,014
	Positiva	,014
	Negativa	-,007
Z de Kolmogorov-Smirnov		,319
Sig. asintót. (bilateral)		1,000

a. La distribución de contraste es la de Poisson.

b. Se han calculado a partir de los datos.

El p-valor (Signatura asintótica bilateral) es 1 mayor que 0,05, indicando que no debe rechazarse la hipótesis nula, de modo que se admite que la distribución del número de accidentes mortales al día se ajusta a una distribución de Poisson.

 Para conocer la opinión de los ciudadanos sobre la actuación del alcalde de una determinada ciudad, se realiza una encuesta a 404 personas, cuyos resultados se recogen en la siguiente tabla:

	Desacuerdo	De acuerdo	No contestan
Mujeres	84	78	37
Varones	118	62	25

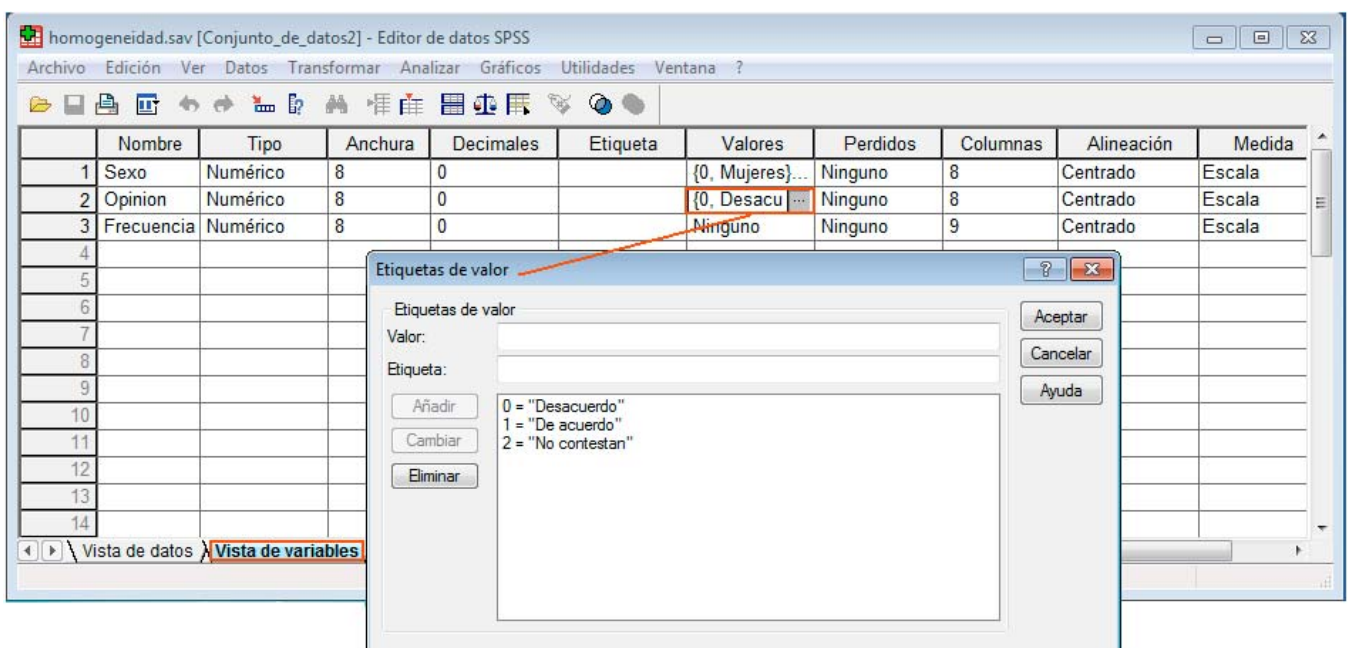
Contrastar, con un nivel de significación del 5%, que no existen diferencias de opinión entre hombres y mujeres ante la actuación del alcalde.

Solución:

Se trata de un contraste de homogeneidad en el que se desea comprobar si las muestras proceden de poblaciones distintas.

Hipótesis nula:

H_0 : No existe diferencia entre hombres y mujeres respecto a la opinión.



The screenshot shows the SPSS 'Editor de datos' window with the following variable list:

Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida
1 Sexo	Númérico	8	0		{0, Mujeres}...	Ninguno	8	Centrado	Escala
2 Opinion	Númérico	8	0		{0, Desacu ...	Ninguno	8	Centrado	Escala
3 Frecuencia	Númérico	8	0		Ninguno	Ninguno	9	Centrado	Escala

The 'Etiquetas de valor' dialog box is open for the 'Opinion' variable, showing the following labels:

- 0 = "Desacuerdo"
- 1 = "De acuerdo"
- 2 = "No contestan"

homogeneidad.sav [Conjunto_de_datos2] - Editor de datos SPSS

Archivo Edición Ver **Datos** Transformar Analizar Gráficos Utilidades Ventana ?

Visible: 3 de 3 var

	Sexo	Opinion	Frecuencia	var	var	var	var	var	var
1	0	0	84						
2	0	1	78						
3	0	2	37						
4	1	0	118						
5	1	1	62						
6	1	2	25						
7									
8									
9									
10									
11									

Ponderar casos

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

homogeneidad.sav [Conjunto_de_datos2] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar **Analizar** Gráficos Utilidades Ventana ?

7 :

	Opinion	Frecuencia	var
1	0	84	
2	1	78	
3	2	37	
4	0	118	
5	1	62	
6	2	25	
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			

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...
 Razón...
 Gráficos P-P...
 Gráficos Q-Q...

Vista de datos / Vista de variables /

Tablas de contingencia

SPSS El procesador está preparado

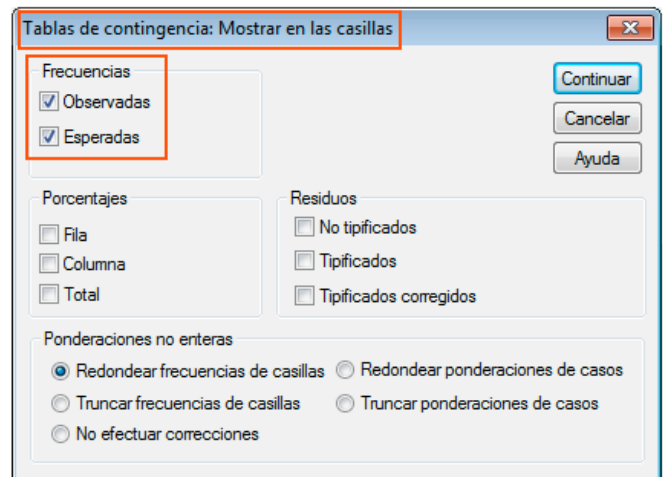
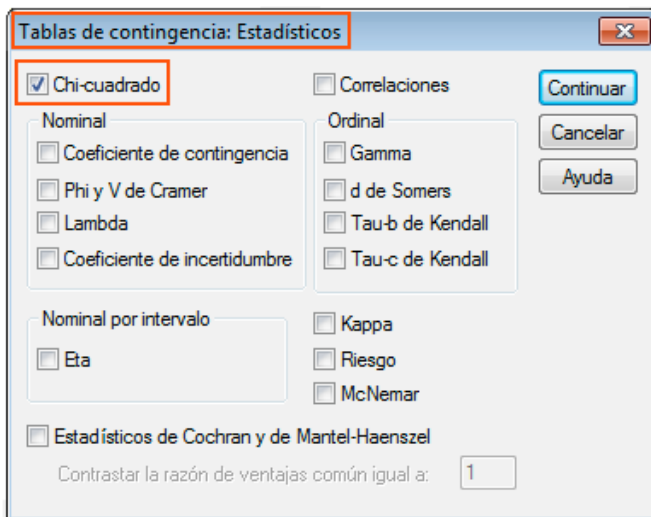
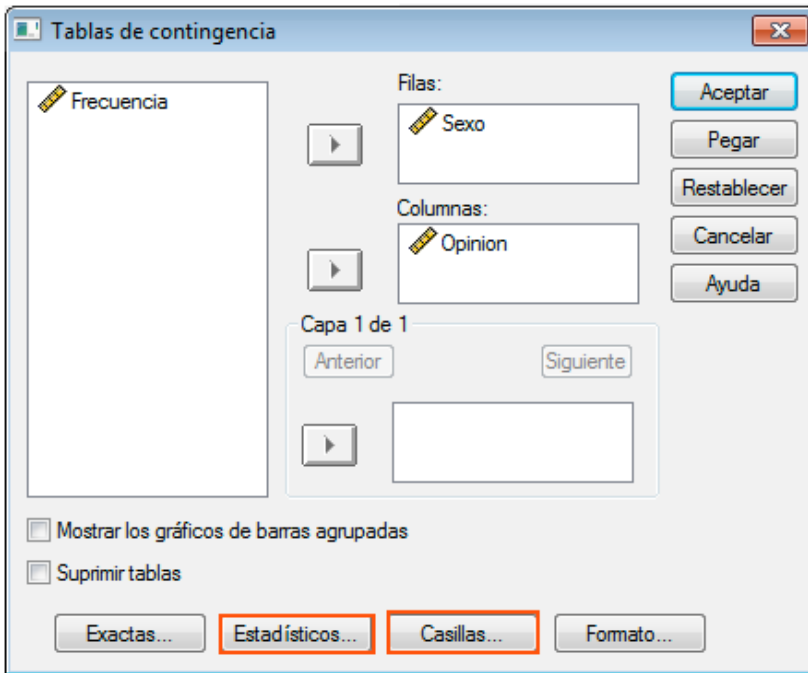


Tabla de contingencia Sexo * Opinion

			Opinion			Total
			Desacuerdo	De acuerdo	No contestan	
Sexo	Mujeres	Recuento	84	78	37	199
		Frecuencia esperada	99,50	68,96	30,54	199
	Hombres	Recuento	118	62	25	205
		Frecuencia esperada	102,50	71	31,46	205
Total		Recuento	202	140	62	404
		Frecuencia esperada	202	140	62	404

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	9,787 ^a	2	,007
Razón de verosimilitudes	9,831	2	,007
Asociación lineal por lineal	8,932	1	,003
N de casos válidos	404		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5.

p-valor (Sig. asintótica bilateral) = 0,007 < 0,05 → Se rechaza H₀

En consecuencia, existe diferencia entre hombres y mujeres sobre la opinión del alcalde con un nivel de confianza del 95%

📄 En un estudio sobre la opinión de fumar en lugares públicos se realiza una encuesta a 350 personas, obteniendo los siguientes resultados:

	Opinión				$n_{i\cdot}$
	Muy en contra	En contra	A Favor	Muy a favor	
Fumador	60	50	20	10	140
No Fumador	10	30	70	100	210
$n_{\cdot j}$	70	80	90	110	350

Con un nivel de significación de 0,05 se desea conocer si existe diferencia de opinión entre fumadores y no fumadores.

Solución:

Se establecen las hipótesis:

H_0 : La opinión es independiente de su condición de fumador o no fumador

H_1 : La opinión no es independiente de su condición de fumador o no fumador

Se acepta H_0 sí: $\chi_c^2 = \overbrace{\sum_{i=1}^2 \sum_{j=1}^4 \frac{(n_{ij} - e_{ij})^2}{e_{ij}}}$ estadístico observado $< \overbrace{\chi_{\alpha, (2-1) \cdot (4-1)}^2}$ estadístico teórico $= \chi_{0,05,3}^2$

	Opinión				$n_{i\cdot}$
	Muy en contra	En contra	A Favor	Muy a favor	
Fumador	60 $e_{11} = 28$	50 $e_{12} = 32$	20 $e_{13} = 36$	10 $e_{14} = 44$	140 140
No Fumador	10 $e_{21} = 42$	30 $e_{22} = 48$	70 $e_{23} = 54$	100 $e_{24} = 66$	210 210
$n_{\cdot j}$	70	80	90	110	350

$$e_{11} = \frac{140 \cdot 70}{350} = 28 \quad e_{12} = \frac{140 \cdot 80}{350} = 32 \quad e_{13} = \frac{140 \cdot 90}{350} = 36 \quad e_{14} = \frac{140 \cdot 110}{350} = 44$$

$$e_{21} = \frac{210 \cdot 70}{350} = 42 \quad e_{22} = \frac{210 \cdot 80}{350} = 48 \quad e_{23} = \frac{210 \cdot 90}{350} = 54 \quad e_{24} = \frac{210 \cdot 110}{350} = 66$$

$$\chi_c^2 = \sum_{i=1}^2 \sum_{j=1}^4 \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^2 \sum_{j=1}^4 \frac{n_{ij}^2}{e_{ij}} - n =$$

$$= \frac{60^2}{28} + \frac{50^2}{32} + \frac{20^2}{36} + \frac{10^2}{44} + \frac{10^2}{42} + \frac{30^2}{48} + \frac{70^2}{54} + \frac{100^2}{66} - 350 = 133,46$$

Estadístico teórico: $\chi_{0,05,3}^2 = 7,815$

Siendo $\chi_c^2 = 133,46 > \chi_{0,05,3}^2 = 7,815$ se rechaza la hipótesis nula, se acepta por tanto la hipótesis alternativa, pudiendo afirmar con una significación 0,05 que la opinión sobre el tabaco depende de sí es o no fumador.

- Coeficiente de contingencia: $C = \sqrt{\frac{\chi_c^2}{\chi_c^2 + n}} = \sqrt{\frac{133,46}{133,46 + 350}} = 0,525$

El grado de dependencia es del 52,5% por lo que la asociación entre las variables es alta.

En las tablas de contingencia $k \times k$ el valor máximo de C es $C_{\text{máximo}} = \sqrt{\frac{k-1}{k}}$

- Coeficiente Phi: $\phi = \sqrt{\frac{\chi_c^2}{n}} = \sqrt{\frac{133,46}{350}} = 0,618$

El estadístico Phi mide el grado de asociación entre las variables.

- Coeficiente V de Cramer:

$$V_{\text{Cramer}} = \sqrt{\frac{\chi_c^2}{n \cdot \min(k-1, m-1)}} = \sqrt{\frac{\chi_c^2}{n}} = \sqrt{\frac{133,46}{350}} = 0,618$$

En las tablas de contingencia 2×2 es idéntico al estadístico Phi, presenta el problema de subestimar el grado de asociación entre las variables.

- Test G de la razón de verosimilitud: $G = 2 \sum_{i=1}^k \sum_{j=1}^m n_{ij} \ln \left(\frac{n_{ij}}{e_{ij}} \right)$

Se acepta la hipótesis nula H_0 sí: $G = 2 \sum_{i=1}^2 \sum_{j=1}^4 n_{ij} \ln \left(\frac{n_{ij}}{e_{ij}} \right) < \chi_{\alpha, (2-1) \cdot (4-1)}^2$

	Opinión				$n_{i\cdot}$
	Muy en contra	En contra	A Favor	Muy a favor	
Fumador	60 $e_{11} = 28$ $g_{11} = 45,7$	50 $e_{12} = 32$ $g_{12} = 22,3$	20 $e_{13} = 36$ $g_{13} = -11,7$	10 $e_{14} = 44$ $g_{14} = -14,8$	140 140
No Fumador	10 $e_{21} = 42$ $g_{21} = -14,3$	30 $e_{22} = 48$ $g_{22} = -14,1$	70 $e_{23} = 54$ $g_{23} = 18,2$	100 $e_{24} = 66$ $g_{24} = 41,6$	210 210
$n_{\cdot j}$	70	80	90	110	350

$$g_{11} = 60 \ln\left(\frac{60}{28}\right) = 45,7 \quad g_{12} = 50 \ln\left(\frac{50}{32}\right) = 22,3 \quad g_{13} = 20 \ln\left(\frac{20}{36}\right) = -11,7 \quad g_{14} = 10 \ln\left(\frac{10}{44}\right) = -14,8$$

$$g_{21} = 10 \ln\left(\frac{10}{42}\right) = -14,3 \quad g_{22} = 30 \ln\left(\frac{30}{48}\right) = -14,1 \quad g_{23} = 70 \ln\left(\frac{70}{54}\right) = 18,2 \quad g_{24} = 100 \ln\left(\frac{100}{66}\right) = 41,6$$

$$G = 2 \sum_{i=1}^2 \sum_{j=1}^4 n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right) =$$

$$= 2[45,7 + 22,3 - 11,7 - 14,8 - 14,3 - 14,1 + 18,2 + 41,6] = 145,475$$

El test G da la razón de verosimilitud es una Prueba de hipótesis de la Chi-cuadrado que presenta mejores resultados que el Test de la Chi-cuadrado de Pearson.

