Capítulo IX

Tablas de contingencia de dos variables

1. Objetivos didácticos del capítulo

Toda la exposición realizada hasta el momento se ha centrado en el análisis de una variable (univariante): el séptimo capítulo se dedicó a cómo interpretar la distribución de una variable, mientras que en el octavo se explicó la creación de nuevas variables y la modificación de los valores de las variables existentes. Hasta este momento todos los análisis se han realizado variable a variable, tomando cada variable por separado. Ahora bien, en los últimos apartados del octavo capítulo (secciones 8.8 y 8.9) se ha realizado una ligera introducción al análisis bivariable al llevar a cabo el análisis de una variable considerando los diferentes valores de una segunda variable. En estos apartados se analizó el número de asignaturas que obligan a leer libros en función del curso del entrevistado, bien utilizando *criterios condicionales* (apartado 8.8) o mediante la *segmentación del archivo* (apartado 8.9), que ha servido de iniciación al análisis bivariable.

El presente capítulo está dedicado al análisis conjunto de dos variables, análisis bivariable, centrando su atención en los *cruces de tablas*, *tablas cruzadas*, o *tablas de contingencia*; una de las herramientas más utilizadas por el analista de encuestas. La utilización de tablas de contingencia aporta información conjunta de dos (o más) variables mostrando las respuestas de una en función de la otra; indicando el valor que toma la primera variable cuando la segunda tiene un determinado valor. Como tendremos ocasión de comprobar a lo largo del capítulo, esta herramienta supone grandes mejoras frente a los *criterios condicionales* o la *segmentación de archivo* presentado en el capítulo anterior.

Al igual que procedimos a lo largo de todo el trabajo, la explicación se llevará a cabo utilizando ejemplos realizados con el archivo de datos obtenido del cuestionario presentado en el segundo capítulo, sección 2.7 (ENCUESTAS ESTUDIANTES 2002_03.SAV). Los ejercicios propuestos en el capítulo nueve de los *materiales complementarios* deben realizarse con el archivo "Encuestas estudiantes (SIETE promociones). sav".

2. Elaboración de tablas de contingencia con dos variables

La elaboración de tablas de contingencia se encuentra en el menú Analizar, dentro del submenú Estadísticos descriptivos, de modo que para elaborar una tabla hay que seleccionar *Analizar⇒Estadísticos descriptivos⇒Tablas de contingencia*, lo que da paso al cuadro de diálogo de la figura 9.1 Realizaremos una tabla muy sencilla para comenzar con la explicación, teniendo presente que el objetivo es conocer las actividades de ocio fuera del hogar que caracterizan y diferencian a los hombres de las mujeres. Para elaborar esta tabla se selecciona una variable para las filas y se hace un clic en la primera de los *botones-flecha* de la figura 9.1. En este caso se ha elegido la primera variable del cuestionario (v01) que recoge la actividad, fuera de casa, que más gusta hacer cuando se dispone de tiempo libre. La variable seleccionada pasará a la ventana *Filas*.

Conviene recordar que esta variable ha sido recodificada en el capítulo, concretamente en el apartado dedicado a la *recodificación* en las mismas variables (8.4). De modo que, estrictamente hablando, no trabaremos con v01 sino con v01bis; puesto que se ha realizado una recodificación *en distintas variables*.

Figura 9.1. Cuadro de diálogo Tablas de contingencia.

Es importante tener claro el resultado del proceso para apreciar la transformación experimentada; pasando de una variable de 18 categorías de respuesta a 8. Esto se ha producido como consecuencia de elaborar un denominador común definido como "otras", que agrupa las categorías "ir al teatro" (codificada con el valor 7), "ir a conciertos" (9), "leer libros" (10), "otras" (14), "ninguna en particular" (15), "quedar con amigos" (16) y "quedar con el novio" (17). Otra de las transformaciones ha consistido en unir dos categorías debido a la similitud temática entre ellas: se trata de "ir de excursión" (codificada con el valor 4) e "ir al monte" (18), que han sido unidos en una única categoría. Es importante destacar que la opción "otras" supera ligeramente el 10% de los casos. El escaso tamaño muestral nos ha llevado a tomar esta decisión.

De esta exposición se deduce que antes de realizar un cruce de tablas es necesario que las variables a cruzar hayan sido analizadas y, cuando sea preciso, proceder a su transformación con el fin de que presenten categorías con un número aceptable de respuestas. Las variables "no agrupadas" presentan dos problemas:

- Por un lado es imposible generalizar tomando en consideración un escaso número de entrevistados. ¿Qué generalización podemos hacer de las personas que han señalado que en su tiempo libre lo que más les gusta es ir al teatro?, opción que ha sido elegida por 2 entrevistados que suponen un 1% de la muestra. Que se trate de dos mujeres, ¿quiere decir que las mujeres prefieren el teatro más que los hombres? El escaso número de elecciones impide realizar tal afirmación; no es posible extraer conclusiones de una categoría elegida únicamente por dos personas.
- En segundo lugar, y como veremos en la sección 3, los tests estadísticos utilizados para conocer la relación entre variables no funcionan correctamente cuando la tabla analizada tiene muchas *celdillas* con pocas respuestas.

Finalizada la explicación de la variable en filas, posteriormente se selecciona la variable V49, que corresponde al sexo, y se coloca en la ventana de las *columnas*[103]. Antes de proceder de esta forma es preciso conocer como se distribuye el sexo, puesto que se trata de una variable que no ha sido tratada con anterioridad. Las frecuencias de v49 desvelan que el 39% de los entrevistados son hombres y un 61% de mujeres. Pulsando el botón *Aceptar* obtenemos una tabla de contingencia, en su formato más sencillo, que se muestra en la tabla 9.1.

103. Las variables situadas en las columnas son conocidas como *variables de identificación*, *cabeceras*, o *variables cabecera*.

Resumen del procesamiento de los casos

	Casos					
	Válidos		Pérdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
(v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género	180	94,2%	111	5,8%	191	100,0%

Tabla de contingencia (v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género

		Genero		
		Hombre	Mujer	Total
v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género	Beber, ir de copas	26	20	46
	Bailar	0	12	12
	Hacer deporte	12	8	20
	Ir de excursión y al monte	6	6	12
	Viajar	12	28	40
	Ir al cine	0	10	10
	Practicar alguna afición o hobby	12	8	20
	Otras[104]	2	18	20
Total		70	110	180

Tabla 9.1. Cruce de tablas entre actividad fuera de casa que más le gusta hacer en su tiempo libre (v01bis) y sexo (v49).

