Capítulo 12
Análisis de variables categóricas
El procedimiento *Tablas de contingencia*

En las ciencias sociales, de la salud y del comportamiento es bastante frecuente encontrarse con variables categóricas. El sexo, la raza, la clase social, el lugar de procedencia, la categoría laboral, participar o no en un programa de intervención, el tipo de tratamiento aplicado, los distintos departamentos de una empresa, padecer o no una enfermedad o un determinado síntoma, etc., son ejemplos de algunas variables categóricas con las que nos podemos encontrar. Son variables sobre las que únicamente es posible obtener una medida de tipo nominal (u ordinal, pero con muy pocos valores).

En este capítulo vamos a estudiar un procedimiento SPSS que permite describir este tipo de variables y detectar posibles pautas de asociación entre ellas.

Tablas de contingencia

Cuando se trabaja con variables categóricas, los datos suelen organizarse en tablas de doble entrada en las que cada entrada representa un criterio de clasificación (una variable categórica). Como resultado de esta clasificación, las frecuencias (el número o porcentaje de casos) aparecen organizadas en casillas que contienen información sobre la relación existente entre ambos criterios. A estas tablas de frecuencias se les llama *tablas de contingencia*.

La tabla 12.1 muestra un ejemplo de tabla de contingencia. En ella, 474 sujetos han sido ordenados con arreglo a dos criterios de clasificación: *sexo* y *salario* (se trata por tanto de una tabla *bidimensional*). Los números que aparecen en la tabla no son puntuaciones, sino frecuencias absolutas (número de casos): 19 varones tienen salarios de menos de 25.000 $; 86 mujeres tienen salarios comprendidos entre 25.000 y 50.000 $; etc.

Tabla 12.1. Tabla de contingencia de las variables *sexo* y *grupos de salario*.

Recuento

		Grupos de salario actual				
		Menos de 25.000 $	Entre 25.000 y 50.000 $	Entre 50.000 y 75.000 $	Más de 75.000 $	Total
Sexo	Hombre	19	174	48	17	258
	Mujer	124	86	6		216
Total		143	260	54	17	474

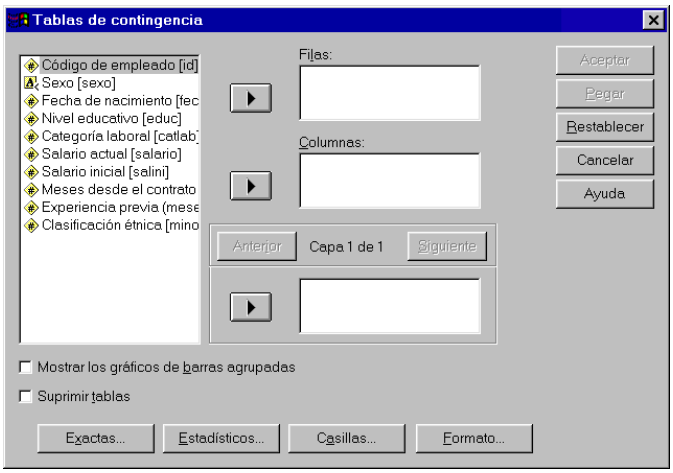
Por supuesto, en lugar de utilizar sólo dos criterios de clasificación para generar una tabla de contingencia *bidimensional*, también podríamos utilizar tres o más criterios, lo que nos llevaría

a obtener tablas *tridimensionales*, *cuatridimensionales*, etc. El procedimiento **Tablas de contingencia** del SPSS permite generar tablas con cualquier número de dimensiones. No obstante, los estadísticos que incluye sólo son útiles para analizar tablas *bidimensionales*. El análisis de tablas de contingencia con más de dos criterios de clasificación se aborda en otros procedimientos SPSS (por ejemplo, en el procedimiento *Modelos loglineales*).

Así pues, el procedimiento **Tablas de contingencia** permite obtener tablas de contingencia *bidimensionales*. Pero, además, incluye la posibilidad de añadir terceras variables (variables de segmentación) para definir subgrupos o capas y obtener así tablas multidimensionales. También incluye varios estadísticos y medidas de asociación que proporcionan la información necesaria para estudiar las posibles pautas de asociación existentes entre las variables que conforman una tabla de contingencia bidimensional. Para utilizar el procedimiento **Tablas de contingencia**:

- Seleccionar la opción **Estadísticos descriptivos** > **Tablas de contingencia** del menú **Analizar** para acceder al cuadro de diálogo *Tablas de contingencia* que muestra la figura 12.1.

Figura 12.1. Cuadro de diálogo *Tablas de contingencia*.



La lista de variables del archivo muestra todas las variables numéricas y de cadena corta del archivo de datos. Para obtener una tabla de contingencia:

- Trasladar una variable categórica a la lista **Filas,** otra a la lista **Columnas** y pulsar el botón **Aceptar**.

- **Mostrar los gráficos de barras agrupadas**. Activando esta opción, el *Visor de resultados* muestra un gráfico de barras con las categorías de la variable *fila* en el eje de abscisas y las categorías de la variable columna anidadas dentro de las categorías de la variable fila. Cada barra, por tanto, representa una casilla, y su altura viene dada por la frecuencia de la casilla.

☐ **Suprimir tablas**. Esta opción puede activarse si no se desea obtener ninguna tabla de contingencia. Esto tendría sentido si sólo estuviéramos interesados en obtener un gráfico de barras o alguno de los estadísticos o medidas de asociación disponibles en el procedimiento **Tablas de contingencia**.

Ejemplo (Tablas de contingencia)

Este ejemplo muestra cómo obtener una tabla de contingencia y un diagrama de barras agrupadas mediante el procedimiento **Tablas de contingencia**. Utilizaremos la variable *sexo* como variable *fila* y la variable *catlab* (categoría laboral) como variable *columna*:

- En el cuadro de diálogo *Tablas de contingencia* (ver figura 12.1), seleccionar la variable *sexo* y trasladarla a la lista **Filas**; seleccionar la variable *categoría laboral* y trasladarla a la lista **Columnas**.
- Marcar la opción **Mostrar los gráficos de barras agrupadas**.

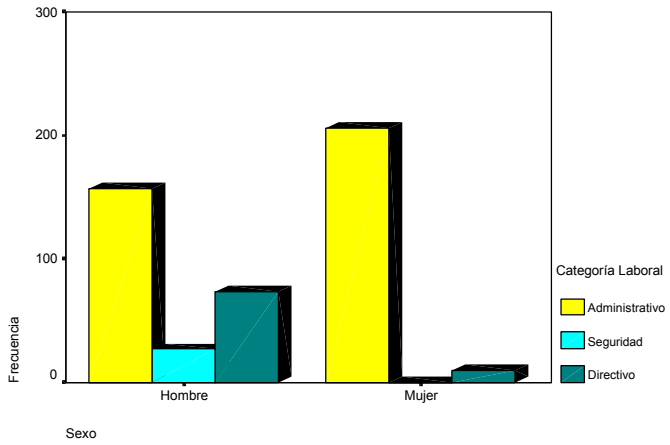
Aceptando estas elecciones el *Visor de resultados* genera la tabla 12.2 y la figura 12.2. La tabla 12.2 ofrece las frecuencias (número de casos) que resultan de cruzar cada categoría de la variable *sexo* con cada categoría de la variable *categoría laboral*. También ofrece las frecuencias (totales) correspondientes a cada variable individualmente considerada

Tabla 12.2. Tabla de contingencia resultante de cruzar las variables *sexo* y *categoría laboral*.

Recuento

		Categoría laboral			
		Administrativo	Seguridad	Directivo	Total
Sexo	Hombre	157	27	74	258
	Mujer	206		10	216
Total		363	27	84	474

Figura 12.2. Gráfico de barras agrupadas de las variables *sexo* y *categoría laboral*.



La figura 12.2 muestra el gráfico de barras agrupadas correspondiente a los datos de la tabla 12.2. Cada barra del gráfico se corresponde con una casilla de la tabla.

Podemos trasladar más de una variable tanto a la lista **filas** como a la lista **columnas**. En ese caso, cada variable *fila* se cruza con cada variable *columna* para formar una tabla de contingencia distinta. Seleccionando, por ejemplo, dos variables *fila* y tres variables *columna*, obtendríamos seis tablas de contingencia diferentes.

Tablas de contingencia segmentadas

El procedimiento **Tablas de contingencia** también permite cruzar variables categóricas teniendo en cuenta los niveles o categorías de una o más variables adicionales. Al cruzar, por ejemplo, las variables *sexo* y *categoría laboral*, podemos solicitar tablas separadas para cada nivel de *nivel de estudios*; en ese caso, la variable *nivel de estudios* actúa como variable de *segmentación*. Para obtener una tabla de contingencia segmentada:

- Trasladar la variable de segmentación a la lista situada en el recuadro **Capa 1 de 1** (ver figura 12.1).

Al seleccionar una variable de segmentación, el SPSS genera una tabla con tres dimensiones: la variable *fila*, la variable *columna* y la variable de *segmentación*. También es posible utilizar más de una variable de segmentación llevando más de una variable a la lista del recuadro **Capa 1 de 1**. Al hacer esto, el SPSS genera una tabla de contingencia separada para cada variable de segmentación seleccionada. Y si se seleccionan variables en distintas capas, la tabla de contingencia pasa a tener una nueva dimensión por cada capa adicional (ver siguiente ejemplo).