■ Coeficiente Lambda (λ) de Goodman y Kruskal, conocido también como coeficiente de Goodman Predicción, se basa en la reducción proporcional del error en la predicción la moda, de es decir el número de aciertos que proporciona el conocer la distribución dividido por el número de errores sin conocerla.

$$\lambda_{yx} = \frac{\sum m_y - M_y}{n - M_y} \begin{cases} M_y \equiv \text{Frecuencia modal global} \\ \sum m_y \equiv \text{Suma de frecuencias modales} \\ n \equiv \text{Número total de casos} \end{cases}$$

$$\text{También, } \lambda = \frac{E_1 - E_2}{E_1} \begin{cases} E_1 = n - M_y \\ E_2 = n - \sum m_y \end{cases}$$

Valores Lambda (λ) próximos a 0 implican baja asociación y valores próximos a 1 denotan fuerte asociación.

Dos variables son independientes cuando $\lambda = 0$. Sin embargo $\lambda = 0$ no implica independencia estadística.

	Opinión				
	Muy en contra	En contra	A Favor	Muy a favor	$n_{i\cdot}$
Fumador	60	50	20	10	140
No Fumador	10	30	70	100	210
$n_{\cdot j}$	70	80	90	110	350

$$\lambda_{yx} = \frac{\sum m_y - M_y}{n - M_y} = \frac{280 - 210}{350 - 210} = 0,5 \quad \left\{ \begin{array}{l} M_y \equiv 210 \\ \sum m_y \equiv 60 + 50 + 70 + 100 = 280 \\ n \equiv 350 \end{array} \right.$$

$$\lambda_{yx} = \frac{E_1 - E_2}{E_1} = \frac{140 - 70}{140} = 0,5 \quad \left\{ \begin{array}{l} E_1 = n - M_y = 350 - 210 = 140 \\ E_2 = n - \sum m_y = 350 - 280 = 70 \end{array} \right.$$

$$\lambda_{xy} = \frac{\sum m_x - M_x}{n - M_x} = \frac{160 - 110}{350 - 110} = 0,208 \quad \left\{ \begin{array}{l} M_x \equiv 110 \\ \sum m_x \equiv 60 + 100 = 160 \\ n \equiv 350 \end{array} \right.$$

$$\lambda_{xy} = \frac{E_1 - E_2}{E_1} = \frac{240 - 190}{240} = 0,208 \quad \left\{ \begin{array}{l} E_1 = n - M_x = 350 - 110 = 240 \\ E_2 = n - \sum m_x = 350 - 160 = 190 \end{array} \right.$$

Un Fumador que estuviera Muy en contra de fumar en lugares públicos acertaría 60 veces de 70, es decir fallaría en 10 ocasiones. Un fumador que estuviera en contra tendría $80 - 50 = 30$ errores.

■ **Coeficiente Tau de Goodman y Kruskal:** Al igual que el coeficiente Lambda (λ) es un coeficiente asimétrico, aunque a diferencia del Lambda parte de los errores cometidos al asignar aleatoriamente los casos a las categorías de la variable dependiente.

$$\tau = \frac{E_1 - E_2}{E_1} \text{ donde } E_1 = \sum_{i=1}^k \left[\frac{(n - n_{i\bullet}) n_{i\bullet}}{n} \right] \text{ y } E_2 = \sum_{j=1}^m \sum_{i=1}^k \left[\frac{(n_{\bullet j} - n_{ij}) n_{ij}}{n_{\bullet j}} \right]$$

© Para conocer los errores sin conocer la distribución de la variable independiente:

Se supone que en cada categoría se clasificaran erróneamente por azar un número de casos, que en cada categoría es igual al número de casos que no pertenecen a la misma.

$$E_1 = \sum_{i=1}^k \left[\frac{(n - n_{i\bullet}) n_{i\bullet}}{n} \right] \begin{cases} n \equiv \text{número total de casos} \\ k \equiv \text{número de categorías de la variable} \\ n_{i\bullet} \equiv \text{frecuencia de la categoría } i\text{-ésima} \end{cases}$$

	Opinión				$n_{i\bullet}$
	Muy en contra	En contra	A Favor	Muy a favor	
Fumador	60	50	20	10	140
No Fumador	10	30	70	100	210
$n_{\bullet j}$	70	80	90	110	350

En la categoría de Fumadores de $n_{1\bullet} = 140$ de un total de $n = 350$ se cometerían $n - n_{1\bullet} = 350 - 140 = 210$ errores.

Intentando designar al azar los $n_{1\bullet} = 140$ casos de Fumadores se cometería

un error promedio de: $\frac{(n - n_{1\bullet})}{n} \times n_{1\bullet} = \frac{350 - 140}{350} \times 140 = 84$

En la categoría de No Fumadores de $n_{2\bullet} = 210$ de un total de $n = 350$ se cometerían $n - n_{2\bullet} = 350 - 210 = 140$ errores.

Intentando designar al azar los $n_{2\bullet} = 210$ casos de No Fumadores se

cometería un error promedio de: $\frac{(n - n_{2\bullet})}{n} \times n_{2\bullet} = \frac{350 - 210}{350} \times 210 = 84$

$$E_1 = \sum_{i=1}^2 \left[\frac{(350 - n_{i\bullet}) n_{i\bullet}}{350} \right] = 84 + 84 = 168$$

⊙ Para conocer los errores conociendo la distribución de la variable independiente:

$$E_2 = \sum_{j=1}^m \sum_{i=1}^k \left[\frac{(n_{\bullet j} - n_{ij}) n_{ij}}{n_{\bullet j}} \right]$$

$n_{ij} \equiv$ frecuencia de cada celdilla en la categoría i -ésima variable dependiente

$m \equiv$ número de categorías de la variable independiente

$n_{\bullet j} \equiv$ total parcial de las categorías de la variable independiente

□ Categoría con la opinión Muy en contra:

$$\text{Fumadores: } \frac{(n_{\bullet 1} - n_{11}) n_{11}}{n_{\bullet 1}} = \frac{(70 - 60) 60}{70} = 8,57$$

$$\text{No Fumadores: } \frac{(n_{\bullet 1} - n_{21}) n_{21}}{n_{\bullet 1}} = \frac{(70 - 10) 10}{70} = 8,57$$

$$\text{Errores en la categoría } E_{21} = 8,57 + 8,57 = 17,14$$

□ Categoría con la opinión En contra:

$$\text{Fumadores: } \frac{(n_{\bullet 2} - n_{12}) n_{12}}{n_{\bullet 2}} = \frac{(80 - 50) 50}{80} = 18,75$$

$$\text{No Fumadores: } \frac{(n_{\bullet 2} - n_{22}) n_{22}}{n_{\bullet 2}} = \frac{(80 - 30) 30}{80} = 18,75$$

$$\text{Errores en la categoría } E_{22} = 18,75 + 18,75 = 37,5$$

□ Categoría con la opinión A favor:

$$\text{Fumadores: } \frac{(n_{\bullet 3} - n_{13}) n_{13}}{n_{\bullet 3}} = \frac{(90 - 20) 20}{90} = 15,56$$

$$\text{No Fumadores: } \frac{(n_{\bullet 3} - n_{23}) n_{23}}{n_{\bullet 3}} = \frac{(90 - 70) 70}{90} = 15,56$$

$$\text{Errores en la categoría } E_{23} = 15,56 + 15,56 = 31,12$$

□ Categoría con la opinión Muy a favor:

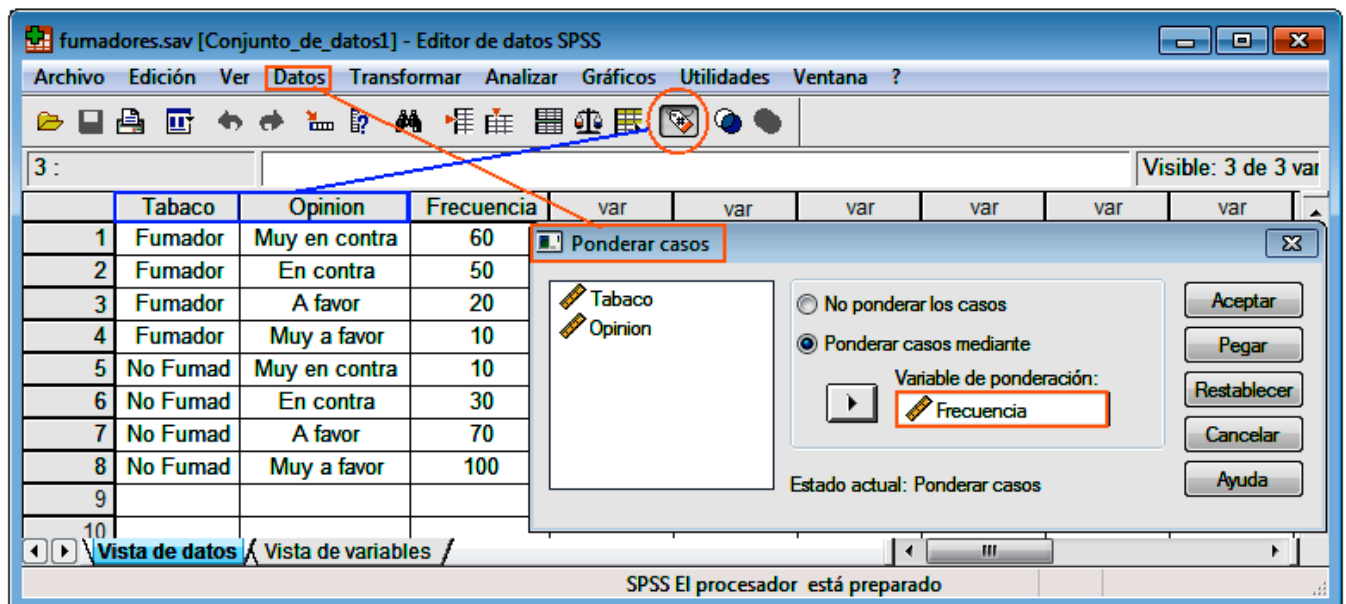
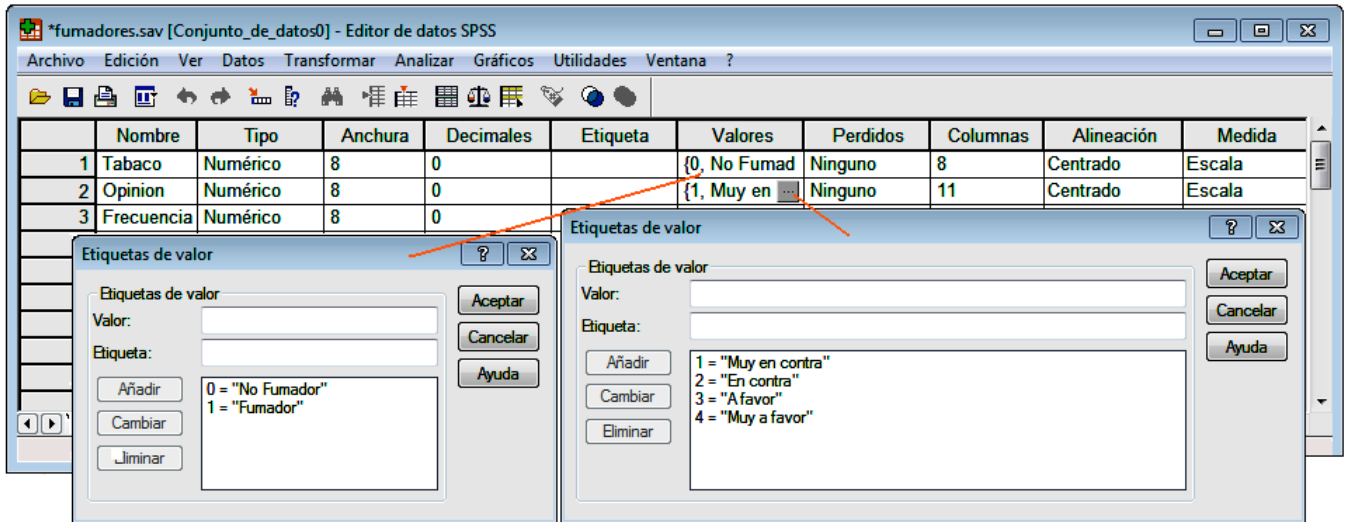
$$\text{Fumadores: } \frac{(n_{\bullet 4} - n_{14}) n_{14}}{n_{\bullet 4}} = \frac{(110 - 10) 10}{110} = 9,09$$

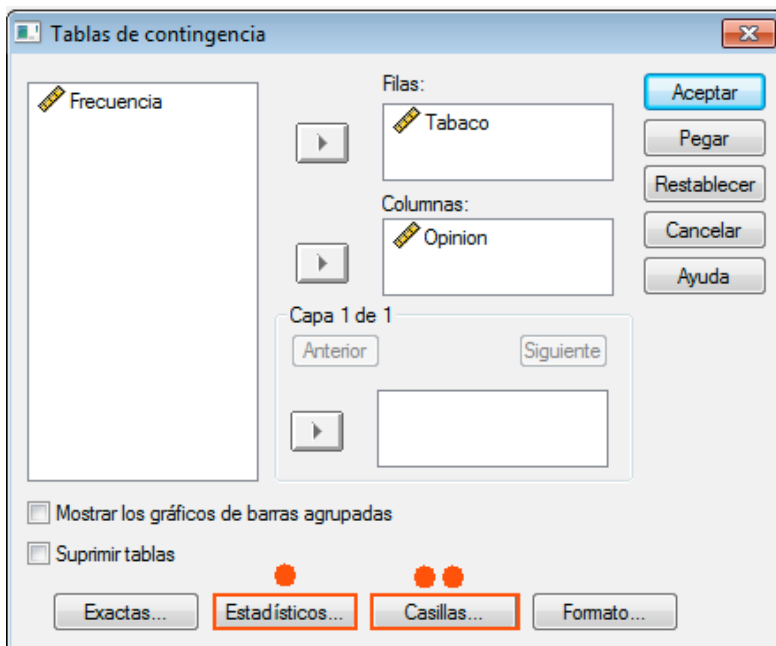
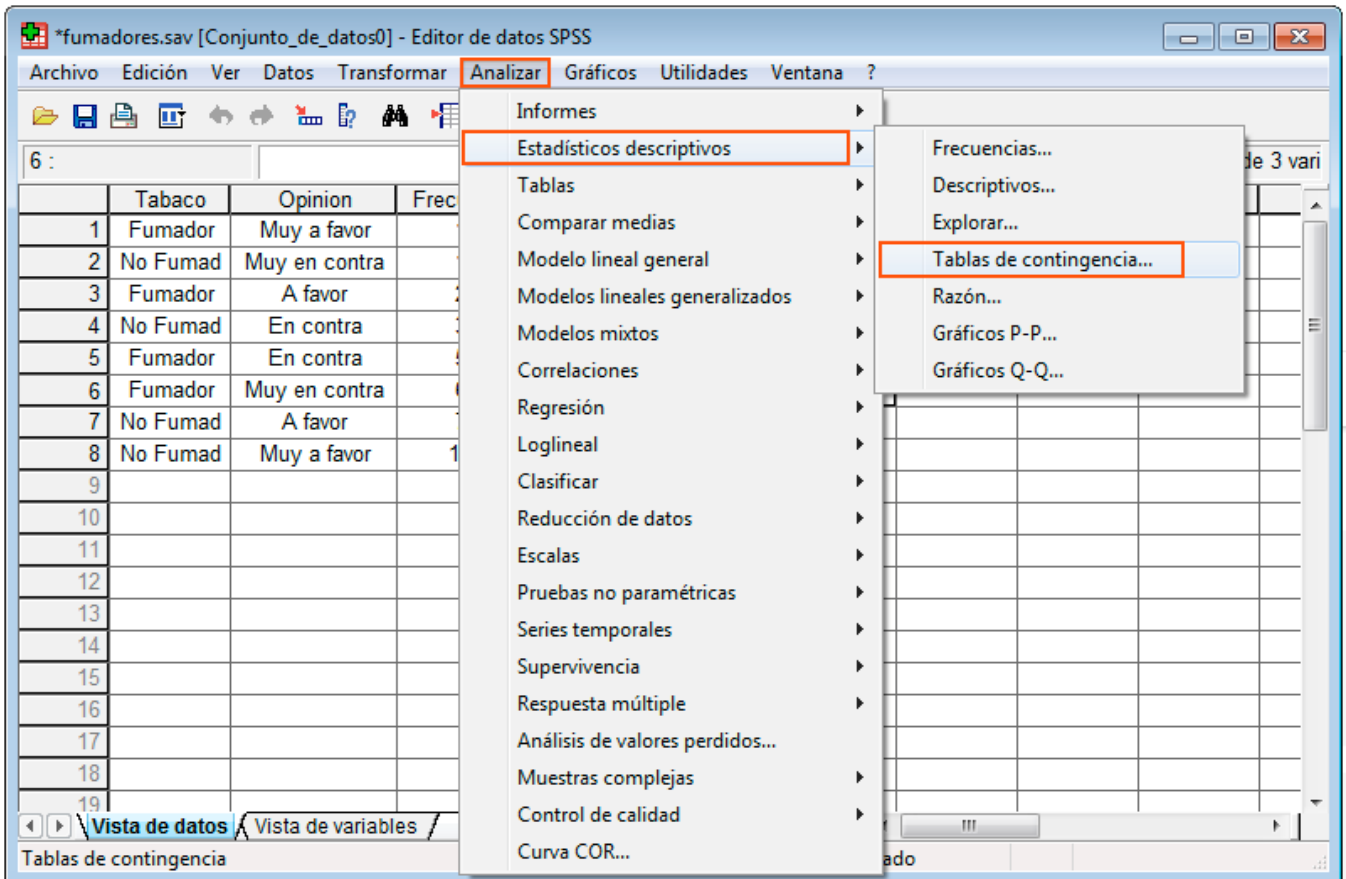
$$\text{No Fumadores: } \frac{(n_{\bullet 4} - n_{24}) n_{24}}{n_{\bullet 4}} = \frac{(110 - 100) 100}{110} = 9,09$$