Comenzaremos la interpretación de esta tabla analizando el "resumen del procesamiento de los casos" que informa sobre el número de datos analizados (191), los casos válidos (180, el 94,2%) y los perdidos (11, un 5,8%). Posteriormente procedemos con

104. Tener en cuenta que dividir más esta categoría supondrían más celdillas con menos de 5 entrevistados.

el análisis de los *marginales*, los totales de filas y columnas, para ocuparnos –más adelante– de las celdillas resultantes de la intersección de filas y columnas. A la derecha de la fila se muestran los valores totales de la variable filas (v01bis): el número 46 que aparece en la parte derecha de la primera fila corresponde al número de personas que declaran que lo que más les gusta hacer fuera de casa es beber, ir de copas; 12 entrevistados eligen bailar, 20 hacer deporte, 40 viajar... Respecto a las columnas, se trata de una muestra compuesta por 70 hombres y 110 mujeres. Esta diferencia en favor de las mujeres precisará, en el apartado 9.4.1, utilizar un determinado tipo de porcentajes que permite *mitigar* tales diferencias.

3. Utilización de test estadísticos para conocer la relación entre variables nominales

Con el fin de conocer hasta que punto las variables utilizadas en la tabla 9.1 están relacionadas se han desarrollado una serie de medidas que –en un solo índice– señalan la existencia de relación entre dos variables, así como el grado de *asociación* y su dirección. Son medidas cuyos valores oscilan entre un valor mínimo indicativo de la ausencia de asociación (normalmente el cero) y un valor máximo que indica asociación perfecta (el uno o el menos uno). Un valor superior a cero (positivo) indicará relación directa, mientras que un valor inferior a cero (negativo) muestra relación inversa[105]. García Ferrando (1985: 217-222) presenta las características que tiene que cumplir una buena *medida de asociación* entre variables: debe indicar *si existe o no una asociación significativa* entre variables, y *cuantificar la fuerza* de esa asociación. El siguiente requisito que debe cumplir una buena medida de asociación es mostrar la *dirección* de la asociación (positiva o negativa), aunque esto únicamente es posible cuando las variables se han medido a nivel ordinal o de intervalo. La cuarta característica es describir la *naturaleza* de la asociación, referida a la distribución de las magnitudes de las variables en cada una de las *celdillas* de la tabla: la comparación de los porcentajes puede mostrar una escasa diferencia en las categorías bajas de las variables, diferencia que se acentúa en las categorías medias y aún más en las altas (relación lineal), o puede tener una tendencia totalmente irregular. Por último tienen que ser medidas *estandarizadas* o *tipificadas* que permitan comparar los índices obtenidos

105. La relación será directa o inversa si las categorías de ambas variables están medidas en el mismo orden. Recordar que en la sección donde se explicaron las escalas se señaló que la codificación de las escalas ordinales (y de intervalo) debe respetar el orden serial.

en distintas tablas. A este respecto García Ferrando (1985: 222) presenta un ejemplo donde señala una supuesta investigación que detecta que la relación entre edad e "interés por la política" es de +0,52; mientras que la relación de está última con el nivel de ingresos es del +0,35. El hecho que estas medidas sean estandarizadas implica que se puedan comparar ambos índices, con el fin de por establecer que el "interés por la política" está más relacionado con la edad que con el nivel de ingresos.

Ilustraremos la explicación del párrafo anterior considerando hasta que punto una de las medidas de asociación más conocidas, el coeficiente de *correlación* lineal de Pearson[106] (la "*r*" de Pearson), cumple cada una de estas propiedades. Supondremos para ello un coeficiente de correlación entre la edad y el nivel de ingresos de –0,87. En primer lugar una medida de asociación debe indicar si existe relación entre variables. Si tenemos en cuenta que el coeficiente de correlación puede oscilar entre -1 y +1, indicando el valor central (0) la no existencia de relación; un valor de –0,87 implicará –sin duda– una relación significativa ente ambas variables[107]. Asimismo, si la máxima relación posible entre variables es 1, un valor de 0,87 indica una relación importante. Por último es posible conocer la dirección de la asociación ya que valores cer-

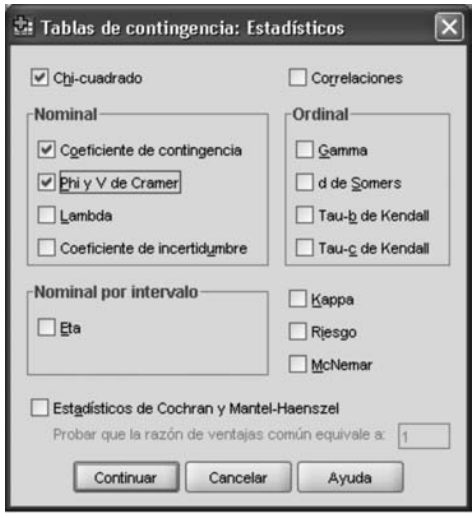

Figura 9.2. *Estadísticos* de tablas de contingencia con variables nominales.

106. En adelante nos referiremos a esta medida como *Coeficiente de Correlación*, si bien es importante indicar que en los resultados del SPSS aparece como "R de Pearson"; tal y como puede apreciarse en la tabla 9.16.

107. Para ello deberemos calcular el nivel de significación de este valor, aspecto que no expondremos aquí puesto que queda fuera de nuestros objetivos; y que puede consultarse en cualquier texto de Estadística.

canos a +1 indicarán relación directa entre variables, mientras que valores cercanos a –1 indican relación inversa, como ocurre en este ejemplo.

Este capítulo está dedicado al análisis de variables nominales y ordinales, que son las que habitualmente se utilizan en tablas de contingencia, de modo que olvidaremos las propiedades del coeficiente de correlación (variables de intervalo) para analizar si los estadísticos que miden la relación entre variables nominales y ordinales cumplen cada uno de estos criterios. A fin de llevar a cabo una exposición práctica utilizaremos el ejemplo expuesto en la tabla 9.1, si bien pulsaremos –en el cuadro de diálogo de la figura 9.1– el botón *"Estadísticos..."* para solicitar aquellos test utilizados para conocer la relación entre dos variables nominales, como son las variables sexo y actividad fuera de casa que más le gusta hacer en su tiempo libre. Tras solicitar *Chi-Cuadrado*, *Coeficiente de Contingencia*, *Phi,* y *V de Cramer* (figura 9.2) obtenemos los resultados mostrados en la tabla 9.2.

Pruebas de chi-cuadrado

	Valor	Gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	36,496(a)	7	,000
Razón de verosimilitud	45,236	7	,000
Asociación lineal por lineal	2,781	1	,095
N de casos válidos	180		

a 3 casillas (18,8%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 3,89.

Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,450	,000
	V de Cramer	,450	,000
	Coeficiente de contingencia	,411	,000
N de casos válidos		180	180

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

Tabla 9.2. *Estadísticos* para variables nominales: Chi-Cuadrado, Phi, V de Cramer y C de Contingencia.