Ejemplo (Tablas de contingencia segmentadas)

Este ejemplo muestra cómo utilizar una variable de segmentación para obtener varias capas de una misma tabla de contingencia. Manteniendo las variables *sexo* y *catlab* como variables *fila* y *columna*, respectivamente:

- Seleccionar la variable *minoría* (clasificación étnica) y trasladarla a la lista del recuadro **Capa 1 de 1** (ver figura 12.1).

Al utilizar una variable de segmentación, el *Visor de resultados* genera la tabla 12.3.

Tabla 12.3. Tabla de contingencia de *sexo* por *categoría laboral*, segmentada por *clasificación étnica*.

Recuento

Clasificación étnica			Categoría Laboral			Total
			Administrativo	Seguridad	Directivo	
Blancos	Sexo	Hombre	110	14	70	194
		Mujer	166		10	176
	Total		276	14	80	370
No blancos	Sexo	Hombre	47	13	4	64
		Mujer	40			40
	Total		87	13	4	104

Con los botones **Siguiente** y **Anterior** del recuadro **Capa # de #** (figura 12.1) podemos obtener tablas de contingencia para los distintos niveles resultantes de combinar dos o más variables de segmentación. Para cruzar, por ejemplo, las variables *sexo* y *categoría laboral* y obtener una tabla separada para cada uno de los niveles resultantes de combinar las variables *minoría* (clasificación étnica) y *estudios* (nivel de estudios):

- Seleccionar la variable *minoría* como variable de segmentación en la *primera capa* y pulsar el botón **Siguiente**.
- Seleccionar la variable *estudios* como variable de segmentación en la *segunda capa*.
- Utilizar el botón **Anterior** para ver o cambiar la variable seleccionada en la capa previa.

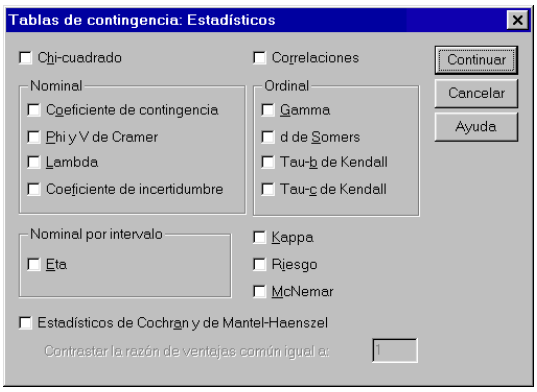
Al hacer esto, obtenemos una tabla de contingencia con cuatro dimensiones: *sexo*, *catlab*, *minoría* y *estudios*. Conforme vamos creando capas, los valores # del recuadro **Capa # de #** van indicando el número de la capa actual y el número total de capas definidas.

Estadísticos

El grado de relación existente entre dos variables categóricas no puede ser establecido simplemente observando las frecuencias de una tabla de contingencia. Incluso aunque la tabla recoja las frecuencias porcentuales en lugar de las absolutas (más tarde veremos que podemos obtener diferentes tipos de frecuencias porcentuales), la simple observación de las frecuencias no puede llevarnos a una conclusión definitiva (aunque sí pueda darnos alguna pista). Para determinar si dos variables se encuentran relacionadas debemos utilizar alguna medida de asociación, preferiblemente acompañada de su correspondiente prueba de significación. Para obtener medidas de asociación:

- Pulsar el botón **Estadísticos...** del cuadro de diálogo *Tablas de contingencia* (ver figura 12.1) para acceder al subcuadro de diálogo *Tablas de contingencia: Estadísticos* que muestra la figura 12.3.

Figura 12.3. Subcuadro de diálogo *Tablas de contingencia: Estadísticos*.



Este cuadro de diálogo contiene una amplia variedad de procedimientos estadísticos (medidas de asociación para variables nominales y ordinales, índices de acuerdo y de riesgo, etc.) diseñados para evaluar el grado de asociación existente entre dos variables categóricas en diferentes tipos de situaciones.

Chi cuadrado

La opción **Chi cuadrado** proporciona un estadístico (también conocido como X^2 o **ji-cuadrado**) propuesto por Pearson (1911) que permite contrastar la hipótesis de que los dos criterios de clasificación utilizados (las dos variables categóricas) son independientes. Para ello, compara las frecuencias *observadas* (las frecuencias de hecho obtenidas) con las frecuencias *esperadas* (las frecuencias que teóricamente deberíamos haber encontrado en cada casilla si los dos criterios de clasificación fueran independientes). Cuando dos criterios de clasificación son independientes, las frecuencias esperadas se estiman de la siguiente manera:

$$(\text{frecuencia esperada})_{ij} = \frac{(\text{total de la fila } i) \times (\text{total de la columna } j)}{\text{nº total de casos}}$$

(*i* se refiere a una fila cualquiera; *j* a una columna cualquiera; *ij* a una casilla cualquiera). Es decir, bajo la condición de independencia, la frecuencia esperada de una casilla concreta se obtiene dividiendo el producto de las frecuencias marginales correspondientes a esa casilla (su total de fila y su total de columna) por el número total de casos.

Obtenidas las frecuencias esperadas para cada casilla, el estadístico X^2 o *chi-cuadrado* de Pearson se obtiene de la siguiente manera:

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

(n_{ij} se refiere a las frecuencias observadas y m_{ij} a las esperadas). De la ecuación se desprende que el estadístico X^2 valdrá cero cuando las variables sean completamente independientes (pues las frecuencias observadas y las esperadas serán iguales), y que el valor del estadístico X^2 será tanto mayor cuanto mayor sea la discrepancia entre las frecuencias observadas y las esperadas (discrepancia que será tanto mayor cuanto mayor sea la relación entre las variables).

El estadístico X^2 sigue el modelo de distribución de probabilidad χ^2 con los grados de libertad resultantes de multiplicar el número de filas menos uno por el número de columnas menos uno ($gl = [J–1][K–1]$). Por tanto, podemos utilizar la distribución χ^2 para establecer el grado de compatibilidad existente entre el valor del estadístico X^2 y la hipótesis de independencia. Si los datos son compatibles con la hipótesis de independencia, la probabilidad asociada al estadístico X^2 será alta (mayor de 0,05). Si esa probabilidad es muy pequeña (menor que 0,05), consideraremos que los datos se muestran incompatibles con la hipótesis de independencia y concluiremos que las variables estudiadas están relacionadas.

Para que las probabilidades de la distribución χ^2 constituyan una buena aproximación a la distribución del estadístico X^2 conviene que se cumplan algunas condiciones; entre ellas, que

las frecuencias esperadas no sean demasiado pequeñas. Suele asumirse que, si existen frecuencias esperadas menores que 5, éstas no deben superar el 20 por ciento del total de frecuencias esperadas. La salida del SPSS muestra un mensaje indicando el valor de la frecuencia esperada más pequeña; si existe alguna casilla con frecuencia esperada menor que 5, la salida también muestra el porcentaje que éstas representan sobre el total de casillas de la tabla. En el caso de que ese porcentaje supere el 20 por ciento, el estadístico de Pearson debe ser interpretado con cautela.

Ejemplo (Tablas de contingencia > Estadísticos > Chi cuadrado)

Este ejemplo muestra cómo obtener e interpretar el estadístico *chi-cuadrado* de Pearson en una tabla de contingencia bidimensional.

- En el cuadro de diálogo *Tablas de contingencia* (ver figura 12.1), seleccionar las variables *sexo* y *categoría laboral* como variables *fila* y *columna*, respectivamente.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Tablas de contingencia: Estadísticos* (ver figura 12.3) y marcar la opción **Chi cuadrado**.

Aceptando estas elecciones, el *Visor de resultados* ofrece los estadísticos que muestra la tabla 12.4.

Tabla 12.4. Tabla de *estadísticos* del procedimiento *Tablas de contingencia*.

	Valor	gl	Sig. asint. (bilateral)
Chi-cuadrado de Pearson	79,277[a]	2	,000
Razón de verosimilitud	95,463	2	,000
N de casos válidos	474		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 12,30.

Vemos que el estadístico *chi-cuadrado* de Pearson toma un valor de 79,277, el cual, en la distribución χ^2 con 2 grados de libertad (*gl*), tiene asociada una probabilidad (*Sig. asint. = Significación asintótica*) de 0,000. Puesto que esta probabilidad (denominada *nivel crítico* o *nivel de significación observado*) es muy pequeña, decidimos rechazar la hipótesis de independencia y concluir que las variables *sexo* y *catlab* están relacionadas.

Además del estadístico *chi-cuadrado*, la tabla muestra otro estadístico denominado **razón de verosimilitud** (Fisher, 1924; Neyman y Pearson, 1928), que se obtiene mediante:

$$\textit{Razón de verosimilitud} = 2\sum_i \sum_j n_{ij} \log\left(\frac{n_{ij}}{m_{ij}}\right)$$

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

Cuando la tabla de contingencia se construye con dos variables dicotómicas (tablas 2×2), los resultados incluyen información adicional. Utilizando, por ejemplo, la variable *sexo* como variable *fila* y la variable *minoría* (clasificación étnica) como variable *columna* obtenemos la tabla 2×2 y los estadísticos que muestran las tablas 12.5.a y 12.5.b.

Tabla 12.5.a. Tabla de contingencia resultante de cruzar *sexo* y *clasificación étnica*.

Recuento

		Clasificación étnica		
		No	Sí	Total
Sexo	Hombre	194	64	258
	Mujer	176	40	216
Total		370	104	474

Tabla 12.5.b. Tabla de estadísticos.