$$\text{Errores en la categoría } E_{24} = 9,09 + 9,09 = 18,18$$

$$E_2 = \sum_{j=1}^4 \sum_{i=1}^2 \left[\frac{(n_{\bullet j} - n_{ij}) n_{ij}}{n_{\bullet j}} \right] = 17,14 + 37,5 + 31,12 + 18,18 = 103,94$$

$$\tau = \frac{E_1 - E_2}{E_1} = \frac{168 - 103,94}{168} = 0,381$$





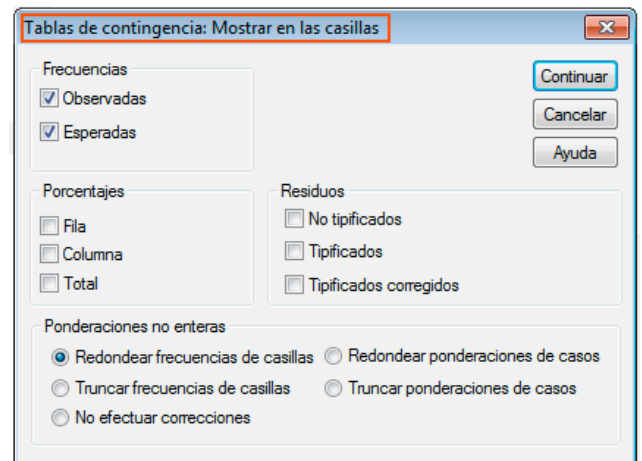
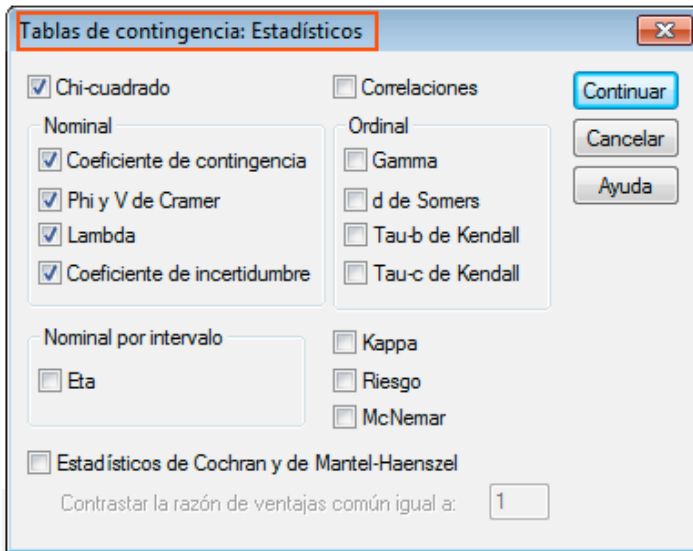


Tabla de contingencia Tabaco * Opinión

			Opinion				Total
			Muy en contra	En contra	A favor	Muy a favor	
Tabaco	No Fumador	Recuento	10	30	70	100	210
		Frecuencia esperada	42	48	54	66	210
	Fumador	Recuento	60	50	20	10	140
		Frecuencia esperada	28	32	36	44	140
Total		Recuento	70	80	90	110	350
		Frecuencia esperada	70	80	90	110	350

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	133,467 ^a	3	,000
Razón de verosimilitudes	145,475	3	,000
Asociación lineal por lineal	128,394	1	,000
N de casos válidos	350		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5.
La frecuencia mínima esperada es 28,00.

La Prueba de Chi-cuadrado de Pearson y la Prueba de Razón de verosimilitudes presenta un p-valor (Sig. asintótica bilateral) = 0,000 < 0,05
En consecuencia, se rechaza la hipótesis nula, concluyendo que la opinión sobre el tabaco depende de sí es o no fumador.

Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,618	,000
	V de Cramer	,618	,000
	Coefficiente de contingencia	,525	,000
N de casos válidos		350	

a. Asumiendo la hipótesis alternativa.

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

Para cuantificar el grado de asociación entre las variables se parte de la hipótesis nula H_0 : Las variables Tabaco y Opinión no están relacionadas, es decir, son independientes.

Los estadísticos Phi, V de Cramer, Coeficiente de contingencia toman valores moderados, aceptando que existe relación de dependencia entre las dos variables para niveles de significación superiores a 0,000.

Medidas direccionales

			Valor	Error típ. asint. ^a	T aproximada ^b	Sig. aproximada
Nominal por nominal	Lambda	Simétrica	,316	,039	6,941	,000
		Tabaco dependiente	,500	,062	6,002	,000
		Opinion dependiente	,208	,031	6,307	,000
	Tau de Goodman y Kruskal	Tabaco dependiente	,381	,048		,000 ^c
		Opinion dependiente	,126	,018		,000 ^c

a. Asumiendo la hipótesis alternativa.

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

c. Basado en la aproximación chi-cuadrado.

Los estadísticos Lambda y Tau de Goodman y Kruskal son índices de asociación que parten de que las variables están relacionadas.

Tomando la variable Tabaco como dependiente el valor de Lambda es 0,500 con un error típico asintótico de 0,062, reflejando que el conocimiento de la variable Opinión permite reducir la incertidumbre en la predicción de los valores de la variable Tabaco en un 50%, afirmación significativa para niveles de significación superiores a 0,000.

El estadístico Tau de Goodman y Kruskal con la variable Tabaco como dependiente toma el valor 0,381 con un error típico asintótico de 0,048, corroborando el diagnóstico anterior.

📄 En la tabla se refleja la edad de los empleados de una empresa y el grado de satisfacción en el trabajo, con un nivel de significación del 5%, contrastar si el grado de satisfacción en el trabajo no depende de la edad de los empleados.

Edad	Satisfacción en el trabajo				
	A	B	C	D	E
< 25	10	10	20	40	70
25 - 36	20	10	15	20	30
> 36	60	50	30	10	5

Solución:

Variables: X= 'edad de los empleados' e Y= 'satisfacción en el trabajo'

Hipótesis nula H_0 : 'El grado de satisfacción en el trabajo no depende de la edad de los empleados'

Se acepta H_0 :

$$\chi_c^2 = \sum_{i=1}^3 \sum_{j=1}^5 \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^3 \sum_{j=1}^5 \frac{n_{ij}^2}{e_{ij}} - n < \chi_{\alpha; (3-1) \cdot (5-1)}^2$$

Se forma la tabla de contingencia 3 x 5 donde cada frecuencia observada $(n_{ij})_{i=1,2,3 ; j=1,\dots,5}$ tiene una frecuencia teórica o esperada en caso de

independencia $e_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$

Edad	Satisfacción en el trabajo					$n_{i\cdot}$
	A	B	C	D	E	
< 25	10 $e_{11} = 33,75$	10 $e_{12} = 26,25$	20 $e_{13} = 24,37$	40 $e_{14} = 26,25$	70 $e_{15} = 39,37$	150 (150)
25 - 36	20 $e_{21} = 21,37$	10 $e_{22} = 16,62$	15 $e_{23} = 15,44$	20 $e_{24} = 16,62$	30 $e_{25} = 24,94$	95 (95)
> 36	60 $e_{31} = 34,87$	50 $e_{32} = 27,12$	30 $e_{33} = 25,19$	10 $e_{34} = 27,12$	5 $e_{35} = 40,69$	155 (155)
$n_{\cdot j}$	90	70	65	70	105	400

$$e_{11} = \frac{150 \cdot 90}{400} = 33,75$$

$$e_{21} = \frac{95 \cdot 90}{400} = 21,37$$

$$e_{31} = \frac{155 \cdot 90}{400} = 34,87$$

$$e_{12} = \frac{150 \cdot 70}{400} = 26,25$$

$$e_{22} = \frac{95 \cdot 70}{400} = 16,62$$

$$e_{32} = \frac{155 \cdot 70}{400} = 27,12$$

$$e_{13} = \frac{150 \cdot 65}{400} = 24,37$$

$$e_{23} = \frac{95 \cdot 65}{400} = 15,44$$

$$e_{33} = \frac{155 \cdot 65}{400} = 25,19$$

$$e_{14} = \frac{150 \cdot 70}{400} = 26,25$$

$$e_{24} = \frac{95 \cdot 70}{400} = 16,62$$

$$e_{34} = \frac{155 \cdot 70}{400} = 27,12$$

$$e_{15} = \frac{150 \cdot 105}{400} = 39,37$$

$$e_{25} = \frac{95 \cdot 105}{400} = 24,94$$

$$e_{35} = \frac{155 \cdot 105}{400} = 40,69$$

Estadístico observado: $\chi_c^2 = \sum_{i=1}^3 \sum_{j=1}^5 \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^3 \sum_{j=1}^5 \frac{n_{ij}^2}{e_{ij}} - n =$

$$= \left(\frac{10^2}{33,75} + \frac{10^2}{26,25} + \frac{20^2}{24,37} + \frac{40^2}{26,25} + \frac{70^2}{39,37} \right) + \left(\frac{20^2}{21,37} + \frac{10^2}{16,62} + \frac{15^2}{15,44} + \frac{20^2}{16,62} + \frac{30^2}{24,94} \right) + \left(\frac{60^2}{34,87} + \frac{50^2}{27,12} + \frac{30^2}{25,19} + \frac{10^2}{27,12} + \frac{5^2}{40,69} \right) - 400 = 143,458$$

Estadístico teórico: $\chi_{0,05; (3-1) \cdot (5-1)}^2 = \chi_{0,05; 8}^2 = 15,507$

Como $\chi_8^2 = 143,458 > 15,507 = \chi_{0,05; 8}^2$ se rechaza la hipótesis nula de independencia entre la edad y la satisfacción en el trabajo. En consecuencia, la edad influye significativamente en la satisfacción en el trabajo.

ESTADÍSTICOS VARIABLES NOMINALES: FUERZA DE LA RELACIÓN

- Coeficiente Phi: $\phi = \sqrt{\frac{\chi_c^2}{n}} = \sqrt{\frac{143,51}{400}} = 0,599$

El estadístico Phi mide el grado de asociación entre las variables.

- Coeficiente V de Cramer:

$$V_{\text{Cramer}} = \sqrt{\frac{\chi_c^2}{n \cdot \min(k-1, m-1)}} = \sqrt{\frac{143,51}{400 \cdot \min(3-1, 5-1)}} = \sqrt{\frac{143,51}{400 \cdot 2}} = 0,423$$

El estadístico V de Cramer es una medida simétrica que cuantifica la relación entre dos o más variables de la escala nominal. Quizás es el estadístico más utilizado.

Un valor del estadístico V de Cramer próximo a 0 indica la falta de asociación de las variables, mientras que próximo a 1 refleja mayor asociación entre las variables en estudio.

Como $V_{\text{Cramer}} = 0,423$ se detecta una relación moderada de las variables.

■ Coeficiente de contingencia: $C = \sqrt{\frac{\chi_c^2}{\chi_c^2 + n}} = \sqrt{\frac{143,51}{143,51 + 400}} = 0,514$

El grado de dependencia es del 51,4% por lo que la asociación entre las variables es alta.

■ Test G de la razón de verosimilitud: $G = 2 \sum_{i=1}^k \sum_{j=1}^m n_{ij} \ln \left(\frac{n_{ij}}{e_{ij}} \right)$

Se acepta la hipótesis nula H_0 sí: $G = 2 \sum_{i=1}^3 \sum_{j=1}^5 n_{ij} \ln \left(\frac{n_{ij}}{e_{ij}} \right) < \chi_{\alpha, (3-1) \cdot (5-1)}^2$

Edad	Satisfacción en el trabajo					$n_{i\cdot}$
	A	B	C	D	E	
< 25	10 $e_{11} = 33,75$ $g_{11} = -12,16$	10 $e_{12} = 26,25$ $g_{12} = -9,65$	20 $e_{13} = 24,37$ $g_{13} = -3,95$	40 $e_{14} = 26,25$ $g_{14} = 16,85$	70 $e_{15} = 39,37$ $g_{15} = 40,28$	150 (150) (31,37)
25 - 36	20 $e_{21} = 21,37$ $g_{21} = -1,33$	10 $e_{22} = 16,62$ $g_{22} = -5,08$	15 $e_{23} = 15,44$ $g_{23} = -0,43$	20 $e_{24} = 16,62$ $g_{24} = 3,7$	30 $e_{25} = 24,94$ $g_{25} = 5,54$	95 (95) (2,4)
> 36	60 $e_{31} = 34,87$ $g_{31} = 32,56$	50 $e_{32} = 27,12$ $g_{32} = 30,59$	30 $e_{33} = 25,19$ $g_{33} = 5,24$	10 $e_{34} = 27,12$ $g_{34} = -9,98$	5 $e_{35} = 40,69$ $g_{35} = -10,48$	155 (155) (47,93)
$n_{\cdot j}$	90	70	65	70	105	400 (81,7)

$$G = 2 \sum_{i=1}^3 \sum_{j=1}^5 n_{ij} \ln \left(\frac{n_{ij}}{e_{ij}} \right) = 2 \cdot 81,667 = 163,334 > 15,507 = \chi_{0,05;8}^2$$

Se rechaza la hipótesis nula de independencia entre la edad y la satisfacción en el trabajo, concluyendo que la edad influye significativamente en la satisfacción en el trabajo.

$$\begin{aligned}
 g_{11} &= 10 \ln\left(\frac{10}{33,75}\right) = -12,16 & g_{21} &= 20 \ln\left(\frac{20}{21,37}\right) = -1,33 & g_{31} &= 60 \ln\left(\frac{60}{34,87}\right) = 32,56 \\
 g_{12} &= 10 \ln\left(\frac{10}{26,25}\right) = -9,65 & g_{22} &= 10 \ln\left(\frac{10}{16,62}\right) = -5,08 & g_{32} &= 50 \ln\left(\frac{50}{27,12}\right) = 30,59 \\
 g_{13} &= 20 \ln\left(\frac{20}{24,37}\right) = -3,95 & g_{23} &= 15 \ln\left(\frac{15}{15,44}\right) = -0,43 & g_{33} &= 30 \ln\left(\frac{30}{25,19}\right) = 5,24 \\
 g_{14} &= 40 \ln\left(\frac{40}{26,25}\right) = 16,85 & g_{24} &= 20 \ln\left(\frac{20}{16,62}\right) = 3,7 & g_{34} &= 10 \ln\left(\frac{10}{27,12}\right) = -9,98 \\
 g_{15} &= 70 \ln\left(\frac{70}{39,37}\right) = 40,28 & g_{25} &= 30 \ln\left(\frac{30}{24,94}\right) = 5,54 & g_{35} &= 5 \ln\left(\frac{5}{40,69}\right) = -10,48
 \end{aligned}$$

El test G da la razón de verosimilitud es una Prueba de hipótesis que presenta mejores resultados que el Test de la Chi-cuadrado de Pearson.



MEDIDAS DE ASOCIACIÓN DE VARIABLES ORDINALES

Edad	Satisfacción en el trabajo					$n_{i\cdot}$
	A	B	C	D	E	
< 25	10	10	20	40	70	150
25 - 36	20	10	15	20	30	95
> 36	60	50	30	10	5	155
$n_{\cdot j}$	90	70	65	70	105	400

Pares Concordantes:

$$\begin{aligned}C &= 10[10 + 15 + 20 + 30 + 50 + 30 + 10 + 5] \\ &+ 10[15 + 20 + 30 + 30 + 10 + 5] \\ &+ 20[20 + 30 + 10 + 5] \\ &+ 40[30 + 5] \\ &+ 20[50 + 30 + 10 + 5] \\ &+ 10[30 + 10 + 5] \\ &+ 15[10 + 5] \\ &+ 20[5] \\ &= 8175\end{aligned}$$

Pares Discordantes:

$$\begin{aligned}D &= 70[20 + 10 + 15 + 20 + 60 + 50 + 30 + 10] \\ &+ 40[20 + 10 + 15 + 60 + 50 + 30] \\ &+ 20[20 + 10 + 60 + 50] \\ &+ 10[20 + 60] \\ &+ 30[60 + 50 + 30 + 10] \\ &+ 20[60 + 50 + 30] \\ &+ 15[60 + 50] \\ &+ 10[60] \\ &= 35600\end{aligned}$$

- La Gamma de Goodman y Kruskal γ mide la fuerza de asociación de los datos cuando las variables se miden en el nivel ordinal.