3.1. Relación entre variables nominales utilizando el Chi-Cuadrado

El primero de los estadísticos solicitados es el Chi-Cuadrado, que aparece en la tabla 9.2 con el nombre de Chi-Cuadrado de Pearson. Este estadístico es un contraste que tiene en cuenta la totalidad de la tabla y se emplea para saber si la relación entre estas dos variables es significativa. En el cuadro 9.1 se muestra que el Chi-Cuadrado se calcula restando en cada *celdilla* las frecuencias observadas menos las esperadas (o teóricas), multiplicando esta diferencia al cuadrado, y dividiéndola entre las frecuencias esperadas. Las frecuencias esperadas son las que hubiera tenido la tabla de no existir relación entre variables (Calvo, 1990: 146), y se obtiene de multiplicar el total de fila por el total de columna, y dividiendo el resultado entre el número de casos. La frecuencia esperada de la *celdilla* superior izquierda de la tabla 9.1, por ejemplo, se obtiene de la multiplicación (70 * 46) / 180. Las frecuencias esperadas de las dos variables utilizadas se presentan en la tabla 9.3.

Fórmula:

$$X^2 = \Sigma \frac{(FO - FE)^2}{FE}$$

FO: Frecuencias observadas
FE: Frecuencias esperadas (o teóricas)

Cálculo con los datos de la tabla 9.3:

$$\frac{(26 - 17{,}9)^2}{17{,}9} + \frac{(0 - 4{,}7)^2}{4{,}7} + \frac{(12 - 7{,}8)^2}{7{,}8} + \frac{(12 - 15{,}6)^2}{15{,}6} + \frac{(6 - 4{,}7)^2}{4{,}7} +$$

$$\frac{(12 - 15{,}6)^2}{15{,}6} + \frac{(0 - 3{,}9)^2}{3{,}9} + \frac{(12 - 7{,}8)^2}{7{,}8} + \frac{(2 - 7{,}8)^2}{7{,}8} + \frac{(20 - 28{,}1)^2}{28{,}1} +$$

$$\frac{(12 - 7{,}3)^2}{7{,}3} + \frac{(8 - 12{,}2)^2}{12{,}2} + \frac{(6 - 7{,}3)^2}{7{,}3} + \frac{(28 - 24{,}4)^2}{24{,}4} + \frac{(10 - 6{,}1)^2}{6{,}1} +$$

$$\frac{(8 - 12{,}2)^2}{12{,}2} + \frac{(18 - 12{,}2)^2}{12{,}2} = 36{,}496$$

Grados de Libertad = (f - 1) (c - 1) = (8 - 1) (2 - 1) = 7

Cuadro 9.1. Cálculo del Chi-Cuadrado.

El sumatorio de las diferencias entre las frecuencias esperadas y las observadas, multiplicadas al cuadrado y dividiéndola entre las frecuencias teóricas será, en este caso 36,496 (cuadro 9.1). Se ha señalado en el párrafo anterior que las frecuencias esperadas son las que hubiera tenido la tabla de no existir relación entre variables, de modo que si a las frecuencias obtenidas se les resta las esperadas, una gran diferencia estará indicando que existe relación entre variables. Como el estadístico Chi-Cuadrado se calcula sumando los valores de estas diferencias (al cuadrado divididas entre la frecuencia esperada), un elevado valor del Chi-Cuadrado indicará importantes diferencias entre las frecuencias observadas y las esperadas, o dicho de otro modo, existencia de relación entre variables.

Tabla de contingencia (v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género

			(v49) Genero		
			Hombre	**Mujer**	**Total**
(v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género	Beber, ir de copas	Recuento	26	20	46
		Frecuencia esperada	17,9	28,1	46,0
	Bailar	Recuento	0	12	12
		Frec. esperada	4,7	7,3	12,0
	Hacer deporte	Recuento	12	8	20
		Frec. esperada	7,8	12,2	20,0
	Ir de excursión y al monte	Recuento	6	6	12
		Frec. esperada	4,7	7,3	12,0
	Viajar	Recuento	12	28	40
		Frec. esperada	15,6	24,4	40,0
	Ir al cine	Recuento	0	10	10
		Frec. esperada	3,9	6,1	10,0
	Practicar alguna afición o hobby	Recuento	12	8	20
		Frec. esperada	7,8	12,2	20,0
	Otras	Recuento	2	18	20
		Frec. esperada	7,8	12,2	20,0
Total		Recuento	70	110	180
		Frec. esperada	70,0	110,0	180,0

Tabla 9.3. Tabla de contingencia con frecuencias observadas (recuento) y frecuencia esperada.

Explicaremos detenidamente lo expuesto en el párrafo anterior con ayuda de dos ejemplos ficticios mostrados en la tabla 9.4; y cuyo objetivo es analizar si existe relación entre el sexo y el tipo de ocio. Es preciso volver insistir que se trata de un ejemplo ficticio, que supone una importante simplificación de la realidad en la que existen únicamente dos tipos de ocio (beber y bailar), pero que consideramos puede resultar muy útil para el tema que nos ocupa (explicar la existencia –o ausencia– de relación entre variables). Observando la tabla de la izquierda se aprecia con claridad la ausencia de relación entre el sexo de los entrevistados y el tipo de ocio: 25 mujeres dedican su ocio a beber, y otras 25 a bailar. Lo mismo ocurre con los hombres: 25 lo emplean su tiempo de ocio en beber, y otros 25 en bailar.

La tabla de la derecha, sin embargo, presenta grandes diferencias en las pautas de ocio de los hombres y las mujeres: 49 hombres (de una muestra de 100) muestran su preferencia por beber, y 48 mujeres se decantan por bailar. Tan sólo 1 hombre emplean su tiempo libre bailando, y 2 mujeres bebiendo. La tabla central, rotulada con el nombre B, muestra una ligera influencia, influencia que será necesario ver hasta que punto es importante (significativa) o no. Considerando estos ejemplos, ¿tiene claro el lector que tabla C está desvelando una asociación entre variables, algo que no ocurre en la tabla A? La tabla A, así concebida, sería similar a la tabla de frecuencias esperadas –tabla de no relación entre variables– de modo que cuanto más diferentes sean los datos obtenidos respecto a esta tabla, mayor será la relación entre variables. Esto es lo que intentamos explicar dos párrafos más arriba.

	Sexo	
	Hombre	**Mujer**
Beber	25	25
Bailar	25	25

Tabla A

	Sexo	
	Hombre	**Mujer**
Beber	30	20
Bailar	20	20

Tabla B

	Sexo	
	Hombre	**Mujer**
Beber	49	2
Bailar	1	48

Tabla C

Tabla 9.4. Tabla de contingencia (ejemplo ficticio).

Dijimos más arriba que un gran valor del Chi-Cuadrado indicará relación entre variables, pero ¿a partir de que límite definimos el *gran valor* del Chi-Cuadrado? Para responder a esta pregunta es preciso considerar la columna *sig. asintótica (bilateral)* o *sig. aproximada* de los estadísticos mostrados en la tabla 9.2. Estos valores están indicando la *probabilidad de equivocarnos* al señalar que existe relación entre variables, probabilidad expresada en tantos por uno. Considerando la significación del

Chi–Cuadrado, por ejemplo, la probabilidad de equivocarnos si decimos que existe relación entre *sexo* y *actividad fuera de casa que más gusta hacer cuando se dispone de tiempo libre* es de 0,000, o lo que es lo mismo del 0,0%. Es decir, al tratarse de una probabilidad de equivocarnos prácticamente nula decimos que hay relación entre variables, que los valores de la variable "actividad fuera de casa..." presenta variación según el sexo del entrevistado.