	Valor	gl	Sig. asint. (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	2,714[b]	1	,099		
Corrección de continuidad [a]	2,359	1	,125		
Razón de verosimilitud	2,738	1	,098		
Estadístico exacto de Fisher				,119	,062
N de casos válidos	474				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 47,39.

Vemos que siguen estando presentes tanto el estadístico de Pearson como la razón de verosimilitud. Pero ahora hay dos líneas nuevas: la *corrección por continuidad de Yates* y el *estadístico exacto de Fisher*. También se muestra, en una nota a pie de tabla, el valor de la frecuencia esperada más pequeña, que en este ejemplo es 47,39.

La ***corrección por continuidad*** de Yates (1934) consiste en restar 0,5 puntos al valor absoluto de las diferencias $n_{ij} - m_{ij}$ del estadístico X^2 (antes de elevarlas al cuadrado). Algunos autores sugieren que, con muestras pequeñas, esta corrección permite que el estadístico X^2 se ajuste mejor a las probabilidades de la distribución χ^2, pero no existe un consenso generalizado sobre la utilización de esta corrección.

El ***estadístico exacto de Fisher*** (1935) ofrece, basándose en la distribución hipergeométrica y en la hipótesis de independencia, la probabilidad exacta de obtener las frecuencias de hecho obtenidas o cualquier otra combinación más alejada de la hipótesis de independencia.

Correlaciones

La opción **Correlaciones** (ver figura 12.3) contiene dos coeficientes de correlación: el de Pearson y el de Spearman (ambos se describen con detalle en el capítulo 16). *El coeficiente de correlación de Pearson* es una medida de asociación *lineal* especialmente apropiada para

estudiar la relación entre variables de intervalo o razón. El *coeficiente de correlación de Spearman* es también una medida de asociación *lineal*, pero para variables ordinales. Ambos coeficientes poseen escasa utilidad para estudiar las pautas de relación presentes en una tabla de contingencia típica, pues lo habitual es utilizar las tablas de contingencia para cruzar variables de tipo nominal o, a lo sumo, de tipo ordinal con sólo unos pocos niveles.

Datos nominales

El estadístico *chi-cuadrado* de Pearson permite contrastar la hipótesis de independencia en una tabla de contingencia, pero no nos dice nada sobre la *fuerza de la asociación* entre las variables estudiadas. Esto es debido a que su valor depende, no sólo del grado en que los datos se ajustan al modelo de independencia, sino del número de casos de que consta la muestra. Con tamaños muestrales muy grandes, diferencias relativamente pequeñas entre las frecuencias observadas y las esperadas pueden dar lugar a valores *chi-cuadrado* demasiado altos. Por esta razón, para estudiar el grado de relación existente entre dos variables se utilizan medidas de asociación que intentan cuantificar ese grado de relación eliminando el efecto del tamaño muestral.

Existen diversas medidas de asociación que, no sólo difieren en la forma de definir lo que es asociación perfecta e intermedia, sino en la forma en que cada una se ve afectada por factores tales como las distribuciones marginales. De hecho, una medida puede arrojar un valor bajo en una situación concreta, no porque las variables estudiadas no estén relacionadas, sino porque esa medida no sea sensible al tipo de relación presente en los datos. Para seleccionar una medida concreta, además de las características particulares de cada medida, hay que tener en cuenta cosas tales como el tipo de variables estudiadas y la hipótesis que interesa contrastar. En ningún caso está justificado obtener todas las medidas disponibles para seleccionar aquella cuyo valor se ajusta mejor a nuestros intereses.

Conviene señalar que las medidas *nominales* sólo aprovechan información nominal. Únicamente informan del grado de asociación existente, no de la dirección o naturaleza de tal asociación.

Medidas basadas en *chi*-cuadrado. Son medidas que intentan corregir el valor del estadístico X^2 para hacerle tomar un valor entre 0 y 1, y para minimizar el efecto del tamaño de la muestra sobre la cuantificación del grado de asociación (Pearson, 1913; Cramer, 1946).

- **Coeficiente de contingencia**: $C = \sqrt{X^2/(X^2 + n)}$. Toma valores entre 0 y 1, pero difícilmente llega a 1. Su valor máximo depende del número de filas y de columnas. Si el número de filas y de columnas es el mismo (k), entonces el valor máximo de C se obtiene de la siguiente manera: $C_{máx.} = \sqrt{(k-1)/k}$. Un coeficiente de 0 indica independencia, mientras que un coeficiente que alcanza su valor máximo indica asociación perfecta.

- **Phi y V de Cramer**. El coeficiente *phi* se obtiene de la siguiente manera: $\phi = \sqrt{X^2/n}$. En tablas de contingencia 2x2, *phi* adopta valores entre 0 y 1, y su valor es idéntico al del coeficiente de correlación de Pearson (ver capítulo 16).

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

La *V* de Cramer incluye una ligera modificación de *phi*: $V_{Cramer} = \sqrt{X^2/[n(k-1)]}$, donde *k* se refiere al menor del número de filas y de columnas. V_{Cramer} nunca excede de 1. En tablas de contingencia 2x2, los valores V_{Cramer} y *phi* son idénticos.

Medidas basadas en la reducción proporcional del error (RPE). Son medidas de asociación que expresan la proporción en que conseguimos reducir la probabilidad de cometer un error de predicción cuando, al intentar clasificar un caso o grupo de casos como pertenecientes a una u otra categoría de una variable, en lugar de utilizar únicamente las probabilidades asociadas a cada categoría de esa variable, efectuamos la clasificación teniendo en cuenta las probabilidades de las categorías de esa variable en cada en cada categoría de una segunda variable.

☐ **Lambda**. Esta opción permite obtener dos medidas de asociación desarrolladas por Goodman y Kruskal: *lambda* y *tau* (ver Goodman y Kruskal, 1979).

La medida de asociación **lambda** parte de la siguiente idea: si al predecir a qué categoría de una determinada variable (*X*) pertenece un caso decimos que pertenece a la categoría más probable de todas, estaremos cometiendo un error de predicción igual a la probabilidad de pertenecer a una cualquiera de las restantes categorías; si, en lugar de esto, clasificamos a ese caso en una u otra categoría de la variable *X* dependiendo de a qué categoría de una segunda variable (*Y*) pertenece, podemos estar consiguiendo una reducción en el error de predicción (lo cual ocurrirá si las dos variables están relacionadas). El coeficiente *lambda* expresa la proporción de error de predicción que conseguimos reducir al proceder de esta segunda manera.

Consideremos los datos de la tabla 12.6.a, que recoge las frecuencias resultantes de cruzar las variables *sexo* y *grupos de salario*. Si conocemos la distribución de la variable *grupos de salario*, al estimar a qué grupo de salario pertenece un sujeto cualquiera, diremos que pertenece al grupo de "entre 25.000 y 50.000 $" porque hay una probabilidad de 260/474 = 0,5485 de pertenecer a ese grupo frente a una probabilidad de (143+54+17)/474 = 0,4515 de pertenecer a cualquiera de los otros grupos. Procediendo de esta manera, estaremos cometiendo un error de clasificación de 0,4515.

Si ahora tenemos en cuenta la variable *sexo* para efectuar esa estimación y clasificamos a los varones en el grupo de "entre 25.000 y 50.000 $" porque ese es el grupo de salario más probable entre los varones (con un error de (19+48+17)/474 = 0,1772), y a las mujeres en el grupo de "menos de 25.000 $" porque ese es el grupo de salario más probable entre las mujeres (con un error de (86+6+0)/474 = 0,1941), estaremos cometiendo un error de clasificación de 0,1772+0,1941 = 0,3713. Actuando de esta segunda manera hemos conseguido reducir el error de clasificación en 0,0802 (de 0,4515 a 0,3713), lo cual representa una proporción de reducción de 0,0802/0,4515 = 0,1776, que es justamente el valor que toma *lambda* en los estadísticos de la tabla 12.6.c cuando consideramos la variable *grupos de salario* como variable dependiente.

Lambda tiene tres versiones: dos *asimétricas* (para cuando una de las dos variables se considera independiente y la otra dependiente) y una *simétrica* (para cuando no existe razón para distinguir entre variable independiente y dependiente). La salida del SPSS incluye las tres versiones.

Lambda toma valores entre 0 y 1. Un valor de 0 indica que la variable independiente (la variable utilizada para efectuar pronósticos) no contribuye en absoluto a reducir el error de predicción. Un valor de 1 indica que el error de predicción se ha conseguido reducir por completo, es decir, que la variable independiente permite predecir con toda precisión a qué categoría de la variable dependiente pertenecen los casos clasificados. Cuando dos variables son estadísticamente independientes, *lambda* vale 0. Pero un valor de 0 no implica independencia estadística, pues *lambda* únicamente es sensible a un tipo particular de asociación: a la derivada de la reducción en el error que se consigue al predecir las categorías de una variable utilizando las de la otra. No existe ningún índice de asociación sensible a todo tipo de asociación posible.

La medida de asociación **tau** se parece a *lambda*, pero se basa en una lógica algo diferente. Al pronosticar a qué categoría de la variable *grupos de salario* pertenece un grupo de sujetos, podemos asignar aleatoriamente el 100(143/474) = 30,17 % a la categoría "menos de 25.000 $", el 100(260/474) = 54,85 % a la categoría "entre 25.000 y 50.000 $", etc., basándonos en la probabilidad de pertenecer a cada categoría, en lugar de considerar sólo la categoría más probable, como hemos hecho con *lambda*. Procediendo de esta manera estaremos clasificando correctamente al 30,17 % de los 143 sujetos del grupo "menos de 25.000 $", al 54,85 % de los 260 sujetos con salarios "entre 25.000 y 50.000 $", etc. Lo cual representa una proporción de clasificación correcta global de 0,4061 y, por tanto, una proporción de clasificación errónea de 1–0,4061 = 0,5939.