$\gamma = 0$ indica la ausencia de asociación.

$$\gamma = \frac{C - D}{C + D} \quad -1 \leq \gamma \leq 1$$

$$\gamma = \frac{C - D}{C + D} = \frac{8175 - 35600}{8175 + 35600} = -0,626$$

- El coeficiente de rango de Kendall (τ_c) a menudo se utiliza como un estadístico control en una prueba de hipótesis estadística para establecer si dos variables pueden considerarse estadísticamente dependientes.

Es una prueba no paramétrica, ya que no se basa en suposiciones sobre las distribuciones de X o Y o la distribución de (X, Y).

Bajo la hipótesis nula de independencia de X e Y, la distribución muestral de Tau-C (τ_c) tiene un valor esperado de cero.

Para muestras pequeñas:
$$\tau_c = \frac{2 \cdot \min(k, m) \cdot (C - D)}{\min(k - 1, m - 1) \cdot n^2}$$

En muestras grandes, se utiliza una aproximación a $N(0, 1)$:
$$\tau_c = \frac{2(2n + 5)}{9n(n - 1)}$$

$$\tau_c = \frac{2 \cdot \min(k, m) \cdot (C - D)}{\min(k - 1, m - 1) \cdot n^2} = \frac{2 \cdot \min(3, 5) \cdot (8175 - 35600)}{\min(3 - 1, 5 - 1) \cdot 400^2} = \frac{2 \cdot 3 \cdot (-27425)}{2 \cdot 400^2} = -0,514$$

- Parejas empatadas en X o en Y:

$$T_x = \sum_{i=1}^k \frac{n_{i\cdot} (n_{i\cdot} - 1)}{2} \quad T_y = \sum_{j=1}^m \frac{n_{\cdot j} (n_{\cdot j} - 1)}{2}$$

$$T_x = \sum_{i=1}^3 \frac{n_{i\cdot} (n_{i\cdot} - 1)}{2} = \frac{1}{2} [150 \cdot 149 + 95 \cdot 94 + 155 \cdot 154] = 27575$$

$$T_y = \sum_{j=1}^5 \frac{n_{\cdot j} (n_{\cdot j} - 1)}{2} = \frac{1}{2} [90 \cdot 89 + 70 \cdot 69 + 65 \cdot 64 + 70 \cdot 69 + 105 \cdot 104] = 16375$$

- El coeficiente Tau-B de Kendall (τ_b) es una medida no paramétrica de la correlación para variables ordinales o de rangos que tiene en consideración los empates.

El signo del coeficiente indica la dirección de la relación y su valor absoluto indica la fuerza de la relación. Varía entre -1 y 1 según sea el sentido de la asociación entre las variables. Los valores mayores indican que la relación es más estrecha.

Cuando la tabla no es cuadrada este coeficiente no puede llegar a valer 1 dado que existirán más pares empatados en la variable que tenga más categorías.

$$\tau_B = \frac{C - D}{\sqrt{\left(\frac{n(n-1)}{2} - T_x\right)\left(\frac{n(n-1)}{2} - T_y\right)}}$$

$$\tau_B = \frac{8175 - 35600}{\sqrt{(79800 - 27575)(79800 - 16375)}} = -0,477$$

■ El estadístico D de Somers establece si las variables ordinales son dependientes o independientes entre sí.

El coeficiente D de Somers varía entre -1 y 1 , es una medida asimétrica como el coeficiente Lambda, los dos valores que se pueden obtener de la tabla dependen de que se tome como independiente la variable X o Y.

Valores del estadístico D cercanos a 0 indican que no hay ninguna o muy poca asociación entre las variables.

$$D \text{ de Somers: } D_x = \frac{C - D}{\frac{n(n-1)}{2} - T_x} \quad D_y = \frac{C - D}{\frac{n(n-1)}{2} - T_y}$$

$$\text{Número de pares: } \binom{n}{2} = \frac{n(n-1)}{2} = \frac{400(400-1)}{2} = 79800$$

$$D_x = \frac{C - D}{\frac{n(n-1)}{2} - T_x} = \frac{8175 - 35600}{79800 - 27575} = -0,525$$

$$D_y = \frac{C - D}{\frac{n(n-1)}{2} - T_x} = \frac{8175 - 35600}{79800 - 16375} = -0,432$$



MEDIDAS BASADAS EN EL ERROR PROPORCIONAL

■ Coeficiente Lambda (λ) de Goodman y Kruskal, conocido también como coeficiente de Goodman Predicción, se basa en la reducción proporcional del error en la predicción la moda.

Estadístico utilizado para determinar si usar los resultados de una de las variables puede utilizarse para predecir los resultados de la otra variable.

Valores Lambda (λ) próximos a 0 implican baja asociación y valores próximos a 1 denotan fuerte asociación.

Dos variables son independientes tienen $\lambda = 0$. Sin embargo $\lambda = 0$ no implica independencia estadística.

Edad	Satisfacción en el trabajo					$n_{i\cdot}$
	A	B	C	D	E	
< 25	10	10	20	40	70	150
25 - 36	20	10	15	20	30	95
> 36	60	50	30	10	5	155
$n_{\cdot j}$	90	70	65	70	105	400

$$\lambda_{yx} = \frac{\sum m_y - M_y}{n - M_y} \begin{cases} M_y \equiv \text{Frecuencia modal global} \\ \sum m_y \equiv \text{Suma de frecuencias modales} \\ n \equiv \text{Número total de casos} \end{cases}$$

$$\text{También, } \lambda = \frac{E_1 - E_2}{E_1} \begin{cases} E_1 = n - M_y \\ E_2 = n - \sum m_y \end{cases}$$

$$\lambda_{yx} = \frac{\sum m_y - M_y}{n - M_y} = \frac{250 - 155}{400 - 155} = 0,388 \begin{cases} M_y \equiv 155 \\ \sum m_y \equiv 60 + 50 + 30 + 40 + 70 = 250 \\ n \equiv 400 \end{cases}$$

$$\lambda_{xy} = \frac{\sum m_x - M_x}{n - M_x} = \frac{160 - 105}{400 - 105} = 0,186 \quad \left\{ \begin{array}{l} M_x \equiv 105 \\ \sum m_x \equiv 70 + 30 + 60 = 160 \\ n \equiv 400 \end{array} \right.$$

■ Tau de Goodman y Kruskal (τ) considera todas las categorías de respuesta y no únicamente la que contempla más casos entre dos variables nominales (variables cualitativas).

El valor de Tau de Goodman y Kruskal (τ) se interpreta como el porcentaje que mejora el error al incluir la variable independiente en la predicción de los valores de la variable dependiente.

Se parece a la Lambda (λ), siendo su cálculo más complejo. Lo mismo que Lambda adopta valores entre 0 y 1, dónde 0 es independencia y 1 el total de dependencia.

$$\tau = \frac{E_1 - E_2}{E_1}$$

❶ Errores sin conocer la distribución de la variable independiente:

$$E_1 = \sum_{i=1}^k \left[\frac{(n - n_{i\cdot}) n_{i\cdot}}{n} \right] \quad \left\{ \begin{array}{l} n \equiv \text{número total de casos} \\ k \equiv \text{número de categorías de la variable} \\ n_{i\cdot} \equiv \text{frecuencia de la categoría } i\text{-ésima} \end{array} \right.$$

❷ Errores conociendo la distribución de la variable independiente:

$$E_2 = \sum_{j=1}^m \sum_{i=1}^k \left[\frac{(n_{\cdot j} - n_{ij}) n_{ij}}{n_{\cdot j}} \right]$$

n_{ij} \equiv frecuencia de cada celdilla en la categoría i -ésima variable dependiente

m \equiv número de categorías de la variable independiente

$n_{\cdot j}$ \equiv total parcial de las categorías de la variable independiente

Edad	Satisfacción en el trabajo					$n_{i\cdot}$
	A	B	C	D	E	
< 25	10	10	20	40	70	150
25 - 36	20	10	15	20	30	95
> 36	60	50	30	10	5	155
$n_{\cdot j}$	90	70	65	70	105	400

$$E_1 = \sum_{i=1}^3 \left[\frac{(n - n_{i\cdot}) n_{i\cdot}}{n} \right] = \frac{(400 - 150)150}{400} + \frac{(400 - 95)95}{400} + \frac{(400 - 155)155}{400} = 261,125$$

$$E_2 = \sum_{j=1}^5 \sum_{i=1}^3 \left[\frac{(n_{\cdot j} - n_{ij}) n_{ij}}{n_{\cdot j}} \right] = \frac{(90 - 10)10}{90} + \frac{(90 - 20)20}{90} + \frac{(90 - 60)60}{90} \\ + \frac{(70 - 10)10}{70} + \frac{(70 - 10)10}{70} + \frac{(70 - 50)50}{70} \\ + \frac{(65 - 20)20}{65} + \frac{(65 - 15)15}{65} + \frac{(65 - 30)30}{65} \\ + \frac{(70 - 40)40}{70} + \frac{(70 - 20)20}{70} + \frac{(70 - 10)10}{70} \\ + \frac{(105 - 70)70}{105} + \frac{(105 - 30)30}{105} + \frac{(105 - 5)5}{105} \\ = 206,93$$

$$\tau = \frac{E_1 - E_2}{E_1} = \frac{261,125 - 206,93}{261,125} = 0,208 \text{ edad variable dependiente}$$

Quando la variable dependiente es la satisfacción en el trabajo:

$$E_1 = \sum_{j=1}^5 \left[\frac{(n - n_{\cdot j}) n_{\cdot j}}{n} \right] = \\ = \frac{(400 - 90)90}{400} + \frac{(400 - 70)70}{400} + \frac{(400 - 65)65}{400} + \frac{(400 - 70)70}{400} + \frac{(400 - 105)105}{400} = 317,125$$

$$\begin{aligned}
E_2 &= \sum_{i=1}^3 \sum_{j=1}^5 \left[\frac{(n_{i\cdot} - n_{ij}) n_{ij}}{n_{i\cdot}} \right] = \\
&= \frac{(150-10)10}{150} + \frac{(150-10)10}{150} + \frac{(150-20)20}{150} + \frac{(150-40)40}{150} + \frac{(150-70)70}{150} \\
&+ \frac{(95-20)20}{95} + \frac{(95-10)10}{95} + \frac{(95-15)15}{95} + \frac{(95-20)20}{95} + \frac{(95-30)30}{95} \\
&+ \frac{(155-60)60}{155} + \frac{(155-50)50}{155} + \frac{(155-30)30}{155} + \frac{(155-10)10}{155} + \frac{(155-5)5}{155} \\
&= 285,38
\end{aligned}$$

$$\tau = \frac{E_1 - E_2}{E_1} = \frac{317,125 - 285,38}{317,125} = 0,100 \quad \text{satisfacción variable dependiente}$$

■ El Coeficiente de Incertidumbre es una medida de asociación basada en la reducción proporcional del error. Es una medida semejante a Lambda en cuanto a su concepción de la asociación de las variables, en relación a la capacidad predictiva y la disminución del error de dicha predicción.

El coeficiente de incertidumbre (I) depende de toda la distribución y no sólo de los valores modales (caso de Lambda), varía entre 0 y 1, tomando el valor 0 en el caso total de independencia. Es más difícil de interpretar que Lambda.

Tiene versiones asimétricas (dependiendo de cual de las dos variables sea dependiente) y una simétrica (donde no se distingue entre variable dependiente e independiente).

La versión asimétrica se interpreta como la proporción de incertidumbre reducida al predecir los valores de una variable a partir de los de valores de la otra variable.

La versión simétrica se interpreta como la proporción de incertidumbre reducida al predecir los valores de cualquiera de las dos variables mediante la tabla de contingencia.

Se obtiene mediante la fórmula:

$$I_{Y/X} = \frac{I(X) + I(Y) - I(XY)}{I(Y)}$$

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

La versión simétrica: $I = \frac{2 [I(X) + I(Y) - I(XY)]}{I(X) + I(Y)}$

donde: $I(X) = \sum_{i=1}^k \frac{n_{i\cdot}}{n} \ln\left(\frac{n_{i\cdot}}{n}\right)$ $I(Y) = \sum_{j=1}^m \frac{n_{\cdot j}}{n} \ln\left(\frac{n_{\cdot j}}{n}\right)$ $I(XY) = \sum_{i=1}^k \sum_{j=1}^m \frac{n_{ij}}{n} \ln\left(\frac{n_{ij}}{n}\right)$

Edad	Satisfacción en el trabajo					$i_{i\cdot}$
	A	B	C	D	E	
< 25	10 $i_{11} = -0,092$	10 $i_{12} = -0,092$	20 $i_{13} = -0,150$	40 $i_{14} = -0,230$	70 $i_{15} = -0,305$	150 $-0,368$
25 - 36	20 $i_{21} = -0,150$	10 $i_{22} = -0,092$	15 $i_{23} = -0,123$	20 $i_{24} = -0,150$	30 $i_{25} = -0,194$	95 $-0,341$
> 36	60 $i_{31} = -0,284$	50 $i_{32} = -0,260$	30 $i_{33} = -0,194$	10 $i_{34} = -0,092$	5 $i_{35} = -0,055$	155 $-0,367$
$i_{\cdot j}$	90 $-0,335$	70 $-0,305$	65 $-0,295$	70 $-0,305$	105 $-0,351$	400

$$i_{1\cdot} = \frac{150}{400} \ln\left(\frac{150}{400}\right) = -0,368 \quad i_{2\cdot} = \frac{95}{400} \ln\left(\frac{95}{400}\right) = -0,341 \quad i_{3\cdot} = \frac{155}{400} \ln\left(\frac{155}{400}\right) = -0,367$$

$$I(X) = \sum_{i=1}^3 \frac{n_{i\cdot}}{n} \ln\left(\frac{n_{i\cdot}}{n}\right) = -0,368 - 0,341 - 0,367 = -1,076$$

$$i_{\cdot 1} = \frac{90}{400} \ln\left(\frac{90}{400}\right) = -0,335 \quad i_{\cdot 2} = \frac{70}{400} \ln\left(\frac{70}{400}\right) = -0,305 \quad i_{\cdot 3} = \frac{65}{400} \ln\left(\frac{65}{400}\right) = -0,295$$

$$i_{\cdot 4} = \frac{70}{400} \ln\left(\frac{70}{400}\right) = -0,305 \quad i_{\cdot 5} = \frac{105}{400} \ln\left(\frac{105}{400}\right) = -0,351$$

$$I(Y) = \sum_{j=1}^5 \frac{n_{\cdot j}}{n} \ln\left(\frac{n_{\cdot j}}{n}\right) = -0,335 - 0,305 - 0,295 - 0,305 - 0,351 = -1,591$$

$$i_{11} = \frac{10}{400} \ln\left(\frac{10}{400}\right) = -0,092 \quad i_{21} = \frac{20}{400} \ln\left(\frac{20}{400}\right) = -0,150 \quad i_{31} = \frac{60}{400} \ln\left(\frac{60}{400}\right) = -0,284$$

$$i_{12} = \frac{10}{400} \ln\left(\frac{10}{400}\right) = -0,092 \quad i_{22} = \frac{10}{400} \ln\left(\frac{10}{400}\right) = -0,092 \quad i_{32} = \frac{50}{400} \ln\left(\frac{50}{400}\right) = -0,260$$

$$i_{13} = \frac{20}{400} \ln\left(\frac{20}{400}\right) = -0,150 \quad i_{23} = \frac{15}{400} \ln\left(\frac{15}{400}\right) = -0,123 \quad i_{33} = \frac{30}{400} \ln\left(\frac{30}{400}\right) = -0,194$$

$$i_{14} = \frac{40}{400} \ln\left(\frac{40}{400}\right) = -0,230 \quad i_{24} = \frac{20}{400} \ln\left(\frac{20}{400}\right) = -0,150 \quad i_{34} = \frac{10}{400} \ln\left(\frac{10}{400}\right) = -0,092$$

$$i_{15} = \frac{70}{400} \ln\left(\frac{70}{400}\right) = -0,305 \quad i_{25} = \frac{30}{400} \ln\left(\frac{30}{400}\right) = -0,194 \quad i_{35} = \frac{5}{400} \ln\left(\frac{5}{400}\right) = -0,055$$

$$I(XY) = \sum_{i=1}^3 \sum_{j=1}^5 \frac{n_{ij}}{n} \ln\left(\frac{n_{ij}}{n}\right) = -2,463$$