Alguien se preguntará por el límite de este valor, ¿cuando se considera que una *probabilidad de equivocarnos* es lo suficientemente alta para que no podamos decir que exista relación entre variables? En el ámbito de la investigación con encuestas se recomienda no considerar valores superiores al 0,05, es decir, probabilidades mayores del 5% son elevadas para afirmar que existe relación entre variables (Avira, 2000: 109). Si algún lector tiene problemas para recordar estos valores recomendamos que piense en el término *nivel de confianza* (que indica la probabilidad de acertar) y recuerde los niveles de confianza manejados en otras asignaturas: es muy probable que haya manejado niveles de confianza del 90% y del 95%. Un *nivel de confianza* del 95% indica una probabilidad de acertar del 95%, esto es, un nivel de *significación* (probabilidad de equivocarnos) del 5%, o dicho en tantos por uno, probabilidad de acertar del 0,95 y de equivocarnos del 0,05.

Veamos, por un momento, el valor del Chi-Cuadrado y la *significación* de la tabla 9.5. La *significación* 0,085 está indicando una probabilidad de equivocarnos del 8,5%, una probabilidad muy superior del 5% (0,05) que recomendamos en el párrafo anterior, lo que implica la ausencia de relación entre *actividad fuera de casa...* y *titulación* (sociología y resto); que son las dos variables que forman esta tabla.

3.2. Consideraciones a tener en cuenta en la utilización del Chi-Cuadrado

Para utilizar correctamente el Chi-Cuadrado los datos deben cumplir una serie de requisitos. Expondremos aquí los señalados por Reynolds (1984: 19): el primero postula que la muestra sea *aleatoria simple,* aspecto que se cumple en contadas ocasiones debido a que la selección de los entrevistados casi nunca se realiza de forma totalmente aleatoria, puesto que los sistemas de rutas y cuotas utilizados para localizar a los individuos elimina la aleatoriedad muestral. Respecto al otro término señalado en cursiva, muy pocas veces se utilizan muestras *simples* en la investigación con encuesta al recurrir normalmente a muestreos estratificados. Como el cumplimiento de ambos supuestos se realiza en contadas ocasiones, el valor del Chi-Cuadrado es el que tendrían nuestros datos si hubiéramos cumplido los citados requisitos, de modo que cuando no se cumplan consideraremos este valor a modo indicativo.

Tabla de contingencia (v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (Titulac) Titulación (Sociología/no sociología)[108]

		Titulación (Sociología/no sociología)		
		Sociología	No sociología	Total
(v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre	Beber, ir de copas	16	30	46
	Bailar	8	4	12
	Hacer deporte	12	8	20
	Viajar	26	14	40
	Ir al cine	6	4	10
	Practicar alguna afición o hobby	12	8	20
	Otras	20	12	32
Total		100	80	180

Pruebas de chi-cuadrado

	Valor	Gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	11,109(a)	6	,085
Razón de verosimilitud	11,153	6	,084
Asociación lineal por lineal	3,160	1	,075
N de casos válidos	180		

a 1 casillas (7,1%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 4,44.

Tabla 9.5. Ejemplo de tabla de contingencia con relación no significativa.

Otro requisito citado por Reynolds es que las categorías de las variables sean exhaustivas y mutuamente excluyentes. El tercer requisito recomienda no considerar el valor del Chi-Cuadrado cuando existan en la tabla muchas celdillas (un 20%) con frecuencias esperadas menores que 5, puesto que en esta situación no se cumple

108. Esta variable fue creada en el ejercicio 12 de la sección 8.10, y diferencia a los estudiantes de sociología del resto.

uno de los supuestos fundamentales de la distribución Chi-Cuadrado. Una de las formas para evitar esta situación es no utilizar el Chi-Cuadrado con pequeños tamaños muestrales. Otra de las estrategias para solucionar este problema es recodificar las variables con muchas categorías, uniendo las celdillas que tienen pocas respuestas con otras categorías similares. Como hemos señalado en el capítulo anterior el criterio utilizado para recodificar es que las categorías unificadas tengan una *significación* temática, eliminando así los errores muestrales altos que tienen las categorías con pocos sujetos. No obstante, en tablas de 2 x 2 no es posible realizar recodificaciones cuando alguna de las celdillas tiene una frecuencia esperada menor que 5, de modo que la única solución es utilizar el Test Exacto de Fisher en vez del Chi-Cuadrado, que el SPSS muestra automáticamente en el momento que se dan estas condiciones.

En la tabla 9.6 se muestra un ejemplo de un cruce de tablas *inaceptable*, un cruce de tablas donde no es posible interpretar el Chi-Cuadrado puesto que el 62% de las celdillas tienen una frecuencia esperada menor que cinco. Una segunda razón –tan importante como la anterior– es el escaso número de entrevistados que eligen algunas categorías (ir al teatro, leer libros), etc.[109]

La siguiente consideración a tener en cuenta en la utilización de este estadístico está relacionada con el hecho que el Chi-cuadrado utiliza una distribución de probabilidad continua como una *aproximación* a una distribución discreta. Esta *aproximación* indica que existe una relativa incorrección en el cálculo del Chi-Cuadrado, incorrección que es mayor a medida que disminuye el número de categorías. Esta incorrección es prácticamente nula en variables discretas con múltiples categorías, pero alcanza valores importantes en variables dicotómicas. Por ello para tablas de 2 x 2 Yates propuso la *Corrección de Continuidad* que lleva su nombre, calculada restando 0,5 al resultado FO - FT del numerador en la fórmula del Chi-cuadrado (cuadro 9.2). Como el objetivo es restar 0,5, si el resultado FO - FT es negativo será necesario sumar 0,5. El programa SPSS calcula automáticamente la corrección de Yates en tablas de contingencia de 2 filas y 2 columnas, de modo que será necesario desviar la atención del valor Chi-Cuadrado a la Corrección de Continuidad de Yates siempre que ésta aparezca.

$$X^2 = \Sigma \frac{([FO - FT] - ó + 0{,}5)^2}{FT}$$

Cuadro 9.2. Corrección de Continuidad de Yates.

109. Obsérvese que no se trata de v01bis, sino de v01. Recordemos que v01 es la variable sin recodificar, tal y como ha sido recogida en el cuestionario.

Tabla de contingencia (v01) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género

		(v49) Genero		
		Hombre	**Mujer**	**Total**
v01) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre	Beber, ir de copas	26	20	46
	Bailar	0	12	12
	Hacer deporte	12	8	20
	Ir de excursión	2	6	8
	Viajar	12	28	40
	Ir al cine	0	10	10
	Ir al teatro	0	2	2
	Ir a conciertos	0	4	4
	Leer libros	0	2	2
	Practicar alguna afición o hobby	12	8	20
	Otras	0	6	6
	Ninguna en particular	2	0	2
	Quedar con amigos/as	0	2	2
	Quedar con el novio/a	0	2	2
	Ir al monte	4	0	4
	Más de una respuesta	1	10	11
Total		71	120	191

Pruebas de chi-cuadrado

	Valor	**Gl**	**Sig. asintótica (bilateral)**
Chi-cuadrado de Pearson	55,209(a)	15	,000
Razón de verosimilitud	70,676	15	,000
Asociación lineal por lineal	3,808	1	,051
N de casos válidos	191		

a 20 casillas (62,5%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es, 74.