En lugar de esto, podemos tener en cuenta la variable *sexo* y, entre los varones, asignar aleatoriamente el 100(19/258) = 7,36 % a la categoría "menos de 25.000 $", el 100(174/258) = 67,44 % a la categoría "entre 25.000 y 50.000 $", etc.; y entre las mujeres, asignar aleatoriamente el 100(124/216) = 57,41 % a la categoría "menos de 25.000 $", el 100(86/216) = 39,81 % a la categoría "entre 25.000 y 50.000 $"; etc. Al final, estaremos clasificando de forma correcta al 49,45 % de los sujetos y, por tanto, estaremos efectuando pronósticos erróneos con una probabilidad de 1–0,4945 = 0,5055.

Procediendo de esta segunda manera reducimos la probabilidad de efectuar pronósticos erróneos en 0,0884 (la diferencia entre 0,5939 y 0,5055). Por lo que habremos conseguido reducir la probabilidad de error en una proporción de 0,0884/0,5939 = 0,149, que es justamente el valor que toma la *tau* de Goodman y Kruskal en los estadísticos de la tabla 12.6.c cuando consideramos la variable *grupos de salario* como dependiente.

Al igual que *lambda*, *tau* también toma valores entre 0 y 1, significando el 0 ausencia de reducción del error de clasificación y el 1 reducción completa.

- **Coeficiente de incertidumbre** (Theil, 1970). Al igual que *lambda* y *tau,* el *coeficiente de incertidumbre* es una medida de asociación basada en la reducción proporcional del error. Por tanto, es una medida que expresa el grado de incertidumbre que conseguimos reducir cuando utilizamos una variable para efectuar pronósticos sobre otra.

  Posee dos versiones *asimétricas* (dependiendo de cuál de las dos variables consideremos dependiente) y una *simétrica* (para cuando no hacemos distinción entre variable independiente y dependiente). Se obtiene de la siguiente manera:

$$I_{Y|X} = [I(X) + I(Y) - I(XY)] \,/\, I(Y)$$

donde: $I(X) = -\sum_i \left[(n_i/n)\ \ln(n_i/n)\right]$ (n_i = *frecuencias marginales de las filas*)

$I(Y) = -\sum_j \left[(n_j/n)\ \ln(n_j/n)\right]$ (n_j = *frecuencias marginales de las columnas*)

$I(XY) = -\sum_i \sum_j \left[(n_{ij}/n)\ \ln(n_{ij}/n)\right]$ (n_{ij} = *frecuencias de las casillas* ($n_{ij} > 0$)

Para obtener $I_{X|Y}$ basta con intercambiar los papeles de *I(X)* e *I(Y)*. Y la versión *simétrica* se obtiene multiplicando $I_{Y|X}$ por 2 después de añadirle *I(X)* al denominador.

Ejemplo (Tablas de contingencia > Estadísticos > Datos nominales)

Este ejemplo muestra cómo obtener e interpretar los estadísticos para datos nominales del procedimiento **Tablas de contingencia**.

- En el cuadro de diálogo *Tablas de contingencia* (figura 12.1), seleccionar las variables *sexo* y *salargr* (grupos de salario) como variables *fila* y *columna*, respectivamente.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Tablas de contingencia: Estadísticos* (ver figura 12.3) y marcar las cuatro opciones del recuadro **Nominal** (coeficiente de contingencia, phi y *V* de Cramer, lambda y coeficiente de incertidumbre).

Aceptando estas elecciones, el *Visor* ofrece los resultados que recogen las tablas 12.6.a, 12.6.b y 12.6.c.

Tabla 12.6.a. Tabla de contingencia de *sexo* por *grupos de salario*.

Recuento

		Grupos de salario actual				
		Menos de 25.000 $	Entre 25.000 y 50.000 $	Entre 50.000 y 75.000 $	Más de 75.000 $	Total
Sexo	Hombre	19	174	48	17	258
	Mujer	124	86	6		216
Total		143	260	54	17	474

Tabla 12.6.b. Medidas de asociación *simétricas* del procedimiento *Tablas de contingencia*.

		Valor	Sig. aproximada
Nominal por nominal	Phi	,570	,000
	V de Cramer	,570	,000
	Coeficiente de contingencia	,495	,000
N de casos válidos		474	

a. No asumiendo la hipótesis nula.

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

Tabla 12.6.c. Medidas de asociación *direccionales* del procedimiento *Tablas de contingencia*.

		Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Lambda	Simétrica	,333	,048	6,054	,000
	Sexo dependiente	,486	,040	9,596	,000
	Grupos de salario dependiente	,178	,061	2,641	,008
Tau de Goodman y Kruskal	Sexo dependiente	,325	,036		,000[c]
	Grupos de salario dependiente	,149	,024		,000[c]
Coeficiente incertidumbre	Simétrica	,210	,026	7,948	,000[d]
	Sexo dependiente	,266	,033	7,948	,000[d]
	Grupos de salario dependiente	,173	,021	7,948	,000[d]

a. No asumiendo la hipótesis nula.

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 verosimilitud.

La primera tabla (12.6.a) recoge las frecuencias resultantes de cruzar las variables *sexo* y *grupos de salario*.

Las otras dos tablas (12.6.b y 12.6.c) muestran las medidas de asociación para datos nominales recién estudiadas. Cada medida aparece acompañada de su correspondiente nivel crítico (*Sig. aproximada*), el cual permite decidir sobre la hipótesis de independencia: puesto que el nivel crítico de todas las medidas listadas es muy pequeño (menor que 0,05 en todos los casos), podemos rechazar la hipótesis nula de independencia y concluir que las variables *sexo* y *grupos de salario* están relacionadas.

En la tabla 12.6.c, junto con el valor concreto adoptado por cada medida de asociación aparece su valor estandarizado (*T aproximada*), que se obtiene dividiendo el valor de la medida entre su error típico (calculado éste suponiendo independencia entre las variables). La tabla también ofrece el error típico de cada medida calculado sin suponer independencia (*Error típico asintótico*).

Además de las medidas de asociación, las tablas recogen *notas* aclaratorias acerca de aspectos tales como bajo qué condiciones se hacen algunos cálculos, cómo se obtienen algunos de los niveles críticos que se ofrecen, cuál es el motivo de que no se puedan efectuar algunos cálculos, etc.

Datos ordinales

El apartado **Datos ordinales** recoge una serie de medidas de asociación que permiten aprovechar la información ordinal que las medidas diseñadas para datos nominales pasan por alto. Con datos ordinales ya tiene sentido hablar de la *dirección* de la relación: una relación *positiva* indica que los valores altos de una variable se asocian con los valores altos de la otra, y los valores bajos, con valores bajos; una relación *negativa* indica que los valores altos de una variable se asocian con los valores bajos de la otra, y los valores bajos con valores altos.

Muchas de las medidas de asociación diseñadas para estudiar la relación entre variables ordinales se basan en el concepto de *inversión* y *no inversión*. Si los dos valores de un caso en ambas variables son mayores (o menores) que los dos valores de otro caso, decimos que entre

esos casos se da una *no inversión* (*P*). Si el valor de un caso en una de las variables es mayor que el de otro caso, y en la otra variable el valor del segundo caso es mayor que el del primero, decimos que se da una *inversión* (*Q*). Si dos casos tienen valores idénticos en una o en las dos variables, decimos que se da un *empate* (*E*). Cuando predominan las *no inversiones*, la relación es positiva: conforme aumentan (o disminuyen) los valores de una de las variables, aumentan (o disminuyen) los de la otra. Cuando predominan las *inversiones*, la relación es negativa: conforme aumentan (o disminuyen) los valores de una de las variables, disminuyen (o aumentan) los de la otra.

Todas las medidas de asociación recogidas en este apartado utilizan en el numerador la diferencia entre el número de *inversiones* y *no inversiones* resultantes de comparar cada caso con cada otro, pero se diferencian en el tratamiento dado a los *empates* (ver Somers, 1962; Kendall, 1963; Goodman y Kruskal, 1979).

- **Gamma**: $\gamma = (n_P - n_Q) / (n_P + n_Q)$. Si la relación entre dos variables es perfecta y positiva, todos los pares (comparaciones entre casos) serán *no inversiones* y, consiguientemente, $n_P - n_Q$ será igual a $n_P + n_Q$, en cuyo caso, $\gamma = 1$. Si la relación entre las variables es perfecta, pero negativa, todos los pares serán *inversiones* y, en consecuencia, $n_P - n_Q$ será igual a $-(n_P + n_Q)$, de donde $\gamma = -1$. Si las variables son independientes, habrá tantas *inversiones* como *no inversiones*: $n_P = n_Q$; de modo que $n_P - n_Q = 0$ y $\gamma = 0$. Así pues, γ oscila entre –1 y 1. Si dos variables son estadísticamente independientes, γ vale cero; pero una γ de cero no implica independencia (excepto en tablas de contingencia 2×2).
- **d de Somers**: cuando una de las variables se considera independiente (*X*) y la otra dependiente (*Y*), Somers ha propuesto una modificación del coeficiente γ que consiste en añadir en el denominador de *gamma* el número de pares empatados en la variable dependiente: $d = (n_P - n_Q)/(n_P + n_Q + n_{E(Y)})$. El SPSS ofrece tres versiones: dos asimétricas y una simétrica. La versión *simétrica* se obtiene utilizando en el denominador de la *d* el promedio de los denominadores correspondientes a las dos versiones asimétricas.
- **Tau-b de Kendall**: tanto el coeficiente *tau-b* como el *tau-c* tienen en cuenta el número de empates, pero de distinta manera: $\tau_b = (n_P - n_Q) \Big/ \sqrt{(n_P + n_Q + n_{E(X)})\,(n_P + n_Q + n_{E(Y)})}$. El coeficiente *tau-b* toma valores entre –1 y +1 sólo en tablas de contingencia cuadradas y si ninguna frecuencia marginal vale cero.
- **Tau-c de Kendall**: $\tau_c = 2m(n_P - n_Q)/[n^2(m-1)]$, donde *m* se refiere al valor menor del número de filas y del número de columnas. *Tau-c* toma valores entre aproximadamente –1 y +1 sea cual sea el número de filas y de columnas de la tabla.