Coefficiente de Incertidumbre, Satisfacción como variable dependiente:

$$I_{Y/X} = \frac{I(X) + I(Y) - I(XY)}{I(Y)} = \frac{-1,076 - 1,591 + 2,463}{-1,591} = 0,128$$

Coefficiente de Incertidumbre, Edad como variable dependiente:

$$I_{X/Y} = \frac{I(X) + I(Y) - I(XY)}{I(X)} = \frac{-1,076 - 1,591 + 2,463}{-1,076} = 0,190$$

Coefficiente de Incertidumbre simétrico:

$$I = \frac{2 [I(X) + I(Y) - I(XY)]}{I(X) + I(Y)} = \frac{2 [-1,076 - 1,591 + 2,463]}{-1,076 - 1,591} = 0,153$$

H₀: Las variables son independientes

Pruebas significación estadística { Chi-cuadrado de Pearson
Razón de verosimilitud Chi-cuadrado

H₀: La asociación entre las variables es nula (son independientes)

Estadísticos Nominales { Phi
Coeficiente de Contingencia
V de Cramer
Variables Cualitativas { Lambda
Coeficiente de Incertidumbre
Q de Yule

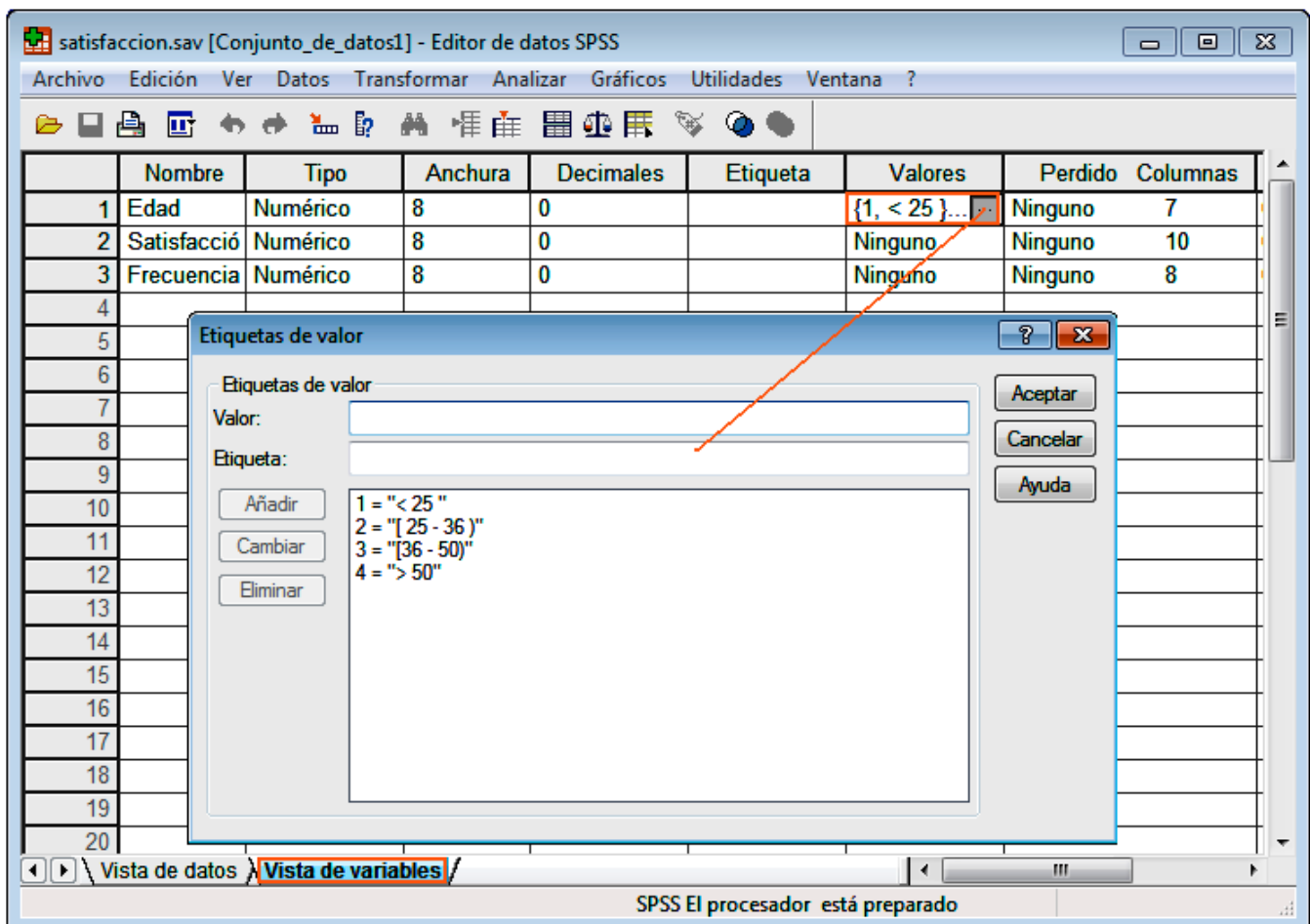
H₀: La asociación entre las variables es nula (son independientes)

Estadísticos Ordinales
Variables Cuantitativas

Gamma de Goodman y Kruskal
D de Somers
Tau-B de Kendall
Tau-C de Kendall
Riesgo relativo

Análogos a las medidas de asociación, aplicables a las variables que se computan en función de acuerdos-desacuerdos o concordancias-discrepancias

Pruebas de Concordancia { **Índice de Concordancia**
Coeficiente Kappa de Cohen



*satisfaccion.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver **Datos** Transformar Analizar Gráficos Utilidades Ventana ?

40 : Visible: 3 de 3 var

	Edad	Satisfacción	Frecuencia	var	var	var	var	var	var
1	1	0	10						
2	1	1	10						
3	1	2	20						
4	1	3	40						
5	1	4	70						
6	2	0	20						
7	2	1	10						
8	2	2	15						
9	2	3	20						
10	2	4	30						
11	3	0	60						
12	3	1	50						
13	3	2	30						
14	3	3	10						
15	3	4	5						

Ponderar casos

Edad
Satisfacción

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

*satisfaccion.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar **Analizar** Gráficos Utilidades Ventana ?

40 : Visible: 3 de 3 var

	Edad	Satisfacción	Frecuencia	var
1	1	0	10	
2	1	1	10	
3	1	2	20	
4	1	3	40	
5	1	4	70	
6	2	0	20	
7	2	1	10	
8	2	2	15	
9	2	3	20	
10	2	4	30	
11	3	0	60	
12	3	1	50	
13	3	2	30	
14	3	3	10	
15	3	4	5	
16				
17				
18				
19				
20				

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...
Razón...
Gráficos P-P...
Gráficos Q-Q...

Vista de datos / Vista de variables /

Tablas de contingencia

SPSS El procesador está preparado

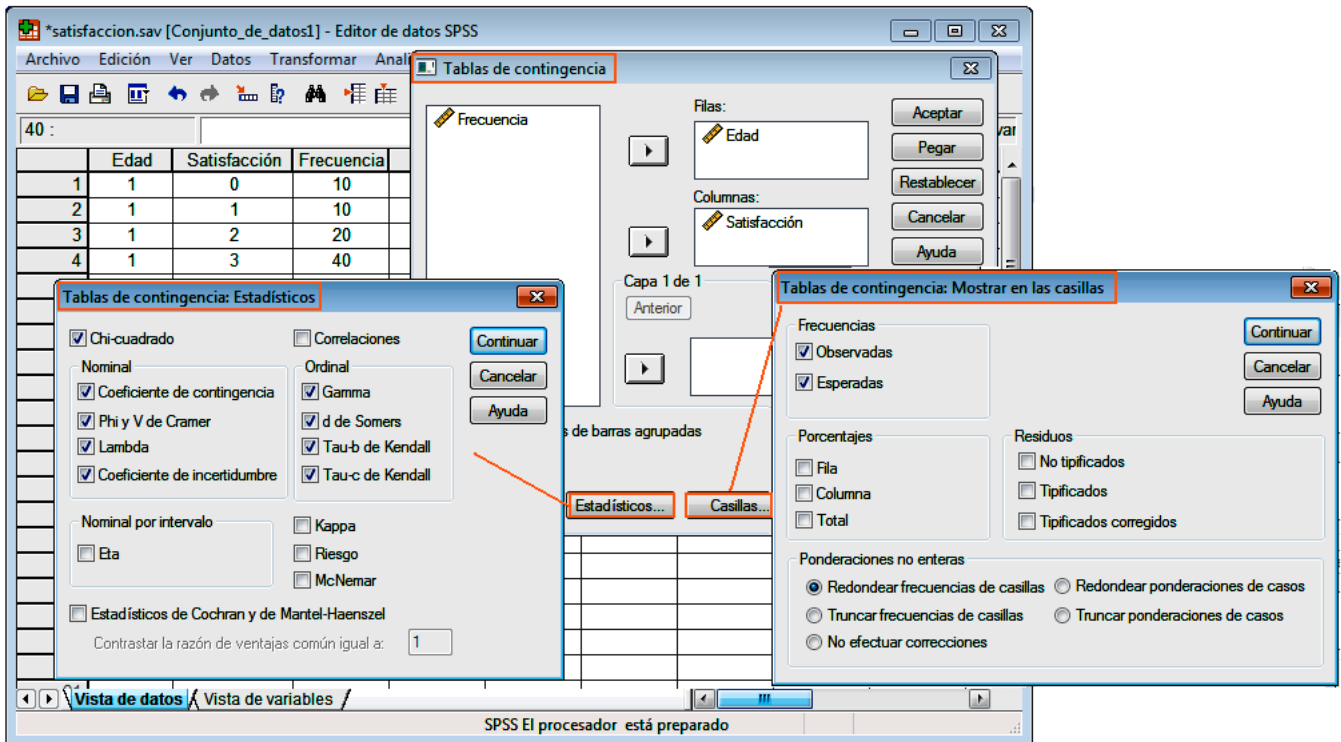


Tabla de contingencia Edad * Satisfacción

			Satisfacción					Total
			0	1	2	3	4	
Edad < 25	Recuento		10	10	20	40	70	150
	Frecuencia esperada		33,75	26,25	24,38	26,25	39,38	150
[25 - 36)	Recuento		20	10	15	20	30	95
	Frecuencia esperada		21,38	16,63	15,44	16,63	24,94	95
[36 - 50)	Recuento		60	50	30	10	5	155
	Frecuencia esperada		34,88	27,13	25,19	27,13	40,69	155
Total	Recuento		90	70	65	70	105	400
	Frecuencia esperada		90	70	65	70	105	400

Todas las casillas presentan frecuencia esperada superior a 5, en estas condiciones los resultados del contraste Chi-cuadrado son fiables. En la práctica se admite sólo el 20% de las frecuencias esperadas inferior a 5. En otro caso, se agrupan clases contiguas, recodificando la variable satisfacción asignando un único valor hasta obtener frecuencias esperadas superiores a 5.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	143,458 ^a	8	,000
Razón de verosimilitudes	163,334	8	,000
Asociación lineal por lineal	128,637	1	,000
N de casos válidos	400		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 15,44.

El valor de Chi-cuadrado de Pearson es 143,458 y la Razón de verosimilitud 163,334.

Estos valores presentan un p-valor (Sig. aproximada) de $0,000 < 0,05$, rechazando la hipótesis nula de independencia entre la edad y la satisfacción en el trabajo.

En consecuencia, la edad influye significativamente en la satisfacción en el trabajo.

Medidas direccionales

			Valor	Error típ. asint. ^a	T aproximada ^b	Sig. aproximada
Nominal por nominal	Lambda	Simétrica	,278	,023	12,511	,000
		Edad dependiente	,388	,036	9,386	,000
		Satisfacción dependiente	,186	,025	7,257	,000
	Tau de Goodman y Kruskal	Edad dependiente	,208	,027		,000 ^c
		Satisfacción dependiente	,100	,013		,000 ^c
	Coeficiente de incertidumbre	Simétrica	,153	,020	7,713	,000 ^d
		Edad dependiente	,190	,025	7,713	,000 ^d
		Satisfacción dependiente	,128	,017	7,713	,000 ^d
	Ordinal por ordinal	d de Somers	Simétrica	-,474	,031	-15,684
Edad dependiente			-,432	,028	-15,684	,000
Satisfacción dependiente			-,525	,035	-15,684	,000

a. Asumiendo la hipótesis alternativa.

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

c. Basado en la aproximación chi-cuadrado.

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

Tomando la variable Edad como dependiente el valor de Lambda es 0,388 con un error típico asintótico de 0,023, lo que indica que el conocimiento de la variable Satisfacción en el trabajo permite reducir la incertidumbre de la variable Edad en un 38,8%. Esta estimación de Lambda es significativa para niveles de significación superiores a 0,000.

Tomando la variable Satisfacción en el trabajo como dependiente el valor de Lambda es 0,186 con un error típico asintótico de 0,025. Así, pues, el conocimiento de los valores de la Edad permite reducir la incertidumbre en la predicción del comportamiento de la variable Satisfacción en el Trabajo en un 18,6%, estimación significativa para niveles de significación superiores a 0,000.

El estadístico Tau de Goodman y Kruskal con la variable Edad como dependiente toma el valor 0,208 con un error típico asintótico de 0,027, confirmando la conclusión anterior.

El Coeficiente de Incertidumbre con la variable Edad como dependiente, con un valor de 0,190 y un error típico asintótico de 0,025, reduce el grado de incertidumbre en el pronóstico de los valores de la Satisfacción en el trabajo en un 19% para niveles de significación superiores a 0,000.

El estadístico D de Somers oscila entre -1 y 1 . Tomando la variable Edad como dependiente tiene un valor de $-0,432$ con un error típico asintótico de 0,028 estableciendo que la Edad tiene un grado de dependencia inversa con la Satisfacción en el trabajo de un 43,2%, resultado con niveles de significación superiores a 0,000.

Elijiendo la variable Satisfacción en el trabajo como dependiente, con un valor de $-0,525$ y un error típico asintótico de 0,035, tiene un grado de dependencia inversa con la Edad de 52,5% a niveles de significación superiores a 0,000.

Medidas simétricas

		Valor	Error típ. asint. ^a	T aproximada ^b	Sig. aproximada
Nominal por nominal	Phi	,599			,000
	V de Cramer	,423			,000
	Coeficiente de contingencia	,514			,000
Ordinal por ordinal	Tau-b de Kendall	-,477	,031	-15,684	,000
	Tau-c de Kendall	-,514	,033	-15,684	,000
	Gamma	-,626	,038	-15,684	,000
N de casos válidos		400			

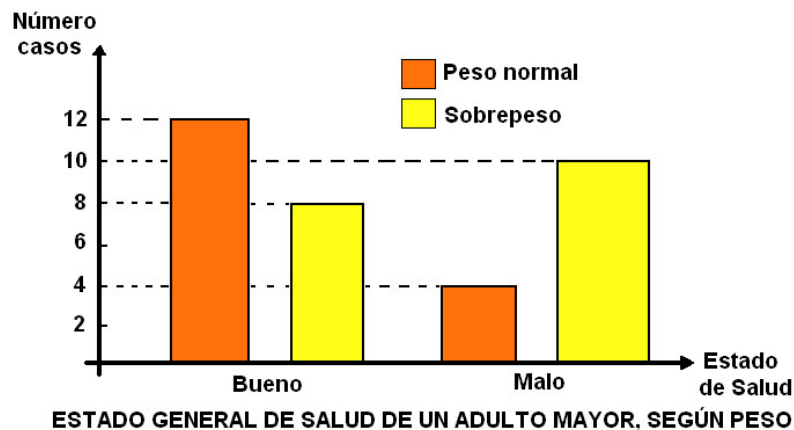
a. Asumiendo la hipótesis alternativa.

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

Los estadísticos Phi, V de Cramer, Coeficiente de contingencia, Tau-b de Kendall, Tau-C de Kendall y Gamma toman valores moderados y en todos los casos con un p-valor (Signatura aproximada) de $0,000 < 0,05$

En consecuencia se rechaza la hipótesis nula de independencia, pudiendo suponerse que existe una relación de dependencia entre la Edad y la Satisfacción en el trabajo, resultado confirmado con los valores de las medidas direccionales.

En el gráfico se presenta la evaluación del estado general de salud de una muestra de personas adultas mayores, según sea su peso normal o sobrepeso.



Analizar la existencia de una relación significativa entre el peso y el estado general de salud en el adulto mayor, con un nivel de significación del 5%,

Solución:

Se trata de dos variables dicotómicas con datos de frecuencia, pudiéndose aplicar una prueba de contraste de asociación con la Chi-cuadrado.