Tabla 9.6. Tabla de contingencia con un valor de Chi–Cuadrado erróneo; con un valor que no debe interpretarse.

En la tabla 9.7 no supone ninguna diferencia considerar el Chi-Cuadrado o la Corrección de Continuidad de Yates, pero en numerosas ocasiones nos encontramos con tablas de 2 x 2 donde el Chi-Cuadrado es significativo y no así la Corrección de Continuidad. En estos casos concluiremos que no existe relación significativa entre las variables de la tabla, puesto que en tablas de 2 x 2 la atención debe desviarse del Chi-Cuadrado a la Corrección de continuidad.

Tabla de contingencia (v28) Dispositivos en el ordenador: lectora DVD * (v49) Género

		(v49) Genero		
		Hombre	**Mujer**	**Total**
(v28) Dispositivos en el ordenador: LECTORA DVD	No/no responde	12	66	78
	Si	56	48	104
Total		68	114	182

Pruebas de chi-cuadrado

	Valor	**Gl**	**Sig. asintótica (bilateral)**	**Sig. exacta (bilateral)**	**Sig. exacta (bilateral)**
Chi-cuadrado de Pearson	28,173(b)	1	,000		
Corrección por continuidad(a)	26,554	1	,000		
Razón de verosimilitud	30,019	1	,000		
Estadístico exacto de Fisher				,000	,000
Asociación lineal por lineal	28,019	1	,000		
N de casos válidos	182				

a Calculado sólo para una tabla de 2x2.
b 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 29,14.

Tabla 9.7. Tabla de contingencia 2 x 2, Corrección de Continuidad de Yates.

Finalizaremos la exposición sobre el Chi-Cuadrado volviendo a recordar los requisitos que debe cumplir una *buena* medida de asociación. En páginas anteriores se ha utilizado el coeficiente de correlación como ejemplo para ilustrar algunas explicaciones, debido a su amplia difusión y conocimiento. Desearíamos utilizar de nuevo su

popularidad para analizar las diferencias entre éste y el Chi-Cuadrado, tal y como se presenta en el cuadro 9.3. Las dos primeras características del cuadro 9.3 se refieren específicamente a los requisitos apuntados. Ambos señalan si existe o no asociación, si bien el Chi-Cuadrado no indica el sentido de la asociación entre variables porque las variables nominales no llevan implícitas ninguna relación de orden entre sus categorías.

Chi-Cuadrado

- Afirma si existe o no asociación
- No indica el sentido de la asociación
- Sirve para variables nominales, ordinales y de intervalo.
- No exige "distribución especial" de las variables.
- No exige función especial entre ambas variables.

Coeficiente de Correlación

- Afirma si existe o no relación.
- Indica el grado de relación.
- Indica el sentido de la asociación
- Sólo sirve para variables de intervalo.
- Exige que ambas variables sigan la curva normal.
- Exige función rectilínea lineal entre las variables.

Cuadro 9.3. Diferencias entre el Chi-Cuadrado y el coeficiente de correlación (Calvo 1990: 145).

Volviendo de nuevo a la primera característica del cuadro 9.3, el Chi-Cuadrado permite afirmar si existe o no asociación (significativa) entre variables, pero no indica el grado de esta asociación. Su propia formulación, una resta al cuadrado entre frecuencias observadas y teóricas, genera que no tenga un límite superior fijo como el coeficiente de correlación. El Chi-Cuadrado es siempre positivo y puede llegar a un valor máximo de N(K-1), donde N es el tamaño de la muestra y K es el número más pequeño de filas o columnas, que en el caso de la tabla 9.1 esto implica 180*(2-1). Si el Chi-Cuadrado de la tabla 9.1 puede alcanzar un valor máximo de 180, ¿cuanta relación entre variables indicará el valor obtenido, el 36,496? Es difícil responder a esta pregunta con la información proporcionada por el Chi-Cuadrado, mucho más cuando la sensibilidad del Chi-Cuadrado al tamaño de la muestra[110] genera que dos tablas con idéntica distribución de porcentajes –pero con distinto número de casos–

110. El valor del Chi-Cuadrado varía en función del tamaño de la muestra.

presentan dos valores diferentes. Volviendo a nuestro ejemplo, una distribución porcentual idéntica a la tabla 9.8, pero basada en 360 casos, proporciona un Chi-Cuadrado de 79,992; el doble que el mostrado en la tabla de 180 casos. Estos motivos han llevado a desarrollar distintas medidas de asociación que serán presentadas en el siguiente apartado.

3.3. Estadísticos basados en el Chi-Cuadrado

Las razones apuntadas en los últimos párrafos del apartado anterior, así como la dificultad para cuantificar la relación entre variables una vez que ya se sabe que ésta es significativa, ha generado la necesidad de utilizar medidas de asociación que permitan solucionar estos problemas. Podemos definir *medida de asociación* como un índice numérico que indica la existencia, grado y dirección de la asociación entre dos variables. Las más utilizadas, con su formulación correspondiente, se presentan en el cuadro 9.4. Algunas de ellas únicamente son aplicables en tablas cuadradas, por lo que será necesario conocer en que momento debemos utilizar cada una, algo que se detalla a continuación.

La primera de estas medidas, conocida como *contingencia cuadrática media* o *Phi-Cuadrado* y calculada como se expone en el cuadro 9.4, oscila entre 0 y 1 y su magnitud indica el grado de asociación entre las variables (García Ferrando, 1985: 224). Así 0 indica ausencia de relación y 1 máxima relación entre variables[111]. Únicamente puede ser utilizado en tablas de 2 x 2 porque en tablas con más de dos categorías su valor máximo puede superar la unidad (García Ferrando, 1985: 224). Además, se trata de una medida muy sensible a la presencia de totales marginales desequilibrados (Ruiz Maya et al, 1990: 264).

En tablas que no sean de 2 x 2 el coeficiente de *Contingencia* permite solucionar algunas de las limitaciones de *Phi*, pero presenta el problema que no llega a la unidad aunque las variables estén perfectamente relacionadas[112], ya que el denominador es siempre superior al numerados (ver cuadro 9.4). En tablas cuadradas (cuando el número de filas es igual al de columnas) se utiliza el coeficiente de Contingencia dividido entre el *C máximo* (Calvo, 1990: 157-8 y Reynolds, 1984: 47). El *C máximo* es la raíz cuadrada del número de filas (o columnas) menos uno dividido entre el número de

111. En la tabla 9.2 se muestra su magnitud en el ejemplo utilizado en este capítulo. En la figura 9.2 se muestra al cuadro de diálogo donde se solicita el cálculo de tales estadísticos.