Ejemplo (Tablas de contingencia > Estadísticos > Datos ordinales)

Este ejemplo muestra cómo obtener e interpretar los estadísticos para datos ordinales del procedimiento **Tablas de contingencia**.

- En el cuadro de diálogo *Tablas de contingencia* (ver figura 12.1), seleccionar las variables *salargr* (grupos de salario) y *estudios* (nivel de estudios) como variables *fila* y *columna*, respectivamente.

▶ Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Tablas de contingencia: Estadísticos* (ver figura 12.3) y marcar las cuatro opciones del recuadro **Ordinal**: *gamma*, *d de Somers*, *tau-b* y *tau-c*).

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las tablas 12.7.a, 12.7.b y 12.7.c.

Tabla 12.7.a. Tabla de contingencia de *salargr* (grupos de salario) por *estudios* (nivel de estudios).

Recuento

		Nivel de estudios				
		Primarios	Secundarios	Medios	Superiores	Total
Grupos de salario	Menos de 25.000 $	29	95	19		143
	Entre 25.000 y 50.000 $	24	94	136	6	260
	Entre 50.000 y 75.000 $		1	21	32	54
	Más de 75.000 $			5	12	17
Total		53	190	181	50	474

Tabla 12.7.b. Medidas *direccionales* (*d de Somers*) del procedimiento *Tablas de contingencia*.

Ordinal por ordinal d de Somer

	Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Simétrica	,557	,029	16,075	,000
Grupos de salario dependiente	,525	,030	16,075	,000
Nivel de estudios dependiente	,593	,030	16,075	,000

a. No asumiendo la hipótesis nula.

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

Tabla 12.7.c. Medidas *simétricas* (*tau-b*, *tau-c* y *gamma*) del procedimiento *Tablas de contingencia*.

		Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Ordinal por ordinal	Tau-b de Kendall	,558	,029	16,075	,000
	Tau-c de Kendall	,469	,029	16,075	,000
	Gamma	,798	,031	16,075	,000
N de casos válidos		474			

a. No asumiendo la hipótesis nula.

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

La primera tabla (12.7.a) contiene las frecuencias absolutas resultantes de cruzar *salargr* y *estudios*. Las otras dos tablas recogen el coeficiente *d* de *Somers* en sus tres versiones (tabla 12.7.b) y los coeficientes *tau-b*, *tau-c* y *gamma* (tabla 12.7.c).

Cada coeficiente aparece con su correspondiente nivel crítico (*Sig. aproximada*), el cual permite tomar una decisión sobre la hipótesis de independencia. Puesto que estos niveles críticos son menores que 0,05, podemos afirmar que las variables *salargr* y *estudios* están relacionadas. Y como el valor de las medidas es positivo (relación positiva), podemos concluir interpretando que a mayor nivel de estudios corresponde mayor salario.

Al igual que ocurría con las medidas de asociación para datos nominales, junto con el valor de cada coeficiente aparece su valor estandarizado (*T aproximada*), que no es otra cosa que el

valor del coeficiente dividido por su error típico. La tabla también ofrece el error típico de cada medida obtenido sin suponer independencia (*Error típico asintótico*).

Nominal por intervalo.

El coeficiente de correlación **eta** sirve para cuantificar el grado de asociación existente entre una variable cuantitativa (medida en escala de intervalo o razón) y una variable categórica (medida en escala nominal u ordinal). Su mayor utilidad no está precisamente asociada a las tablas de contingencia, pues éstas se construyen, según sabemos ya, utilizando variables categóricas. Pero, puesto que este coeficiente se encuentra disponible en el procedimiento SPSS *Tablas de contingencia*, podemos marcar la opción **Eta** (figura 12.2) y obtener el valor de la relación entre dos variables cuando una de ellas es cuantitativa y la otra categórica. Se trata de un coeficiente de correlación que no supone *linealidad* y cuyo cuadrado puede interpretarse como la proporción de varianza de la variable cuantitativa que está explicada por la variable categórica.

Índice de acuerdo (kappa)

La opción **Kappa** (ver figura 12.2) proporciona una medida del grado de acuerdo existente entre dos observadores o jueces al avaluar una serie de sujetos u objetos (Cohen, 1960). La tabla 12.8 muestra el resultado obtenido por dos jueces al clasificar una muestra de 200 pacientes neuróticos según el tipo de neurosis padecida.

Tabla 12.8. Resultado obtenido por dos *jueces* al diagnosticar una muestra de 200 pacientes histéricos.

Recuento

		Juez 2				
		Fóbica	Histérica	Obsesiva	Depresiva	Total
Juez 1	Fóbica	20	8	6	1	35
	Histérica	7	36	14	4	61
	Obsesiva	1	8	43	7	59
	Depresiva	2	6	4	33	45
Total		30	58	67	45	200

Una forma intuitiva de medir el grado de acuerdo entre los dos jueces consiste en hacer un recuento del número de coincidencias existentes entre ambos (es decir, del número de casos que ambos jueces han clasificado de la misma manera). Sumando las frecuencias que indican acuerdo, es decir, las que se encuentran en la diagonal que va desde la parte superior izquierda de la tabla a la parte inferior derecha, obtenemos 132 coincidencias, lo que representa un porcentaje de acuerdo del 100(132/200) = 66 %.

El problema de utilizar este porcentaje como índice de acuerdo es que no tiene en cuenta la probabilidad de obtener acuerdos por azar. Si suponemos que ambos jueces son independientes, los casos que cabría esperar por azar en las casillas de la diagonal pueden obtenerse multiplicando las correspondientes frecuencias marginales y dividiendo por el total de casos.

Así, en la primera casilla de la diagonal cabría esperar, por azar, 35(30)/200 = 5,25 casos; en la segunda casilla, 61(58)/200 = 17,69 casos; etc. Repitiendo la operación para todas las casillas de la diagonal obtenemos un total de 52,83 casos, lo que representa un 26,42 % de acuerdo esperado por azar. La diferencia entre la proporción de *acuerdo observado* (0,66) y la proporción de *acuerdo esperado por azar* (0,2642) es 0,3958.

La **kappa** de Cohen se obtiene dividiendo esa diferencia por la proporción de acuerdo máximo que los dos jueces podrían alcanzar. Esta proporción máxima se obtiene restando a 1 la proporción de acuerdo esperado por azar: 1–0,2642 = 0,7358. Dividiendo el acuerdo observado (0,3958) entre el acuerdo máximo posible (0,7358), obtenemos una proporción de acuerdo de 0,5379. Este valor debe interpretarse teniendo en cuenta que *kappa* toma valores entre 0 (acuerdo nulo) y 1 (acuerdo máximo).

Ejemplo (Tablas de contingencia > Estadísticos > Kappa)

Este ejemplo muestra cómo obtener e interpretar el índice de acuerdo *kappa* del procedimiento **Tablas de contingencia**.

- Reproducir en el *Editor de datos* los datos de la tabla 12.7 (utilizar la opción **Ponderar** del menú **Datos**).
- En el cuadro de diálogo *Tablas de contingencia* (ver figura 12.1), seleccionar las variables *juez_1* y *juez_2* como variables *fila* y *columna*, respectivamente.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Tablas de contingencia: Estadísticos* (ver figura 12.3) y marcar la opción **Kappa**.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestra la tabla 12.9.

Tabla 12.9. Índice de acuerdo *kappa*.

		Valor	Error típ. asint.[a]	T aproximada[b]	Sig. aproximada
Medida de acuerdo	Kappa	,538	,046	12,921	,000
N de casos válidos		200			

a. No asumiendo la hipótesis nula.

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

La tabla 12.9 recoge el valor del estadístico *kappa* y su nivel crítico (*Significación aproximada*), el cual nos permite decidir sobre la hipótesis de acuerdo nulo: puesto que el nivel crítico es muy pequeño (0,000), podemos rechazar la hipótesis de acuerdo nulo y concluir que existe un acuerdo significativamente más alto que el esperado por azar.

Al igual que con el resto de medidas de asociación estudiadas, junto con el valor del índice *kappa* aparece su valor estandarizado (*T aproximada*), resultado de dividir el valor de *kappa* entre su error típico, calculado éste bajo el supuesto de acuerdo nulo; también recoge la tabla el error típico de *kappa* calculado sin suponer acuerdo nulo (*Error típico asintótico*).