La hipótesis nula H_0 : El estado de salud y el peso son independientes

Llevando la información a una tabla de contingencia de 2x2

Estado de Salud	Peso		$n_{i\cdot}$
	Normal	Sobrepeso	
Buena	12 9,41	8 10,59	20 20
Mala	4 6,59	10 7,41	14 14
$n_{\cdot j}$	16	18	34

La frecuencia observada $n_{21} = 4$ es menor que lo aconsejable en cada celda (≥ 5), lo que podría hacer pensar en una inestabilidad del cálculo.

Como la frecuencia esperada $e_{21} = 6,59$, todas las celdas cumplen con el mínimo aconsejable de 5 en su valor esperado. En la práctica se acepta hasta un 20% de las celdas que no cumplen con el requisito de que la frecuencia esperada sea ≥ 5

Se calculan los valores de χ^2 correspondientes a las dos observaciones,

siendo la frecuencia esperada $e_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$

$$e_{11} = \frac{20 \cdot 16}{34} = 9,41 \quad e_{21} = \frac{14 \cdot 16}{34} = 6,59$$

$$e_{12} = \frac{20 \cdot 18}{34} = 10,59 \quad e_{22} = \frac{14 \cdot 18}{34} = 7,41$$

Estadístico de contraste:

$$\chi_{(2-1) \cdot (2-1)}^2 = \chi_1^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{n_{ij}^2}{e_{ij}} - n = \frac{12^2}{9,41} + \frac{8^2}{10,59} + \frac{4^2}{6,59} + \frac{10^2}{7,41} - 34 = 3,265$$

Estadístico teórico: $\chi_{0,05, 1}^2 = 3,841$

Como $\chi_1^2 = 3,265 < 3,841 = \chi_{0,05, 1}^2$ se acepta la hipótesis nula, concluyendo que el estado general de salud del adulto mayor no está asociado a su peso.

 Adviértase que como la muestra $n < 40$ se hace aconsejable el uso de la Chi-cuadrado con el factor de corrección de continuidad de Yates:

$$\text{Factor corrección} \begin{cases} n_{ij} < e_{ij} & \mapsto n_{ij} + 0,5 \\ n_{ij} > e_{ij} & \mapsto n_{ij} - 0,5 \end{cases}$$

Para una tabla de contingencia de 2×2 la corrección de Yates:

$$\chi_1^2 = \frac{n \left(\left| n_{11} \cdot n_{22} - n_{12} \cdot n_{21} \right| - \frac{n}{2} \right)^2}{n_{1\cdot} \cdot n_{2\cdot} \cdot n_{\cdot 1} \cdot n_{\cdot 2}}$$

La corrección no es válida cuando $\left| n_{11} \cdot n_{22} - n_{12} \cdot n_{21} \right| \leq \frac{n}{2}$

En general, la corrección de Yates se hace cuando el número de grados de libertad es 1.

$$\text{En este caso, } \chi_1^2 = \frac{34 \left(\left| 12 \times 10 - 8 \times 4 \right| - \frac{34}{2} \right)^2}{20 \times 14 \times 16 \times 18} = 2,125$$

Como $\chi_1^2 = 2,125 < 3,841 = \chi_{0,05,1}^2$ se acepta la hipótesis nula.

La validez del contraste también se puede hacer con el p-valor (α_p):

$$\alpha_p = P(\chi_{p,1}^2 > 2,125) = 0,273$$

0,90	α_p	0,10
0,0158	2,125	2,706

$$0,90 - 0,10 \longrightarrow 0,0158 - 2,706$$

$$\alpha_p - 0,10 \longrightarrow 2,125 - 2,706$$

$$(\alpha_p - 0,10) \times (0,0158 - 2,706) = (0,90 - 0,10) \times (2,125 - 2,706) \mapsto \alpha_p = 0,273$$

Al ser $\alpha_p = 0,273 > 0,05 = \alpha$ se acepta la hipótesis nula, afirmando que el estado general de salud del adulto mayor es independiente de su peso.

■ Test G de la razón de verosimilitud: $G = 2 \sum_{i=1}^2 \sum_{j=1}^2 n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right) =$

$$= 2 \left[12 \ln\left(\frac{12}{9,41}\right) + 8 \ln\left(\frac{8}{10,59}\right) + 4 \ln\left(\frac{4}{6,59}\right) + 10 \ln\left(\frac{10}{7,41}\right) \right] = 3,344$$

■ Coeficiente Phi: $\phi = \sqrt{\frac{\chi_c^2}{n}} = \sqrt{\frac{3,265}{34}} = 0,310$

El estadístico Phi mide el grado de asociación entre las variables.

■ Coeficiente V de Cramer:

$$V_{\text{Cramer}} = \sqrt{\frac{\chi_c^2}{n \cdot \min(k-1, m-1)}} = \sqrt{\frac{3,265}{34 \cdot \min(2-1, 2-1)}} = \sqrt{\frac{3,265}{34}} = 0,310$$

En tablas de contingencia 2x2 el estadístico Phi y V de Cramer tienen el mismo valor.

- Gamma de Goodman y Kruskal: $\gamma = \frac{C - D}{C + D} = \frac{120 - 32}{120 + 32} = 0,579$

Estado de Salud	Peso		$n_{i\cdot}$
	Normal	Sobrepeso	
Bueno	12	8	20
Malo	4	10	14
$n_{\cdot j}$	16	18	34

Pares Concordantes: $C = 12[10] = 120$

Pares Discordantes: $D = 8[4] = 32$

Parejas empatadas en X: $T_x = \sum_{i=1}^2 \frac{n_{i\cdot} (n_{i\cdot} - 1)}{2} = \frac{1}{2} [20 \cdot 19 + 14 \cdot 13] = 281$

Parejas empatadas en Y: $T_y = \sum_{j=1}^2 \frac{n_{\cdot j} (n_{\cdot j} - 1)}{2} = \frac{1}{2} [16 \cdot 15 + 18 \cdot 17] = 273$

- Tau-C de Kendall: $\tau_c = \frac{2 \cdot \min(k, m) \cdot (C - D)}{\min(k - 1, m - 1) \cdot n^2} = \frac{2 \cdot 2 \cdot (120 - 32)}{34^2} = 0,304$

- Tau-B de Kendall: $\tau_B = \frac{C - D}{\sqrt{\left(\frac{n(n-1)}{2} - T_x\right) \left(\frac{n(n-1)}{2} - T_y\right)}}$

$$\tau_B = \frac{120 - 32}{\sqrt{\left(\frac{34 \times 33}{2} - 281\right) \left(\frac{34 \times 33}{2} - 273\right)}} = 0,310$$

- Lambda de Goodman y Kruskal: $(X, Y) \equiv (\text{Estado Salud}, \text{Peso})$

$$\lambda_{yx} = \frac{\sum m_Y - M_Y}{n - M_Y} = \frac{22 - 20}{34 - 20} = 0,143 \quad \begin{cases} M_Y \equiv 20 \\ \sum m_Y \equiv 12 + 10 = 22 \\ n \equiv 34 \end{cases}$$

$M_Y \equiv$ Frecuencia modal global $\sum m_Y \equiv$ Suma de frecuencias modales

$$\lambda_{xy} = \frac{\sum m_x - M_x}{n - M_x} = \frac{22 - 18}{34 - 18} = 0,250 \quad \begin{cases} M_x \equiv 18 \\ \sum m_x \equiv 12 + 10 = 22 \\ n \equiv 34 \end{cases}$$

■ Tau de Goodman y Kruskal:

Peso dependiente: $\tau_{yx} = \frac{E_1 - E_2}{E_1} = \frac{16,47 - 14,89}{16,47} = 0,096$

$$E_1 = \sum_{i=1}^2 \left[\frac{(n - n_{i\cdot}) n_{i\cdot}}{n} \right] = \frac{(34 - 20)20}{34} + \frac{(34 - 14)14}{34} = 16,47$$

$$E_2 = \sum_{j=1}^2 \sum_{i=1}^2 \left[\frac{(n_{\cdot j} - n_{ij}) n_{ij}}{n_{\cdot j}} \right] =$$

$$= \frac{(16 - 12)12}{16} + \frac{(16 - 4)4}{16} + \frac{(18 - 8)8}{18} + \frac{(18 - 10)10}{18} = 14,89$$

Estado Salud dependiente: $\tau_{yx} = \frac{E_1 - E_2}{E_1} = \frac{16,94 - 15,31}{16,94} = 0,096$

$$E_1 = \sum_{j=1}^2 \left[\frac{(n - n_{\cdot j}) n_{\cdot j}}{n} \right] = \frac{(34 - 16)16}{34} + \frac{(34 - 18)18}{34} = 16,94$$

$$E_2 = \sum_{j=1}^2 \sum_{i=1}^2 \left[\frac{(n_{i\cdot} - n_{ij}) n_{ij}}{n_{i\cdot}} \right] =$$

$$= \frac{(20 - 12)12}{20} + \frac{(20 - 8)8}{20} + \frac{(14 - 4)4}{14} + \frac{(14 - 10)10}{14} = 15,31$$

■ Coeficiente de Incertidumbre

$$I(X) = \sum_{i=1}^2 \frac{n_{i\cdot}}{n} \ln \left(\frac{n_{i\cdot}}{n} \right) = \frac{20}{34} \ln \left(\frac{20}{34} \right) + \frac{14}{34} \ln \left(\frac{14}{34} \right) = -0,677$$

$$I(Y) = \sum_{j=1}^2 \frac{n_{\cdot j}}{n} \ln \left(\frac{n_{\cdot j}}{n} \right) = \frac{16}{34} \ln \left(\frac{16}{34} \right) + \frac{18}{34} \ln \left(\frac{18}{34} \right) = -0,691$$

$$I(XY) = \sum_{i=1}^2 \sum_{j=1}^2 \frac{n_{ij}}{n} \ln\left(\frac{n_{ij}}{n}\right) = \frac{12}{34} \ln\left(\frac{12}{34}\right) + \frac{8}{34} \ln\left(\frac{8}{34}\right) + \frac{4}{34} \ln\left(\frac{4}{34}\right) + \frac{10}{34} \ln\left(\frac{10}{34}\right) = -1,319$$

Coefficiente simétrico:

$$I = \frac{2 [I(X) + I(Y) - I(XY)]}{I(X) + I(Y)} = \frac{2 [-0,677 - 0,691 + 1,319]}{-0,677 - 0,691} = 0,072$$

Estado de salud como variable dependiente:

$$I_{X/Y} = \frac{I(X) + I(Y) - I(XY)}{I(X)} = \frac{-0,677 - 0,691 + 1,319}{-0,677} = 0,073$$

Peso como variable dependiente:

$$I_{Y/X} = \frac{I(X) + I(Y) - I(XY)}{I(Y)} = \frac{-0,677 - 0,691 + 1,319}{-0,691} = 0,071$$

■ El coeficiente o índice de Kappa κ es una medida de concordancia propuesta por Cohen en 1960, se basa en comparar la concordancia observada en un conjunto de datos, respecto a lo que podría ocurrir por pura casualidad. Se puede calcular en tablas de cualquier dimensión, en el caso de tablas de 2x2 tiene algunas peculiaridades.

X	Y		
	y_1	y_2	
x_1	n_{11}	n_{12}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	n

$$\text{Índice de Kappa: } \kappa = \frac{p_0 - p_e}{1 - p_e}$$

$$p_0 = \frac{1}{n} \sum_i n_{ii} \quad p_e = \frac{1}{n^2} \sum_i n_{i\cdot} \times n_{\cdot i}$$

Donde p_0 es la proporción de concordancia observada y p_e es la proporción de concordancia esperada por azar.

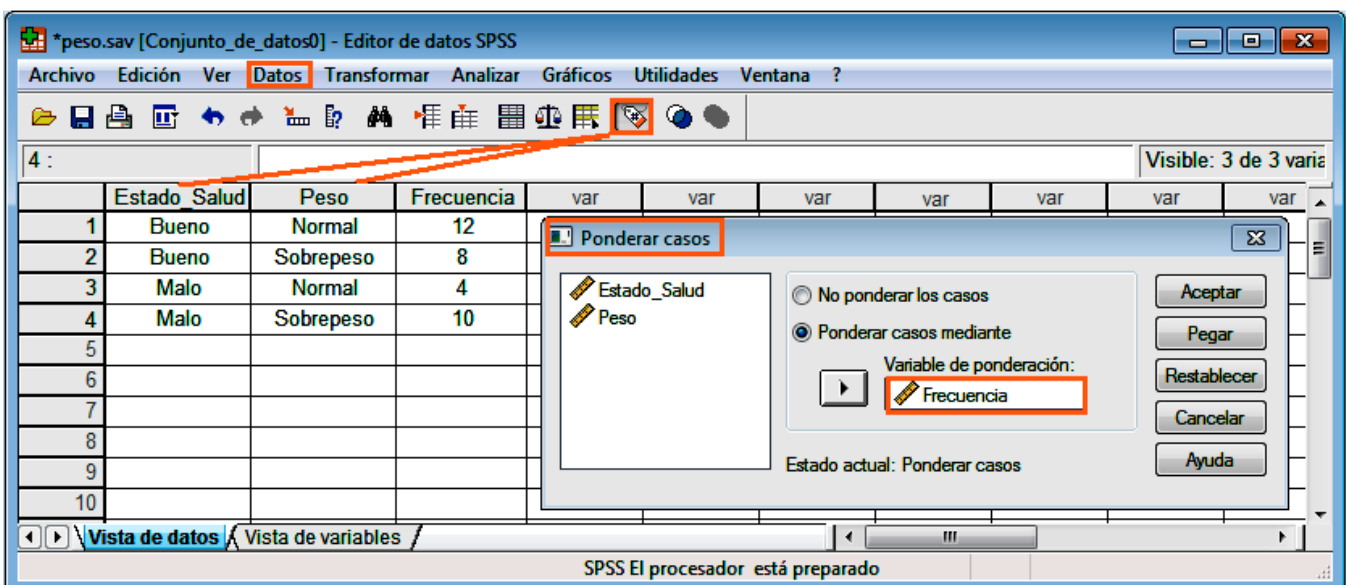
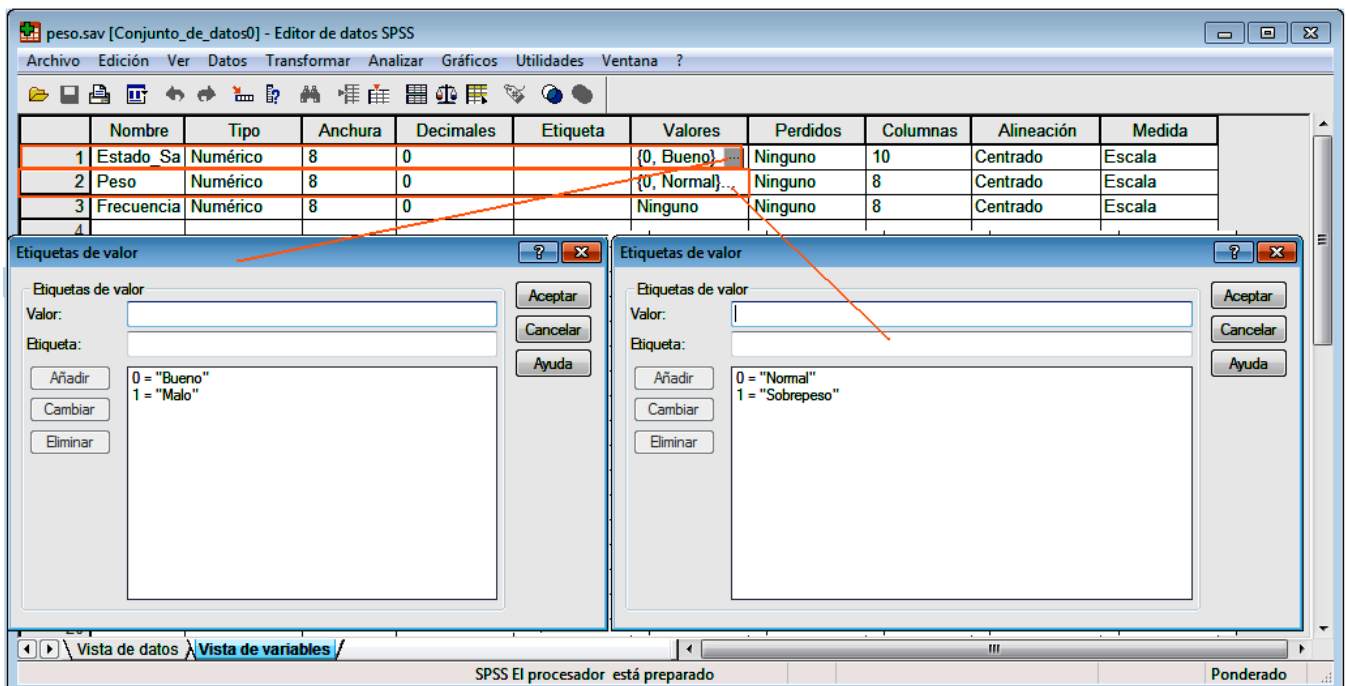
Cuando $\kappa = 1$ se da la máxima concordancia posible. El valor $\kappa = 0$ indica que la concordancia observada es precisamente la que se espera por pura casualidad.

$$p_o = \frac{1}{n} \sum_i n_{ii} = \frac{12 + 10}{34} = 0,647$$

$$p_e = \frac{1}{n^2} \sum_i n_{i.} \times n_{.i} = \frac{20 \times 16 + 14 \times 18}{34^2} = 0,495$$

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0,647 - 0,495}{1 - 0,495} = 0,301$$

En el caso de más de dos evaluadores, clasificaciones, métodos, etc., Joseph L. Fleiss generalizó el método de Cohen, dando lugar a la Kappa de Fleiss.