112. A los interesados en ampliar conocimientos sobre el tema recomendamos la lectura del trabajo de A. Camarero Rioja (2002), en especial las páginas 381-384 donde se presenta el origen y los problemas asociados a cada una de estas medidas.

A) *Phi:* (tablas de 2 * 2)

Datos tabla 9.1:

$$\varphi = \sqrt{\frac{X^2}{N}} = \sqrt{\frac{31{,}447}{180}} = 0{,}418$$

B) *Coeficiente de Contingencia o Coef. C de Pearson*:

Datos tabla 9.1:

$$C = \sqrt{\frac{X^2}{X^2 + N}} = \sqrt{\frac{31{,}447}{31{,}447 + 180}} = 0{,}3856$$

C) *V de Cramer*:

Datos tabla 9.1:

$$V = \sqrt{\frac{X^2}{N^* \text{ mínimo } (f-1) \text{ o } (c-1)}} = \sqrt{\frac{31{,}447}{180^*}} = 0{,}418$$

Cuadro 9.4. Estadísticos basados en el Chi-Cuadrado.

filas (o columnas[113]), de modo que en tablas de 2 x 2 este valor es 0,707 (√ [K-1/k]; √ [2-1/2]), en tablas de 3 x 3 llega a 0,81 (√ [K-1/K]; √ [3-1/3]), en tablas 4 x 4 de 0,87, y en tablas 5 x 5 el C máximo llega a 0,89 (García Ferrando, 1985: 225). El ratio C / $C_{máx}$ se interpreta de modo similar al coeficiente de correlación al cuadrado, esto es, el tanto por uno que C es de $C_{máx}$ (Calvo, 1990: 158). Esta medida, por su parte, presenta el problema que no es posible comparar tablas de diferentes tamaños.

Para solucionar el problema de la tablas rectangulares Cramer desarrollo el estadístico "V" que oscila entre 0 y 1, con independencia del tamaño de la tabla. Como puede apreciarse, el valor del *V de Cramer* en tablas cuadradas de 2 x 2 es el mismo que el obtenido por la *Phi*, como sucede en el cuadro 9.4 y en la tabla 9.2.

Veamos, mediante un ejemplo la mejora que suponen estas medidas respecto a la utilización del Chi–cuadrado, considerando para ello las tablas mostradas dentro de la tabla 9.4. La tabla A presenta un Chi–Cuadrado de 0,000, que indica la inexistencia de relación entre variables, por lo que no procede calcular el valor *Phi*. La tabla B tiene un Chi-Cuadrado de 4, un valor Phi de 0,2, con una significación de 0,046;

113. Es indiferente considerar el número de filas o de columnas puesto que se trata de una tabla cuadrada, una tabla que presenta igual número de filas que de columnas.

que indica una relación significativa *justa*, casi en el límite. La tabla "C", por su parte, presenta un Chi-Cuadrado de 88,395, con un valor Phi de 0,940 que es significativo al 0,000[114]. Un valor de Phi cercano a la unidad (o casi uno) indica que existe una gran relación entre sexo y tipo de ocio; esto es, que el beber o el bailar está muy relacionado con el sexo del entrevistado. Además, esto implica que el sexo es *determinante* en el tipo de ocio, que no influye ninguna otra variable. Estaremos de acuerdo que este valor de 0,94 es fruto de un ejemplo, que en el *mundo real* en difícil encontrar una relación de tal magnitud, una relación tan determinante. Por este motivo recomendamos –más que valorar el valor del coeficiente evaluando cuanto se aproxima a 0 o a 1– considerar la magnitud de un coeficiente en relación con el promedio; esto es, valorar si un coeficiente es el mayor o menor de todos los considerados.

Pese a las mejoras que suponen la utilización de estas medidas frente al uso del Chi-Cuadrado, Reynolds (1984: 34-49) señala los problemas que surgen a la hora de interpretar estos coeficientes: estas magnitudes se interpretan según su proximidad a 1 o 0, de modo que si están cercanas a uno la relación será importante, mientras que ésta será despreciable si están cercanas a 0. Esta es toda la información que suministran estos coeficientes, no siendo posible interpretarlos como la variación porcentual de una variable que es explicada por otra, ni reducción del error al predecir una variable mediante el conocimiento de la otra[115]. En palabras de este autor, carecen de una interpretación intuitiva: ¿Cómo interpretamos un valor de 0,29? Parece una relación débil pero no hay una medida que ayude a decidir sobre la debilidad de esta relación (Reynolds 1984: 49). Por otro lado, estas medidas se han desarrollado para solucionar algunas de las limitaciones del Chi-Cuadrado, y como éste todas son simétricas, no distinguen entre variables dependientes e independientes.

¿Y si el Chi–Cuadrado no muestra relación entre variables?, como ocurre en el ejemplo de la tabla 9.5. En este caso no se procede al cálculo de estos estadísticos. Una vez constatado que el Chi-Cuadrado no es significativo, el análisis de la tabla termina concluyendo que no existe relación entre *las actividades que más te gusta hacer (fuera de casa) cuando dispones de tiempo libre* y el hecho de *estudiar sociología u otra* carrera.

Buscando *fijar* los conocimientos aprendidos en esta sección, antes de considerar nuevos contenidos, proponemos varios ejercicios utilizando la investigación sobre *Vida Cotidiana*. El primero plantea si existe diferencia en el grado de felicidad (a54) con-

114. Debe tenerse en cuenta que en tablas de 2*2 no se interpreta el Chi-Cuadrado, sino la corrección de continuidad propuesta por Yates (cuadro 9.2 y tabla 9.7); que en estos casos es de 3,20 (significación 0,072) y de 84,67 (significación 0,000). Lo que indica que no existe relación significativa entre variables en la *tabla B*, y sí en la *tabla C*.

115. Volviendo al ejemplo anterior del coeficiente de correlación, el coeficiente de 0,8 antes apuntado está indicando que una variable explica un 64% (8 * 8 = 64) de la varianza de la otra. Esta interpretación no es posible con los coeficientes basados en el Chi-Cuadrado.

siderando el sexo (e9) de los entrevistados, esto es, si es posible afirmar que los hombres son más (o menos) felices que las mujeres. ¿Y respecto al estado civil (e12)?; ¿existe diferencia en el grado de felicidad considerando el estado civil de los entrevistados?

El tercer ejercicio propone considerar hasta que punto está relacionada la variable "clase social de pertenencia" (e26) con el nivel de equipamiento del hogar considerando conjuntamente todos los equipamientos. Nos referimos, concretamente, a la variable creada al final del apartado 8.7 utilizando la pregunta 20, que clasifica a la población entrevistada según el número de equipamientos presentes en su hogar.

4. Análisis del interior de la tabla

Una vez comprobado que existe relación entre variables se procede con el estudio del interior de la tabla con el fin de interpretar en que consiste esa relación. Se trata, siguiendo el ejemplo empleado a lo largo de este apartado, de señalar cuáles son las actividades de ocio que caracterizan a los hombres y a las mujeres. Para ello presentamos dos estrategias: cálculo e interpretación de porcentajes, e interpretación de la tabla utilizando *residuo*s.