Índices de riesgo

Las frecuencias de una tabla de contingencia pueden obtenerse utilizando dos estrategias básicas de recogida de datos. En la estrategia habitual, que es la que hemos supuesto al aplicar todas las medidas de asociación estudiadas hasta aquí, los datos representan un corte temporal **transversal**: se recogen en el mismo o aproximadamente el mismo punto temporal.

Si, en lugar de esto, medimos una o más variables en una muestra de sujetos y hacemos seguimiento a esos sujetos para volver a tomar una medida de esas mismas variables o de otras diferentes, nos encontramos en una situación **longitudinal**: las medidas se toman en diferentes puntos temporales. Los *índices de riesgo* que estudiaremos en este apartado resultan especialmente útiles para diseños longitudinales en los que medimos dos variables *dicotómicas* (ver Pardo y San Martín, 1998, págs. 511-514).

El seguimiento de los estudios longitudinales puede hacerse ***hacia adelante*** o ***hacia atrás***. En los diseños longitudinales *hacia adelante*, llamados diseños de *prospectivos* o de *cohortes*, los sujetos son clasificados en dos grupos con arreglo a la presencia o ausencia de algún factor desencadenante (por ejemplo, el hábito de fumar, –fumadores y no fumadores) y se les hace seguimiento durante un espacio de tiempo hasta determinar la proporción de sujetos de cada grupo en los que se da un determinado desenlace objeto de estudio (por ejemplo, problemas vasculares). En los diseños longitudinales *hacia atrás*, también llamados *retrospectivos* o de *caso-control*, se forman dos grupos de sujetos a partir de la presencia o ausencia de una determinada condición objeto de estudio (por ejemplo, sujetos sanos y pacientes con problemas vasculares) y se hace seguimiento hacia atrás intentando encontrar información sobre la proporción en la que se encuentra presente en cada muestra un determinado factor desencadenante (por ejemplo, el hábito de fumar).

Lógicamente, cada diseño de recogida de datos permite dar respuesta a diferentes preguntas y requiere la utilización de unos estadísticos particulares. En los diseños de **cohortes**, en los que se establecen dos grupos de sujetos a partir de la presencia o ausencia de una condición que se considera desencadenante y se hace seguimiento hacia adelante para determinar qué proporción de sujetos de cada grupo alcanza un determinado desenlace, la medida de interés suele ser el *riesgo relativo*: el grado en que la proporción de desenlaces es más alta en un grupo que en el otro.

Consideremos los datos de la tabla 12.10 referidos a un estudio sobre la relación entre el hábito de fumar, *tabaquismo*, y la presencia de *problemas vasculares* en una muestra de 240 sujetos.

Tabla 12.10. Tabla de contingencia de *tabaquismo* por *problemas vasculares*.

Recuento

		Problemas vasculares		
		Con problemas	Sin problemas	Total
Tabaquismo	Fuman	23	81	104
	No fuman	9	127	136
Total		32	208	240

Entre los fumadores, la proporción de casos con problemas vasculares vale 23/104 = 0,221. Entre los no fumadores, esa proporción vale 9/136 = 0,066. El riesgo relativo se obtiene divi-

diendo ambas proporciones: 0,221/0,066 = 3,34. Este índice de riesgo (3,34) informa sobre el número de veces que es más probable encontrar problemas vasculares en sujetos fumadores que en sujetos no fumadores. Un índice de riesgo de 1 indica que los grupos considerados no difieren en la proporción de desenlaces.

En los diseños de **caso-control**, tras formar dos grupos de sujetos a partir de alguna condición de interés, se va hacia atrás buscando la presencia de algún factor desencadenante. El mismo estudio sobre tabaquismo y problemas vasculares podría diseñarse seleccionando dos grupos de sujetos diferenciados por la presencia de problemas vasculares y buscando en la historia clínica la presencia o no del hábito de fumar. Puesto que el tamaño de los grupos se fija a partir de la presencia o ausencia de un determinado desenlace, no tiene sentido calcular un índice de riesgo basado en las proporciones de desenlaces (incidencias), pues el número de fumadores y no fumadores no ha sido previamente establecido sino que es producto del muestreo. Pero podemos calcular la *ratio* fumadores/no-fumadores tanto en el grupo de sujetos con problemas vasculares como en el grupo de sujetos sin problemas, y utilizar el cociente entre ambas *ratios* como una estimación del riesgo relativo.

Basándonos en los datos de la tabla 12.10, la *ratio* fumadores/no-fumadores en el grupo de sujetos con problemas vasculares vale: 23/9 = 2,5556; y en el grupo de sujetos sin problemas: 81/127 = 0,6378. El índice de riesgo en un diseño *caso-control* se obtiene dividiendo ambas *ratios*: 2,5556/0,6378 = 4,007. Este valor se interpreta de la misma manera que el índice de riesgo relativo (pues es una estimación del mismo), pero también admite esta otra interpretación: entre los sujetos con problemas vasculares, es 4,007 veces más probable encontrar fumadores que no fumadores. Un índice de riesgo de 1 indica que la probabilidad de encontrarnos con el factor desencadenante es la misma en las dos cohortes estudiadas.

Ejemplo (Tablas de contingencia > Estadísticos > Riesgo)

Este ejemplo explica cómo obtener e interpretar los índices de riesgo del procedimiento **Tablas de contingencia**.

- Reproducir en el *Editor de datos* los datos de la tabla 12.10.
- En el cuadro de diálogo *Tablas de contingencia* (ver figura 12.1), seleccionar las variables *tabaco* (tabaquismo) y *vascular* (problemas vasculares) como variables *fila* y *columna*, respectivamente.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Tablas de contingencia: Estadísticos* (ver figura 12.3) y marcar la opción **Riesgo**.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestra la tabla 12.11.

Tabla 12.11. Índices de riesgo (*cohortes* y *caso-control*) del procedimiento *Tablas de contingencia*.

		Intervalo de confianza al 95%	
	Valor	Inferior	Superior
Razón de las ventajas para: Tabaco (Fuman / No fuman)	4,007	1,766	9,093
Para la cohorte: Problemas vasculares = Con problemas	3,342	1,615	6,915
Para la cohorte: Problemas vasculares = Sin problemas	,834	,746	,933
N de casos válidos	240		

La primera fila de la tabla indica que el riesgo estimado se refiere al de fumadores sobre el de no fumadores (*Fuman/No fuman*) en un diseño de *caso-control* (*Razón de las ventajas*). Su valor (4,007) significa que, entre los sujetos con problemas vasculares, la probabilidad (el riesgo) de encontrar fumadores es 4 veces mayor que la de encontrar no fumadores. La razón de ventajas también puede interpretarse como una estimación del riesgo relativo (particularmente si la proporción de desenlaces es pequeña): el riesgo de padecer problemas vasculares es 4 veces entre fumadores que entre no fumadores.

Los límites del intervalo de confianza calculado al 95 por ciento indican que el riesgo obtenido es mayor que 1: concluiremos que el riesgo es significativamente mayor que 1 cuando, como en el ejemplo, el valor 1 no se encuentre entre los límites obtenidos.

Las dos filas siguientes ofrecen dos índices de riesgo para un diseño de *cohortes* (dos índices porque el desenlace que interesa evaluar puede encontrarse en cualquiera de las dos categorías de la variable). Si el desenlace que interesa estudiar es la *presencia* de problema vascular (*Problemas vasculares = Con problemas*), la probabilidad o riesgo de encontrar tal desenlace entre los fumadores es 3,34 veces mayor que la de encontrarlo entre los no fumadores: por cada sujeto *con* problema vascular entre los no fumadores, podemos esperar encontrar 3,34 sujetos *con* problema vascular entre los fumadores. Si el desenlace que interesa estudiar es la *ausencia* de problema vascular (*Problemas vasculares = Sin problemas*), la probabilidad o riesgo de encontrar tal desenlace entre los fumadores es menor que entre los no fumadores: por cada sujeto *sin* problema vascular entre los no fumadores, podemos esperar encontrar 0,83 sujetos *sin* problema vascular entre los fumadores.

Proporciones relacionadas (McNemar)

Una variante de los diseños longitudinales recién estudiados consiste en medir una misma variable dicotómica (éxito-fracaso, acierto-error, a favor-en contra, etc.) en dos momentos temporales diferentes. Esta situación es propia de diseños *antes-después* y resulta especialmente útil para medir el cambio. Se procede de la siguiente manera: se toma una medida de una variable dicotómica, se aplica un tratamiento (o simplemente se deja pasar el tiempo) y se vuelve a tomar una medida de la *misma variable* a los *mismos sujetos*.

Estos diseños permiten contrastar la hipótesis nula de *igualdad de proporciones antes-después*, es decir, la hipótesis de que la proporción de *éxitos* es la misma en la medida *antes* y en la medida *después* (la categoría *éxito* se refiere a una cualquiera de las dos categorías de la variable dicotómica estudiada). La tabla 12.12 muestra un ejemplo de diseño *antes-después* con una muestra de 240 sujetos a los que se les ha preguntado sobre su intención de voto antes y después de un debate televisado entre los líderes de los partidos A y B.

Tabla 12.12. Tabla de contingencia de *Intención de voto antes* por *Intención de voto después*.

Recuento

		Intención de voto (después)		
		Partido A	Partido B	Total
Intención de voto (antes)	Partido A	51	45	96
	Partido B	80	64	144
Total		131	109	240

En este ejemplo, la hipótesis sobre *igualdad de proporciones antes-después* puede formularse así: la proporción de sujetos que tienen intención de votar al *partido A* en la medida *antes* es la misma que la proporción de sujetos que tienen intención de votar al *partido A* en la medida *después*. (Sería equivalente referir la hipótesis al *partido B* en lugar de al *partido A*; es decir, es irrelevante cuál de las dos categorías de la variable dicotómica se considere *éxito*).