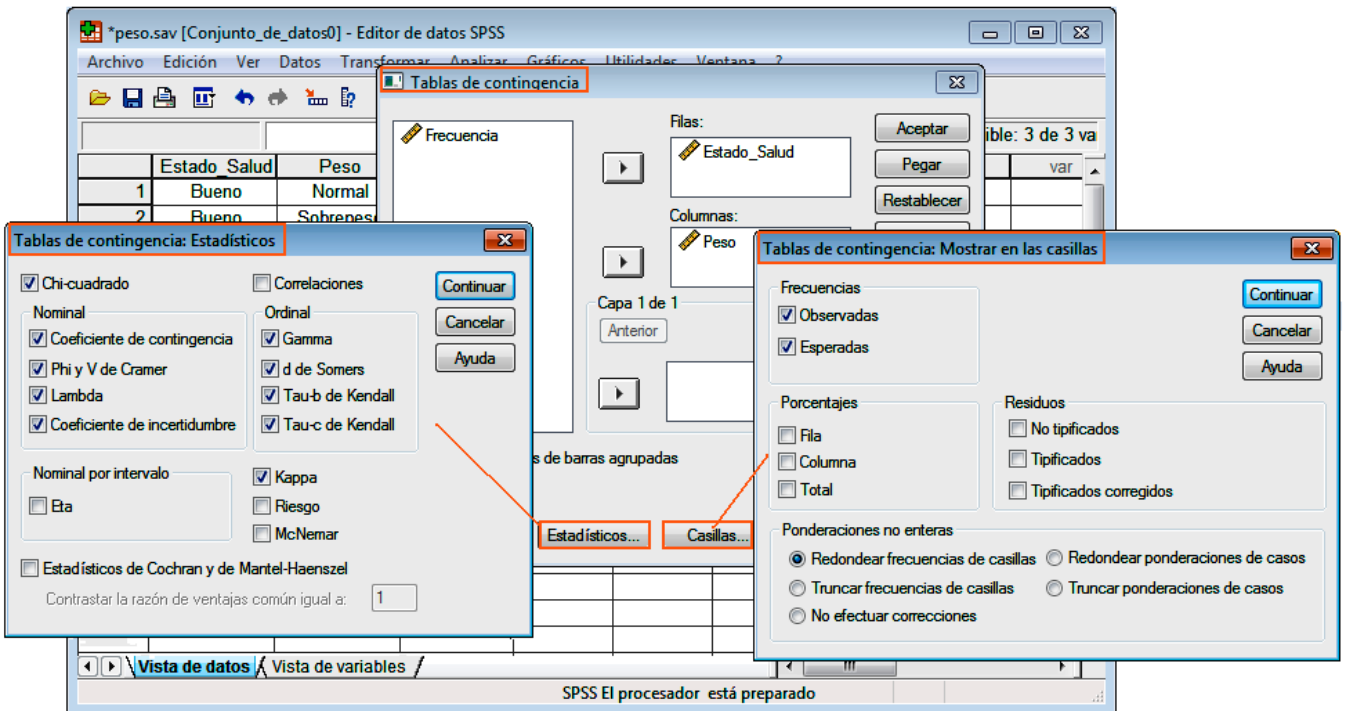
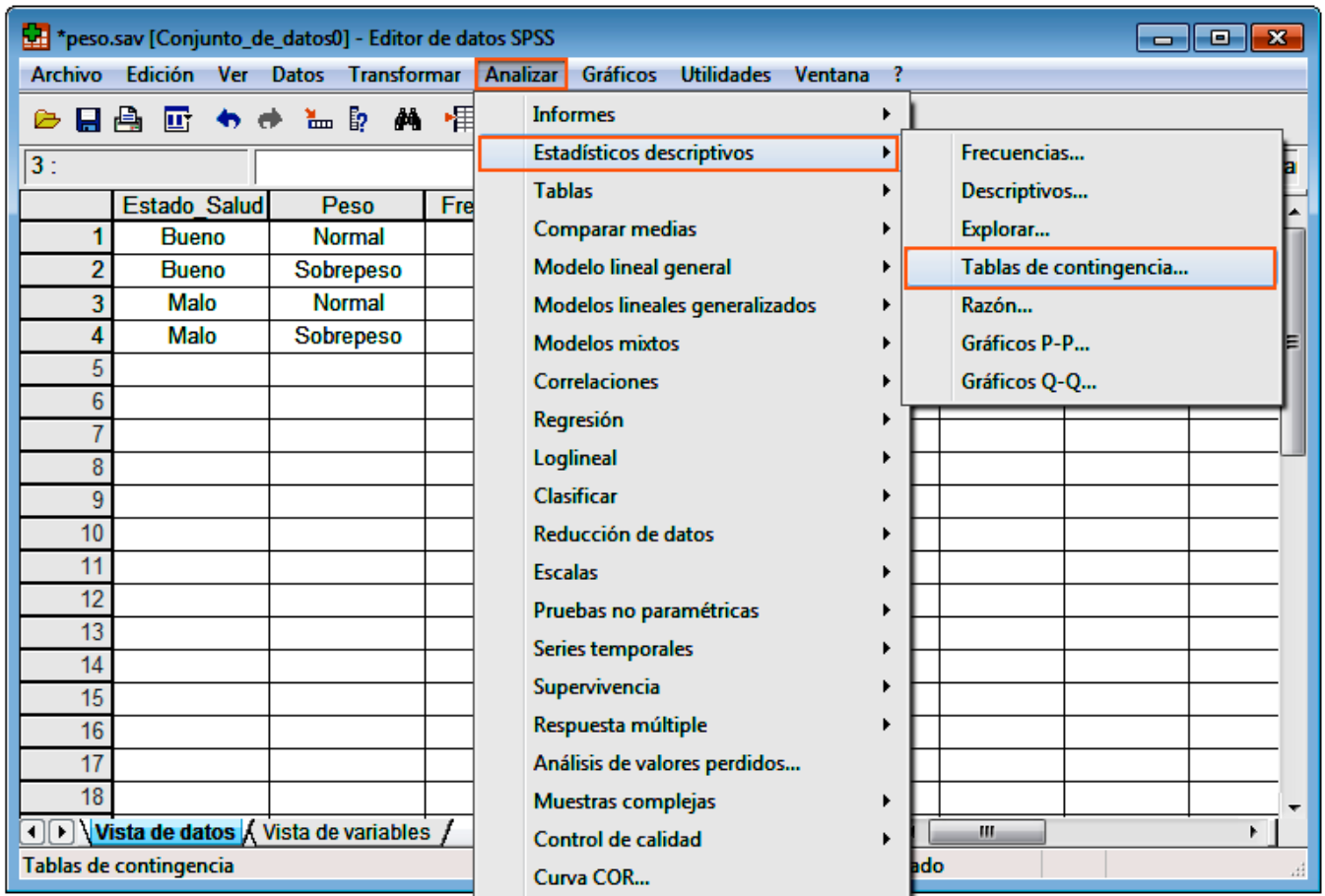


Tabla de contingencia Estado_Salud * Peso

			Peso		Total
			Normal	Sobrepeso	
Estado_Salud	Bueno	Recuento	12	8	20
		Frecuencia esperada	9,41	10,59	20,0
	Malo	Recuento	4	10	14
		Frecuencia esperada	6,59	7,41	14,0
Total		Recuento	16	18	34
		Frecuencia esperada	16,0	18,0	34,0

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	3,265 ^b	1	,071		
Corrección por continuidad	2,125	1	,145		
Razón de verosimilitudes	3,344	1	,067		
Estadístico exacto de Fisher				,092	,072
Asociación lineal por lineal	3,169	1	,075		
N de casos válidos	34				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 6,59.

El valor de Chi-cuadrado de Pearson es 3,265 con un p-valor de 0,071 mayor que 0,05. Lo mismo sucede con la corrección por continuidad de Yates (2,125) y el estadístico G de la razón de verosimilitudes (3,344).

Por tanto, se acepta la hipótesis nula de que las dos variables son independientes, concluyendo que el estado general de salud del adulto mayor y su peso no están relacionados entre sí.

El resto de estadísticos basados en la Chi-cuadrado (Phi, V de Cramer y Coeficiente de contingencia) son significativos.

Medidas simétricas

		Valor	Error típ. asint. ^a	T aproximada ^b	Sig. aproximada
Nominal por nominal	Phi	,310			,071
	V de Cramer	,310			,071
	Coeficiente de contingencia	,296			,071
Ordinal por ordinal	Tau-b de Kendall	,310	,161	1,914	,056
	Tau-c de Kendall	,304	,159	1,914	,056
	Gamma	,579	,248	1,914	,056
Medida de acuerdo	Kappa	,301	,158	1,807	,071
N de casos válidos		34			

a. Asumiendo la hipótesis alternativa.

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

La relación entre las dos variables es moderada ya que la V de Cramer tiene un valor de 0,310. Generalmente, en investigación social se suele considerar una relación fuerte cuando la V de Cramer es mayor que 0,240. Esta norma no es fija y es preferible examinar detalladamente el resto de información de la tabla de medidas direccionales.

Los estadísticos para variables ordinales muestran una ligera relación de carácter positivo. Los estadísticos Tau-b y Tau-c de Kendall presentan valores indicando una dependencia moderada entre las variables, con un p-valor mayor que 0,05 haciendo posible su cálculo.

La Gamma vale 0,579 lo que indica cierto grado de relación positiva, reflejando que Gamma puede inflar la relación existente entre las dos variables (Estado de Salud y Peso) al no tener en cuenta el número de casos empatados.

El índice de Kappa presenta un valor de $0,301 < 0,40$ pudiendo interpretar la concordancia entre las variables no aceptable. Generalmente se considera un grado aceptable cuando su valor es mayor o igual que 0,40 y excelente para valores superiores a 0,75.

Medidas direccionales

			Valor	Error típ. asint. ^a	T aproximada ^b	Sig. aproximada
Nominal por nominal	Lambda	Simétrica	,200	,228	,825	,410
		Estado_Salud dependiente	,143	,281	,473	,636
		Peso dependiente	,250	,242	,905	,365
	Tau de Goodman y Kruskal	Estado_Salud dependiente	,096	,100		,075 ^c
		Peso dependiente	,096	,099		,075 ^c
	Coeficiente de incertidumbre	Simétrica	,072	,076	,942	,067 ^d
		Estado_Salud dependiente	,073	,077	,942	,067 ^d
		Peso dependiente	,071	,075	,942	,067 ^d
	Ordinal por ordinal	d de Somers	Simétrica	,310	,161	1,914
Estado_Salud dependiente			,306	,159	1,914	,056
Peso dependiente			,314	,163	1,914	,056

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.

En los estadísticos basados en el error se obtiene menor fuerza de relación.

El estadístico Lambda presenta valores pequeños indicando una pequeña dependencia entre las variables (Estado de Salud y Peso).

Según el estadístico Tau de Goodman y Kruskal conocer el Estado de Salud ayuda a reducir el error de la variable Peso en un 9,6%, idéntico resultado se obtiene al conocer el Peso a la hora de reducir el error de la variable Estado de Salud, un porcentaje pequeño. Siendo el p-valor de 0,075, mayor que 0,05, induce a hacer posible el cálculo.

El estadístico D de Somers, con la variable Estado de Salud como dependiente, tiene un valor de 0,306 con un error típico asintótico de 0,159 estableciendo que el Estado de Salud tiene un grado de dependencia con el Peso de un 30,6%.

📄 La tabla adjunta refleja un análisis de la obesidad en 14 sujetos. Con un nivel de significación de 0,05, se desea analizar si existen diferencias en la prevalencia de obesidad entre hombres y mujeres o si, por el contrario, el porcentaje de obesos no varía entre sexos.

Sexo	Obesidad		Total
	Sí	No	
Mujeres	1	4	5
Hombres	7	2	9
Total	8	6	14

Solución:

El *test exacto de Fisher* permite analizar si dos variables dicotómicas están asociadas cuando la muestra a estudiar es demasiado pequeña y no cumple las condiciones necesarias para que la aplicación del test de la Chi-cuadrado sea idónea.

Las condiciones necesarias para aplicar el test de la Chi-cuadrado exigen que al menos el 80% de los valores esperados de las celdas sean mayores que 5. De este modo, en una tabla de contingencia de 2 x 2 será necesario que todas las celdas verifiquen esta condición, si bien en la práctica suele permitirse que una de ellas tenga frecuencias esperadas ligeramente por debajo de 5.

Si las dos variables que se están analizando son dicotómicas, y la frecuencia esperada es menor que 5 en más de una celda, no resulta adecuado aplicar el test de la χ^2 , aunque sí el test exacto de Fisher.

El test exacto de Fisher se basa en evaluar la probabilidad asociada a cada una de las tablas 2 x 2 que se pueden formar manteniendo los mismos totales de filas y columnas que los de la tabla observada.

Cada uno de estas probabilidades se obtiene bajo la hipótesis de independencia de las dos variables que se están analizando.

Probabilidad asociada a los datos que han sido observados:

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

La fórmula general de la probabilidad descrita deberá calcularse para todas las tablas de contingencia que puedan formarse con los mismos totales de filas y columnas de la tabla observada.

El valor de la p asociado al test exacto de Fisher puede calcularse sumando las probabilidades de las tablas que resulten menores o iguales a la probabilidad de la tabla que ha sido observada.

El contraste bilateral asume que la hipótesis alternativa establezca la dependencia entre las variables dicotómicas, pero sin especificar de antemano en qué sentido se producen dichas diferencias.

Hipótesis nula H_0 : El sexo y ser obeso son independientes

Sexo	Obesidad		Total
	Sí	No	
Mujeres	1 (a)	4 (b)	5 (a+b)
Hombres	7 (c)	2 (d)	9 (c+d)
Total	8 (a+c)	6 (b+d)	14 (n)

$$p = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{n! a! b! c! d!} = \frac{5! 9! 8! 6!}{14! 1! 4! 7! 2!} = 0,0599$$

Las siguientes tablas muestran todas las posibles combinaciones de frecuencias que se pueden obtener con los mismos totales de filas y columnas:

Sexo	Obesidad		Total
	Sí	No	
Mujeres	4 (a)	1 (b)	5 (a+b)
Hombres	4 (c)	5 (d)	9 (c+d)
Total	8 (a+c)	6 (b+d)	14 (n)

$$p = 0,2098$$

$$p = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{n! a! b! c! d!} = \frac{5! 9! 8! 6!}{14! 4! 1! 4! 5!} = 0,2098$$

Sexo	Obesidad		Total
	Sí	No	
Mujeres	2 (a)	3 (b)	5 (a+b)
Hombres	6 (c)	3 (d)	9 (c+d)
Total	8 (a+c)	6 (b+d)	14 (n)

$$p = 0,2797$$

Sexo	Obesidad		Total
	Sí	No	
Mujeres	3 (a)	2 (b)	5 (a+b)
Hombres	5 (c)	4 (d)	9 (c+d)
Total	8 (a+c)	6 (b+d)	14 (n)

$$p = 0,4196$$

$$p = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{n! a! b! c! d!} = \frac{5! 9! 8! 6!}{14! 3! 2! 5! 4!} = 0,4196$$

Sexo	Obesidad		Total
	Sí	No	
Mujeres	0 (a)	5 (b)	5 (a+b)
Hombres	8 (c)	1 (d)	9 (c+d)
Total	8 (a+c)	6 (b+d)	14 (n)

$$p = 0,0030$$

Sexo	Obesidad		Total
	Sí	No	
Mujeres	5 (a)	0 (b)	5 (a+b)
Hombres	3 (c)	6 (d)	9 (c+d)
Total	8 (a+c)	6 (b+d)	14 (n)

$$p = 0,0280$$

Sumando las probabilidades de las tablas que son menores o iguales a la probabilidad de la tabla observada ($p = 0,0599$) se tiene:

$$p = 0,0599 + 0,0030 + 0,0280 = 0,0909$$

Siendo $p - \text{valor} = 0,0909 > 0,05$ se acepta la hipótesis nula, concluyendo que el sexo y el hecho de ser obeso son independientes, es decir, no existe asociación entre las variables en estudio, con un nivel de significación $\alpha = 0,05$

Otro método de calcular el p-valor consiste en sumar las probabilidades asociadas a aquellas tablas que sean más favorables a la hipótesis

alternativa de los datos observados. La tabla extrema de los datos observados es la que no se observa ninguna mujer obesa, $p = 0,0030$

$$p = 0,0599 + 0,0030 = 0,0629$$

SPSS para el cómputo del test de Fisher, calcula el p-valor correspondiente a un contraste bilateral ($p = 0,0909$) y el p-valor asociado a un contraste unilateral ($p = 0,0629$).