4.1. Cálculo y diferencia de porcentajes

Volvamos de nuevo a la tabla 9.1 con el fin de analizar una de las formas más populares de interpretación de los valores de las tablas de contingencia. Basados en la idea que la interpretación de los números absolutos de las celdillas es complicada, la mayor parte de las veces se recurre al estudio de porcentajes, dejando en un lugar secundario el análisis de los números absolutos. Comenzando, por ejemplo, con el análisis de los hombres que emplean su tiempo de ocio en beber, aparecen dificultades al comparar las frecuencias absolutas de los 26 hombres que manifiestan este comportamiento (celdilla superior izquierda), con las 20 mujeres que también lo declaran (celdilla superior derecha); en la medida que cada columna tiene un distinto número de personas (70 y 110 respectivamente). Por este motivo no estamos interesados en el número absoluto puesto que para poder comparar ambas celdillas es necesario *ponderar* el número de respuestas de cada celdilla respecto al número de respuestas de esta categoría en la variable columna. Así las 26 entrevistados que declaran emplear su tiempo de ocio en beber son un 37,1% de todos los hombres entrevistados, mientras que las 20 mujeres que eligen esta misma opción son un 18,2% del total de mujeres.

Para solicitar los porcentajes en el programa SPSS será necesario, dentro del Cuadro de diálogo *Tablas de contingencia* mostrado en la figura 9.1, pulsar el recuadro *Casillas...* para obtener el cuadro de diálogo expuesto en la figura 9.3. Centraremos nuestra atención en la parte inferior izquierda titulada *Porcentajes*, que permite obtener tres tipos de porcentajes haciendo clic en el lugar correspondiente:

Figura 9.3. Botón *Casillas* dentro de cuadro de diálogo de la Tablas de contingencia. Mostrar en las casillas.

- Porcentajes de columna: calculados dividiendo la frecuencia absoluta de la celdilla entre el número de respuestas de cada una de las categorías de la variable columna (v049); tal y como se presenta en el primer ejemplo de la tabla 9.8. La variable situada en la columna se considera como independiente, y la colocada en la fila como dependiente.
  Cálculo de los porcentajes de la segunda columna: 20/110 = 0,182; 12/110 = 0,109; 8/110 = 0,073...
  Estos porcentajes se interpretan de la siguiente forma: el 18,2% de las mujeres señalan que la actividad que más les gusta hacer cuando disponen de tiempo libre es beber, porcentaje que se reduce al 10,9% cuando se trata de bailar, y al 7,3% respecto a hacer deporte.
  Si observamos la columna de los hombres se aprecia que el 37,1% elige beber cuando dispone de tiempo libre, y un 17,1% muestra su preferencia por hacer deporte. Es importante indicar que bailar no es elegida por ningún hombre.

Hasta ahora únicamente se han interpretado los porcentajes que componen cada columna, pero el análisis de tablas de contingencia se ve notablemente enriquecido al poder comparar *transversalmente* estos porcentajes, realizando comparaciones entre cada una de las celdas de las distintas columnas. Para esto es fundamental la observación de la columna total, que no es otra cosa que la frecuencia de la variable v01bis (el *marginal*). Esta columna es vital para la interpretación de tablas de contingencia puesto que permite localizar los porcentajes de cada categoría (hombres y mujeres) que se encuentran por encima y por debajo del valor marginal[116]. Así es posible apreciar que los hombres destacan (respecto del total) por sus mayores elecciones de beber (37,1% hombres y 25,6% total), hacer deporte (17,1% y 11,1%) y practicar alguna afición o hobby (17,1% y 11,1%); mientras que las mujeres eligen (más que el promedio) bailar (10,9% y 6,7%), ir al cine (9,1% y 5,6%) y otras[117] (16,4% y 11,1%). A la hora de señalar la preferencia por un tipo y otro de ocio únicamente deben considerarse las diferencias *importantes* de porcentajes. Cea D'Ancona (2004: 407), por ejemplo, considera que únicamente las diferencias superiores a cinco puntos pueden considerarse relevantes[118].

- Porcentajes de fila. Calculados con el mismo criterio pero considerando el porcentaje respecto a la variable fila (v01bis). En este caso La variable situada en la fila se considera como independiente, y la colocada en la columna como dependiente.

  Cálculo de la primera fila: 26/46 = 0,5652; 20/46 = 0, 4348.

  El segundo ejemplo de la tabla 9.8 muestra esta situación. La interpretación de estos datos se realiza de modo similar, pero teniendo que cuenta que la fila es el total. De aquellos que lo que más les gusta hacer en su tiempo libre es beber, el 56,5% son hombres y un 43,5% mujeres. De los que eligen hacer deporte, el 60% son hombres y un 40% mujeres. Del total de la muestra un 39% (exactamente 38,9%) son hombres y el 61% mujeres (exactamente 61,1%).

116. Se trataría, en definitiva, de elegir una magnitud que nos ayude a determinar "¿qué es ser alto?", "¿qué es ser rico?", "¿qué es ser gordo"... En la vida cotidiana normalmente definimos como altos a los que sobrepasan la "normalidad", esto es, el promedio de la población.
117. Téngase en cuenta que de las 20 respuestas obtenidas en la categoría "otras", 6 proceden de acudir a espectáculos (concretamente ir al teatro y a conciertos), y otras cuatro a "quedar" (con los amigos/as o con el novio/a).
118. Estas diferencias dependerán el tamaño muestral, puesto que con pequeños tamaños muestrales se produce un aumento del error típico de las estimaciones y una pérdida de significatividad de las diferencias porcentuales detectadas (Cea D'Ancona, 2004: 407). Reproducios una cita literal de López Pintor y Wert, por la precisión de la explicación: "...si la muestra tiene, como es usual, un margen de error del 3%, sólo tiene sentido empezar a pensar en diferencias por encima de los cuatro o cinco puntos de porcentaje, y en el caso de una distribución de frecuencias del conjunto de la muestra" (López Pintor y Wert, 2000: 537).

Tabla de contingencia (v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género

Porcentajes de columna:

			(v49) Genero		
			Hombre	**Mujer**	**Total**
(v01bis) Actividad, fuera de casa...	Beber, ir de copas	Recuento	26	20	46
		% de (v49) Género	[119]**37,1%**	18,2%	25,6%
	Bailar	Recuento	0	12	12
		% de (v49) Género	,0%	**10,9%**	6,7%
	Hacer deporte	Recuento	12	8	20
		% de (v49) Género	**17,1%**	7,3%	11,1%
	Ir de excursión y al monte	Recuento	6	6	12
		% de (v49) Género	8,6%	5,5%	6,7%
	Viajar	Recuento	12	28	40
		% de (v49) Género	17,1%	25,5%	22,2%
	Ir al cine	Recuento	0	10	10
		% de (v49) Género	,0%	**9,1%**	5,6%
	Practicar alguna afición o hobby	Recuento	12	8	20
		% de (v49) Género	**17,1%**	7,3%	11,1%
	Otras	Recuento	2	18	20
		% de (v49) Género	2,9%	**16,4%**	11,1%
Total		Recuento	70	110	180
		% de (v49) Género	100,0%	100,0%	100,0%

Tabla 9.8. Porcentajes de la tabla 9.1 (parte 1).