Para contrastar esta hipótesis, el estadístico de McNemar (1947) compara los cambios que se producen entre el *antes* y el *después* en ambas direcciones y determina la probabilidad de encontrar ese número concreto de cambios si las proporciones *antes-después* fueran iguales. De acuerdo con la hipótesis nula, los cambios de intención de voto en una dirección (del partido A al partido B) deben ser los mismos que los cambios de intención de voto en la otra dirección (del partido B al partido A). Podremos rechazar la hipótesis de igualdad de proporciones cuando los cambios en una dirección sean significativamente más numerosos que en la otra.

Si el número de cambios (en ambas direcciones) no es demasiado grande, el SPSS intenta calcular la probabilidad exacta de encontrar un número de cambios como el observado o más alejado del valor esperado. Para obtener esta probabilidad exacta se basa en la distribución *binomial* con parámetros n = *número de cambios* y $\pi = 0{,}5$.

Si el número de cambios es muy grande, en lugar de obtener la probabilidad exacta del número de cambios observados, el SPSS ofrece una probabilidad aproximada basada en el estadístico de McNemar y en la distribución *chi-cuadrado*:

$$X^2_{\text{McNemar}} = \frac{(\text{nº de cambios en una dirección} - \text{nº de cambios en la otra dirección})^2}{\text{nº total de cambios}}$$

Este estadístico se distribuye según el modelo de probabilidad *chi-cuadrado* con 1 grado de libertad. En el ejemplo de la tabla 12.12 obtenemos:

$$X^2_{\text{McNemar}} = \frac{(45 - 80)^2}{45 + 80} = 9{,}8$$

La probabilidad de encontrar un valor de 9,8 o mayor en la distribución *chi-cuadrado* con 1 grado de libertad es menor que 0,01, lo que debe llevarnos a rechazar la hipótesis de igualdad de proporciones.

Ejemplo (Tablas de contingencia > Estadísticos > McNemar)

Este ejemplo muestra cómo obtener e interpretar el estadístico de McNemar del procedimiento **Tablas de contingencia**.

- Reproducir en el *Editor de datos* los datos de la tabla 12.12.
- En el cuadro de diálogo *Tablas de contingencia* (ver figura 12.1), seleccionar las variables *antes* (intención de voto antes) y *después* (intención de voto después) como variables *fila* y *columna*, respectivamente.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Tablas de contingencia: Estadísticos* (ver figura 12.3) y marcar la opción **McNemar**.

Aceptando estas elecciones, el *Visor* ofrece los resultados que recoge la tabla 12.13.

Tabla 12.13. *Prueba de McNemar* del procedimiento *Tablas de contingencia*.

	Valor	Sig. exacta (bilateral)
Prueba de McNemar		,002[a]
N de casos válidos	240	

a. Utilizando la distribución binomial

La tabla muestra el nivel crítico asociado al número de cambios observados (*Significación exacta bilateral*) y el número de casos válidos. El hecho de que la tabla no muestre el valor del estadístico de McNemar significa que el nivel crítico se ha calculado utilizando la distribución *binomial* (la cual permite obtener la probabilidad exacta en lugar de la aproximada). Según hemos señalado ya, cuando el tamaño muestral no constituye una barrera computacional, el SPSS intenta obtener el nivel crítico exacto en lugar del aproximado (este último es el que se obtendría con el estadístico de McNemar y la distribución *chi-cuadrado*).

Cualquiera que sea la forma de obtenerlo, el nivel crítico indica el grado de compatibilidad existente entre los datos muestrales y la hipótesis nula de *igualdad de proporciones antes-después*. Puesto que el nivel crítico es menor que 0,05, podemos rechazar la hipótesis nula y concluir que la proporción de sujetos que tenían intención de votar al *partido A* antes del debate televisado (96/240 = 0,40) ha cambiado significativamente –ha aumentado– tras el debate (131/240 = 0,55).

Combinación de tablas 2x2 (Cochran y Mantel-Haenszel)

En ocasiones, puede interesar analizar los diseños de *cohortes* y de *caso-control* (descritos ya en el apartado sobre los índices de riesgo) controlando el efecto de terceras variables. Estas situaciones se producen, por ejemplo, cuando se desea evaluar el efecto de un tratamiento sobre una determinada respuesta utilizando distintos grupos de pacientes.

En general, se trata de estudiar si existe o no asociación entre una variable *factor* y una variable *respuesta*, ambas dicotómicas, cuando se dispone de información referida a varios *estratos* (distintos grupos de edad o de sexo, pacientes con distinta sintomatología, distintas dosis de fármaco, distintos grupos étnicos, etc.). La tabla 12.14 recoge datos referidos a las variables *Tabaquismo* y *Problemas vasculares* en dos estratos: *varones* y *mujeres*.

Tabla 12.14. Tabla de contingencia de *Tabaquismo* por *Problemas vasculares* en *varones* y *mujeres*.

Recuento

			Problemas vasculares		
Sexo			Con problemas	Sin problemas	Total
Varones	Tabaquismo	Fuman	22	103	125
		No fuman	17	151	168
	Total		39	254	293
Mujeres	Tabaquismo	Fuman	23	81	104
		No fuman	9	127	136
	Total		32	208	240

En estas situaciones, utilizar el estadístico *chi-cuadrado* de Pearson sobre el conjunto de datos agrupados, puede arrojar resultados equívocos. Y analizar separadamente cada estrato no proporciona una idea global del efecto de la variable factor. Se obtiene información más ajustada utilizando los estadísticos de Cochran y Mantel-Haenszel para contrastar la hipótesis de independencia condicional, es decir, la hipótesis de independencia entre las variables *factor* y *respuesta* una vez que se ha controlado el efecto de los *estratos*. El estadístico de Cochran (1954) adopta la siguiente forma:

$$X^2_{\text{Cochran}} = \frac{\left(\sum_k n_k - \sum_k m_k\right)^2}{\sum_k \sigma^2_{n_k}}$$

donde: k = cada uno de los estratos.
n_k = frecuencia observada en una cualquiera de las casillas del estrato k (sólo una y siempre la misma en todos los estratos).
m_k = frecuencia esperada correspondiente a n_k.
$\sigma^2_{n_k}$ = $n_{1+k}\, n_{2+k}\, n_{+1k}\, n_{+2k} / n^3$.

(n_{1+k}, n_{2+k}, n_{+1k}, y n_{+2k} son las cuatro frecuencias marginales asociadas a las tablas 2x2 de cada estrato). El estadístico de Mantel-Haenszel (1959) es idéntico al de Cochran, excepto en lo que se refiere a dos detalles: 1) utiliza la corrección por continuidad (resta medio punto al numerador antes de elevar al cuadrado), y 2) cambia ligeramente el denominador de la varianza, utilizando $n^2\,(n-1)$ en lugar de n^3.

Tanto el estadístico de Cochran como el de Mantel-Haenszel se distribuyen según el modelo de probabilidad χ^2 con 1 grado de libertad. Si el nivel crítico asociado a ellos es menor que 0,05, deberemos rechazar la hipótesis de independencia condicional y concluir que, una vez controlado el efecto de los estratos, las variables *factor* y *respuesta* están asociadas.

Ejemplo (Tablas de contingencia > Estadísticos > Cochran-Mantel-Haenszel)

Este ejemplo muestra cómo obtener los estadísticos de Cochran y Mantel-Haenszel utilizando los datos de la tabla 12.14:

- Reproducir en el *Editor de datos* los datos de la tabla 12.14.
- En el cuadro de diálogo *Tablas de contingencia* (ver figura 12.1), seleccionar las variables *Tabaquismo* y *Problemas vasculares* como variables *fila* y *columna*, respectivamente, y la variable *Sexo* como variable de *capa*.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Tablas de contingencia: Estadísticos* (ver figura 12.3) y marcar la opción **Estadísticos de Cochran y de Mantel-Haenszel**.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestra la tabla 12.15.

Tabla 12.15. Pruebas de homogeneidad de la razón de las ventajas.

Estadísticos		Chi-cuadrado	gl	Sig. asintótica (bilateral)
Independencia condicional	De Cochran	13,933	1	,000
	Mantel-Haenszel	12,939	1	,000
Homogeneidad	Breslow-Day	1,911	1	,167
	De Tarone	1,910	1	,167

El estadístico de Cochran vale 13,933 y tiene un nivel crítico asociado (*Sig. asintótica bilateral*) de 0,000, lo que nos permite rechazar la hipótesis de independencia condicional y concluir que, una vez controlado el efecto de la variable *Sexo*, las variables *Tabaquismo* y *Problemas vasculares* están relacionadas. Idéntica conclusión se obtiene con el estadístico de Mantel-Haenszel.

Si se rechaza la hipótesis de independencia condicional, el interés del investigador debe orientarse hacia la cuantificación del grado de relación existente entre las variables *factor* y *respuesta*. Para ello, El SPSS ofrece una estimación del riesgo (*odds-ratio*) común para todos los estratos. Pero esta estimación *común* sólo tiene sentido si no existe interacción triple, es decir, si la relación detectada es homogénea en todos los estratos.