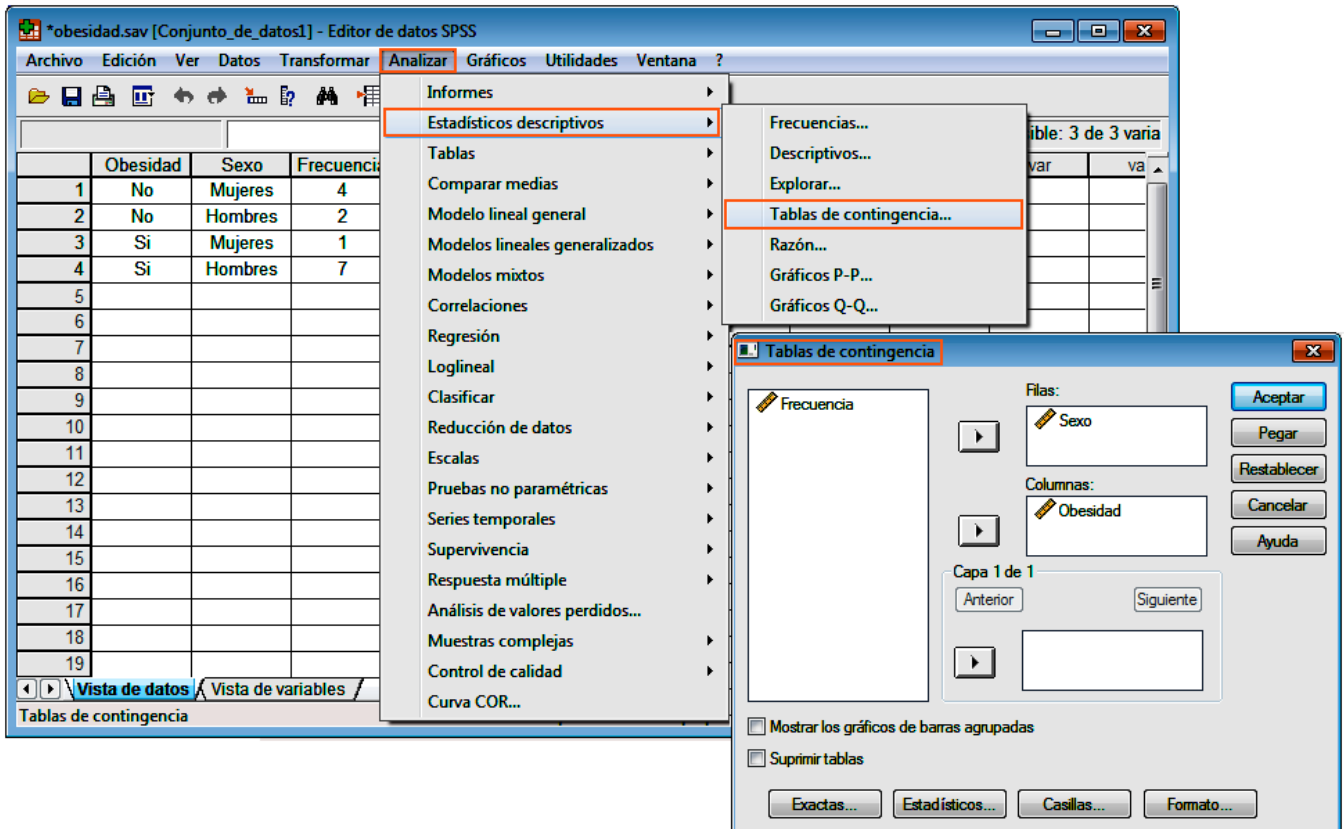
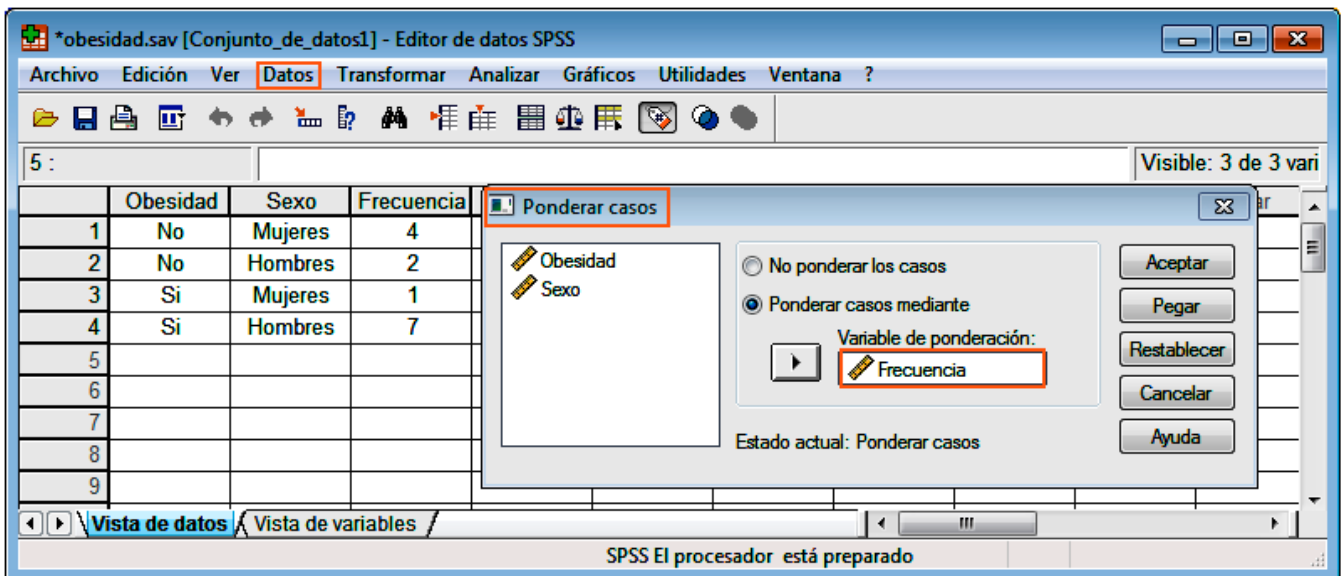


Tabla de contingencia Sexo * Obesidad

			Obesidad		Total
			No	Si	
Sexo	Mujeres	Recuento	4	1	5
		Frecuencia esperada	2,1	2,9	5,0
	Hombres	Recuento	2	7	9
		Frecuencia esperada	3,9	5,1	9,0
Total		Recuento	6	8	14
		Frecuencia esperada	6,0	8,0	14,0

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	4,381 ^b	1	,036		
Corrección por continuidad	2,340	1	,126		
Razón de verosimilitudes	4,583	1	,032		
Estadístico exacto de Fisher				,0909	,0629
Asociación lineal por lineal	4,069	1	,044		
N de casos válidos	14				

a. Calculado sólo para una tabla de 2x2.

b. 3 casillas (75,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 2,14.

p – valor (Signatura asintótica bilateral) = 0,0909 > 0,05, por tanto se acepta la hipótesis nula, concluyendo que el sexo y el hecho de ser obeso son independientes, es decir, no existe asociación entre las variables en estudio, con un nivel de significación $\alpha = 0,05$

📄 Para analizar la repercusión que tienen los debates televisivos en la intención de voto, un equipo de investigación recogió datos entre 240 individuos antes y después del debate, resultando la siguiente tabla:

Antes del debate (candidatos)	Después del debate (candidatos)		Total
	A	B	
A	46	50	96
B	85	59	144
Total	131	109	240

Se desea saber si el debate televisivo cambió la intención de voto, con un nivel de significación del 5%.

Solución:

Se trata de una muestra pareada en una situación antes-después, con lo que es idóneo un contraste estadístico Chi-cuadrado de McNemar.

Antes del debate (candidatos)	Después del debate (candidatos)		Total
	A	B	
A	46 (a)	50 (b)	96 (a+b)
B	85 (c)	59 (d)	144 (c+d)
Total	131 (a+c)	109 (b+d)	240 (n)

Hipótesis nula

H_0 : La intención de voto es la misma antes y después del debate

En esta prueba para la significación de cambios solo interesa conocer las celdas que presentan cambios (celdas b y c) y siendo (b + c) el número de personas que cambiaron, de acuerdo con la hipótesis nula planteada se espera que $\left(\frac{b+c}{2}\right)$ casos cambien en una dirección y $\left(\frac{b+c}{2}\right)$ casos a otra dirección.

- Estadístico de contraste sí $b + c < 20$

Se acepta H_0 sí $\chi_{McNemar}^2 = b < \chi_{\alpha/2, 1}^2$

- Estadístico de contraste si $b + c \geq 20$: $\chi_{McNemar}^2 = \chi_1^2 = \frac{[|b - c| - 1]^2}{b + c}$

La aproximación muestral a la distribución Chi-cuadrado llega a ser muy buena si se realiza una corrección por continuidad, considerando que se utiliza una distribución continua para aproximar una distribución discreta (binomial), por lo que se realiza la corrección de Yates.

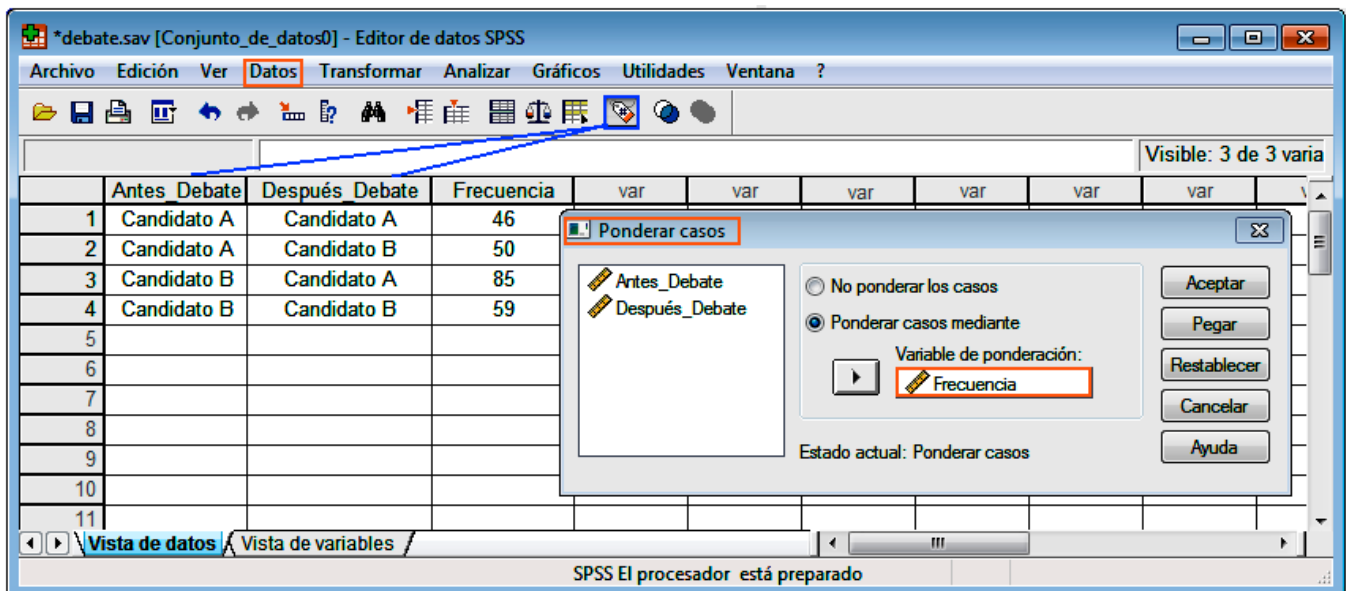
Se acepta H_0 si $\chi_{McNemar}^2 = \chi_1^2 = \frac{[|b - c| - 1]^2}{b + c} < \chi_{\alpha/2, 1}^2$

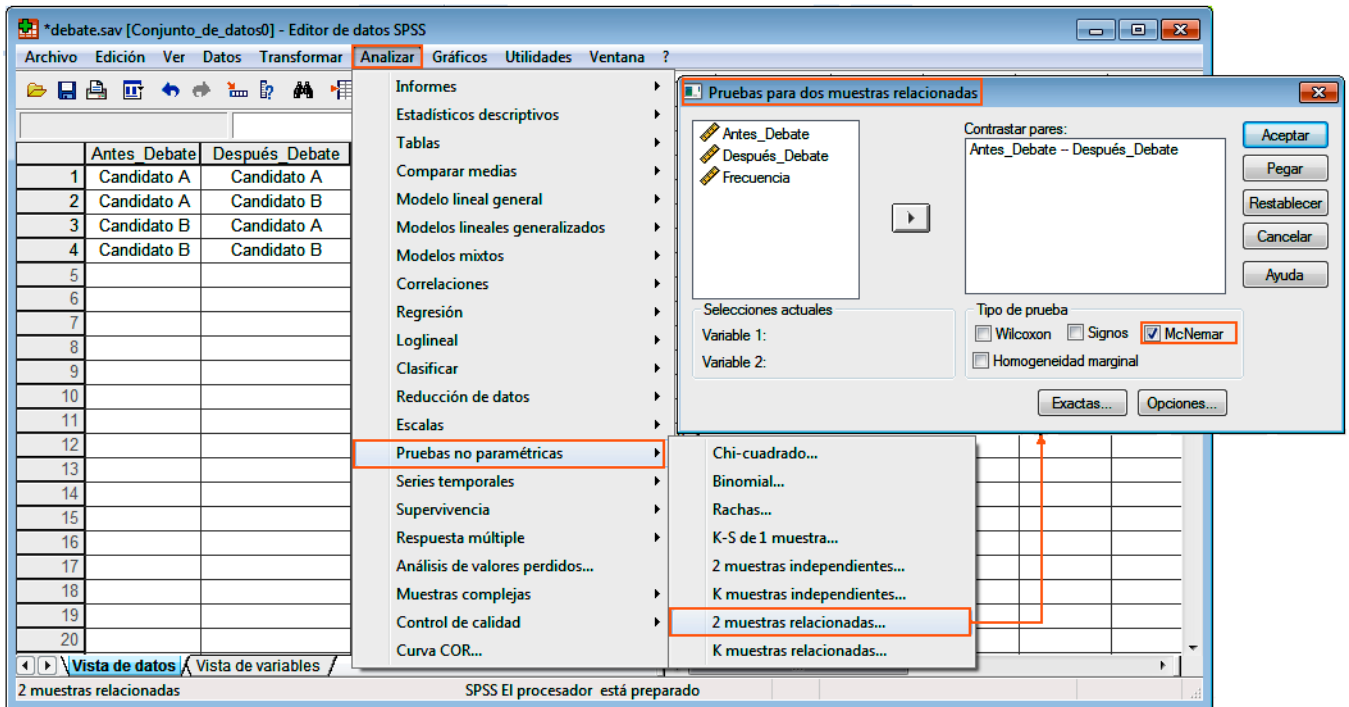
En este caso, $b + c = 50 + 85 = 135 > 20$

Estadístico muestral: $\chi_{McNemar}^2 = \frac{[|50 - 85| - 1]^2}{50 + 85} = 8,563$

Estadístico teórico: $\chi_{\alpha/2, 1}^2 = \chi_{0,025, 1}^2 = 5,024$

Como $\chi_{McNemar}^2 = 8,563 > 5,024 = \chi_{0,025, 1}^2$ se rechaza la hipótesis nula, concluyendo que la intención de voto cambió significativamente después del debate, con un nivel de significación del 5%.





Estadísticos de contraste^{a,b}

Antes_Debate	Después_Debate	
	0	1
0	46	50
1	85	59

	Antes_Debate y Después_Debate
N	240
Chi-cuadrado ^a	8,563
Sig. asintót.	,003

a. Corregido por continuidad
b. Prueba de McNemar

Como $p - \text{valor}(\text{Signatura asintótica}) = 0,003 < 0,05$ con lo que se rechaza la hipótesis nula, en consecuencia la intención de voto cambió significativamente después del debate, con un nivel de significación del 5%.