Esta hipótesis de homogeneidad de las *odds-ratio* puede contrastarse utilizando los estadísticos de Breslow-Day (1980, 1987) y de Tarone (Tarone *et al.* 1983). En la tabla 12.15 vemos que el nivel crítico asociado a ambos estadísticos vale 0,167, por lo que podemos mantener la hipótesis de homogeneidad. Y, puesto que podemos asumir que el riesgo es homogéneo en todos los estratos, tiene sentido obtener una estimación común del riesgo. La tabla 12.16 ofrece una solución basada en un estadístico debido a Mantel-Haenszel (1959):

$$RV_{\text{común}} = \frac{\sum_k (n_{11k} n_{22k} / n_{++k})}{\sum_k (n_{12k} n_{21k} / n_{++k})}$$

En el ejemplo, el valor del riesgo común (*Estimación*) es 2,068, con un intervalo de confianza definido por los límites 1,555 y 4,373. Puesto que el intervalo de confianza no abarca el valor 1, podemos concluir que, el riesgo común (el de todos los estratos tomados juntos), es significativamente mayor que 1.

Tabla 12.16. Estimación de la razón de las ventajas común de Mantel-Haenszel.

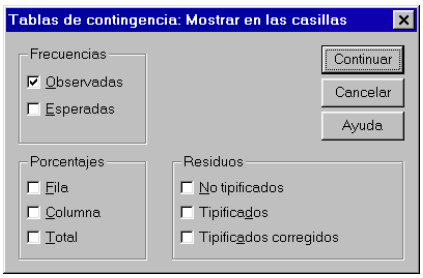
Estimación			2,608
ln(estimación)			,959
Error típ. de ln(estimación)			,264
Sig. asintótica (bilateral)			,000
Intervalo de confianza asintótico al 95%	Razón de ventajas común	Límite inferior	1,555
		Límite superior	4,373
	ln(Razón de ventajas común)	Límite inferior	,442
		Límite superior	1,475

Contenido de las casillas

Todas las tablas de contingencia estudiadas hasta ahora se han construido utilizando las frecuencias absolutas, es decir, el número de casos resultantes de la clasificación. Pero las casillas o celdas de una tabla de contingencia pueden contener información muy variada (frecuencias observadas, porcentajes, residuos, etc.). Parte de esta información es esencial para poder interpretar apropiadamente las pautas de asociación presentes en una tabla dada. Para controlar el contenido de las casillas:

- Pulsar el botón **Casillas...** del cuadro de diálogo *Tablas de contingencia* (ver figura 12.1) para acceder al subcuadro de diálogo *Tablas de contingencia: Mostrar en las casillas* que recoge la figura 12.4.

Figura 12.4. Subcuadro de diálogo *Tablas de contingencia: Mostrar en las casillas*.



Frecuencias. Podemos elegir uno o los dos siguientes tipos de frecuencias absolutas:

- **Observadas**. Número de casos resultantes de la clasificación.
- **Esperadas**. Número de casos que deberíamos encontrar en cada casilla si las variables utilizadas fueran independientes.

Porcentajes. Podemos elegir una o más de las siguientes frecuencias porcentuales:

- **Fila**. Porcentaje que la frecuencia observada de una casilla representa respecto al total marginal de su fila.
- **Columna**. Porcentaje que la frecuencia observada de una casilla representa respecto al total marginal de su columna.
- **Total**. Porcentaje que la frecuencia observada de una casilla representa respecto al número total de casos.

Residuos. Los residuos son las diferencias existentes entre las frecuencias observadas y esperadas de cada casilla. Son especialmente útiles para interpretar la pautas de asociación presentes en una tabla. Podemos elegir una o más de las siguientes opciones:

- **No tipificados**. Diferencia entre la frecuencia observada y la esperada.

- **Tipificados**. Residuo no tipificado dividido por la raíz cuadrada de su correspondiente frecuencia esperada. Su valor esperado vale 0, pero su desviación típica es menor que 1, lo cual hace que no puedan interpretarse como puntuaciones *Z*. Sin embargo, sirven como indicadores del grado en que cada casilla contribuye al valor del estadístico *chi-cuadrado*. De hecho, sumando los cuadrados de los residuos tipificados obtenemos el valor del estadístico *chi-cuadrado*.

- **Tipificados corregidos**. Residuos tipificados corregidos de Haberman (1973). Estos residuos se distribuyen normalmente con media 0 y desviación típica 1. Se calculan dividiendo el residuo de cada casilla por su *error típico*, que en tablas bidimensionales se obtiene como la raíz cuadrada de: $\hat{m}_{ij}(1-n_i)(1-n_j)/n^2$.

  La gran utilidad de los residuos tipificados corregidos radica en que, puesto que se distribuyen normalmente con media cero y desviación típica uno, $N(0, 1)$, son fácilmente interpretables: utilizando un nivel de confianza de 0,95, podemos afirmar que los residuos mayores de 1,96 delatan casillas con más casos de los que debería haber en esa casilla si las variables estudiadas fueran independientes; mientras que los residuos menores de –1,96 delatan casillas con menos casos de los que cabría esperar bajo la condición de independencia.

  En tablas de contingencia con variables nominales, una vez que hemos establecido que entre dos variables existe asociación significativa (mediante el estadístico *chi-cuadrado*) y que hemos cuantificado esa asociación con algún índice de asociación (coeficiente de contingencia, etc.), los residuos tipificados corregidos constituyen la mejor herramienta disponible para poder interpretar con precisión el significado de la asociación detectada.

Ejemplo (Tablas de contingencia > Celdas > Frecuencias, Porcentajes y Residuos)

Este ejemplo muestra cómo obtener e interpretar los diferentes tipos de frecuencias y residuos que ofrece el procedimiento **Tablas de contingencia**.

- En el cuadro de diálogo *Tablas de contingencia* (ver figura 12.1), seleccionar las variables *sexo* y *catlab* como variables *fila* y columna, respectivamente.
- Pulsar el botón **Casillas...** para acceder al subcuadro de diálogo *Tablas de contingencia: Casillas* (ver figura 12.3) y marcar todas las opciones de los recuadros **Frecuencias**, **Porcentajes** y **Residuos** (ver figura 12.4).

Aceptando estas elecciones, el *Visor de resultados* ofrece la información que muestra la tabla 12.14.

Cada casilla de la tabla contiene ocho valores que se corresponden exactamente a con las ocho opciones del subcuadro de diálogo *Tablas de contingencia: Mostrar en las casillas* (ver figura 12.4). Los distintos porcentajes pueden ayudarnos a intuir posibles pautas de asociación, pero son los residuos tipificados corregidos los que nos permiten interpretar de forma precisa la relación existente entre las variables *sexo* y *catlab*. Basta con fijarnos en aquellos que son mayores que +1,96 o menores que –1,96: en el grupo de *Administrativos*, existe una proporción

significativamente más alta de mujeres que de varones (8,8 frente a –8,8), mientras que en los grupos de *Seguridad* y *Directivos* existe una proporción significativamente más alta de varones que de mujeres (4,9 y 6,8 frente a –4,9 y –6,8).

Tabla 12.14. Tabla de contingencia de *sexo* por *categoría laboral*.

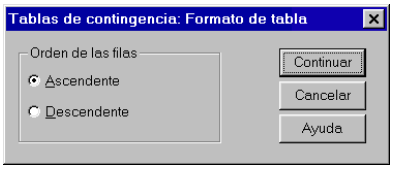
			Categoría Laboral			
			Administrativo	Seguridad	Directivo	Total
Sexo	Hombre	Frecuencia observada	157	27	74	258
		Frecuencia esperada	197,6	14,7	45,7	258,0
		% de Sexo	60,9%	10,5%	28,7%	100,0%
		% de Categoría Laboral	43,3%	100,0%	88,1%	54,4%
		% del total	33,1%	5,7%	15,6%	54,4%
		Residuos	-40,6	12,3	28,3	
		Residuos tipificados	-2,9	3,2	4,2	
		Residuos corregidos	-8,8	4,9	6,8	
	Mujer	Frecuencia observada	206	0	10	216
		Frecuencia esperada	165,4	12,3	38,3	216,0
		% de Sexo	95,4%	,0%	4,6%	100,0%
		% de Categoría Laboral	56,7%	,0%	11,9%	45,6%
		% del total	43,5%	,0%	2,1%	45,6%
		Residuos	40,6	-12,3	-28,3	
		Residuos tipificados	3,2	-3,5	-4,6	
		Residuos corregidos	8,8	-4,9	-6,8	
Total		Frecuencia observada	363	27	84	474
		Frecuencia esperada	363,0	27,0	84,0	474,0
		% de Sexo	76,6%	5,7%	17,7%	100,0%
		% de Categoría Laboral	100,0%	100,0%	100,0%	100,0%
		% del total	76,6%	5,7%	17,7%	100,0%

Formato

Para controlar algunos detalles relacionados con el aspecto de las tablas de contingencia generadas:

- Pulsar el botón **Formato...** del cuadro de diálogo *Tablas de contingencia* (ver figura 12.1) para acceder al subcuadro de diálogo *Tablas de contingencia: Formato de tabla* que muestra la figura 12.5.

Figura 12.5. Subcuadro de diálogo *Tablas de contingencia: Formato de tabla*.



Orden de filas. Las opciones de este recuadro permiten controlar el orden en el que aparecen las categorías de la variable *fila*:

- **Ascendente**. Muestra las categorías de la variable fila ordenadas de menor a mayor. Es la opción por defecto.
- **Descendente**. Muestra las categorías de la variable fila ordenadas de mayor a menor.