119. Se han resaltado en negrilla los porcentajes superiores a la columna "total".

Porcentajes de fila:

			(v49) Genero		
			Hombre	**Mujer**	**Total**
(v01bis) Actividad, f uera de casa...	Beber, ir de copas	Recuento	26	20	46
		% de v01bis	[122]**56,5%**	*43,5%*	100,0%
	Bailar	Recuento	0	12	12
		% de v01bis	,0%	**100,0%**	100,0%
	Hacer deporte	Recuento	12	8	20
		% de v01bis	**60,0%**	*40,0%*	100,0%
	Ir de excursión y al monte	Recuento	6	6	12
		% de v01bis	50,0%	50,0%	100,0%
	Viajar	Recuento	12	28	40
		% de v01bis	*30,0%*	**70,0%**	100,0%
	Ir al cine	Recuento	0	10	10
		% de v01bis	,0%	**100,0%**	100,0%
	Practicar alguna afición o hobby	Recuento	12	8	20
		% de v01bis	**60,0%**	*40,0%*	100,0%
	Otras	Recuento	2	18	20
		% de v01bis	*10,0%*	**90,0%**	100,0%
Total		Recuento	70	110	180
		% de v01bis	38,9%	61,1%	100,0%

Tabla 9.8. Porcentajes de la tabla 9.1 (parte 2).

122. Se han resaltado en negrilla los porcentajes superiores a la fila "total" (última línea). Los subrayados indican porcentajes notablemente inferiores al total.

Porcentajes respecto del total:

			(v49) Genero		
			Hombre	**Mujer**	**Total**
(v01bis) Actividad, fuera de casa...	Beber, ir de copas	Recuento	26	20	46
		% del total	14,4%	11,1%	25,6%
	Bailar	Recuento	0	12	12
		% del total	,0%	6,7%	6,7%
	Hacer deporte	Recuento	12	8	20
		% del total	6,7%	4,4%	11,1%
	Ir de excursión y al monte	Recuento	6	6	12
		% del total	3,3%	3,3%	6,7%
	Viajar	Recuento	12	28	40
		% del total	6,7%	15,6%	22,2%
	Ir al cine	Recuento	0	10	10
		% del total	,0%	5,6%	5,6%
	Practicar alguna afición o hobby	Recuento	12	8	20
		% del total	6,7%	4,4%	11,1%
	Otras	Recuento	2	18	20
		% del total	1,1%	10,0%	11,1%
Total		Recuento	70	110	180
		% del total	38,9%	61,1%	100,0%

Tabla 9.8. Porcentajes de la tabla 9.1 (parte 3).

Con todos los porcentajes disponibles:

			(v49) Genero		
			Hombre	**Mujer**	**Total**
(v01bis) Actividad, fuera de casa...	Beber, ir de copas	Recuento	26	20	46
		% de v01bis	56,5%	43,5%	100,0%
		% de v49 Género	37,1%	18,2%	25,6%
		% del total	14,4%	11,1%	25,6%
	Bailar	Recuento	0	12	12
		% de v01bis	,0%	100,0%	100,0%
		% de v49 Género	,0%	10,9%	6,7%
		% del total	,0%	6,7%	6,7%
	Hacer deporte	Recuento	12	8	20
		% de v01bis	60,0%	40,0%	100,0%
		% de v49 Género	17,1%	7,3%	11,1%
		% del total	6,7%	4,4%	11,1%
	Ir de excursión y al monte	Recuento	6	62	
		% de v01bis	50,0%	50,0%	100,0%
		% de v49 Género	8,6%	5,5%	6,7%
		% del total	3,3%	3,3%	6,7%
	Viajar	Recuento	12	28	40
		% de v01bis	30,0%	70,0%	100,0%
		% de v49 Género	17,1%	25,5%	22,2%
		% del total	6,7%	15,6%	22,2%
	Ir al cine	Recuento	0	10	10
		% de v01bis	,0%	100,0%	100,0%
		% de v49 Género	,0%	9,1%	5,6%
		% del total	,0%	5,6%	5,6%
	Practicar alguna afición o hobby	Recuento	12	8	20
		% de v01bis	60,0%	40,0%	100,0%
		% de v49 Género	17,1%	7,3%	11,1%
		% del total	6,7%	4,4%	11,1%
	Otras	Recuento	2	18	20
		% de v01bis	10,0%	90,0%	100,0%
		% de v49 Género	2,9%	16,4%	11,1%
		% del total	1,1%	10,0%	11,1%
Total		Recuento	70	110	180
		% de v01bis	38,9%	61,1%	100,0%
		% de v49 Género	100,0%	100,0%	100,0%
		% del total	38,9%	61,1%	100,0%

Tabla 9.8. Porcentajes de la tabla 9.1 (parte 4).

El análisis *transversal*, que en este caso será comparando las filas con el total de fila (puesto que se trata de porcentajes de fila), desvela que los hombres emplean fundamentalmente su tiempo libre en beber (56,5% y 38,9%), hacer deporte (60% y 38,9%) y practicar alguna afición o hobby (60% y 38,9%); mientras que las actividades de ocio más elegidas por las mujeres son bailar (100% y 61,1%), viajar (70% y 61,1%), ir al cine (100% y 61,1%), y otras (75% y 61,1%). Ir de excursión y al monte es la única actividad que no presenta ninguna diferencia por sexo. Evidentemente cada investigador puede pedir la opción que le sea más cómoda, puesto que bastará con cambiar la posición de cada variable en las filas y en las columnas para que la tabla cambie de sentido. Si optamos por los porcentajes de fila, pero queremos obtener una tabla para comparar los hábitos de hombres y mujeres (v01bis) bastará con colocar esta variable en las filas, y la otra en las columnas en el cuadro de diálogo de la figura 9.1.

- Porcentajes totales calculados dividiendo el número de respuestas en cada celdilla entre el total de la tabla.
  Cálculo de la segunda columna: 20/180 = 0,111; 12/180 = 0,067; 8/180 = 0,044...
  El tercer ejemplo de la tabla 9.8 se interpreta utilizando conjuntamente la información de las filas y las columnas, es decir, un 14,4% de todos los entrevistados señalan que la actividad fuera de casa que más les gusta hacer cuando disponen de tiempo libre es beber, un 5,6% (de todos los entrevistados) apuestan por ir al cine, etc. En este caso el porcentaje "total de filas" corresponde a la frecuencia de filas (ver similitud con el primer ejemplo presentado en la tabla 9.8), mientras que el porcentaje total de columnas son las frecuencias de la variable columna.

El cálculo del porcentaje, como se ha expuesto, es tremendamente sencillo y no reviste complicación alguna. Algo más difícil es la decisión sobre que porcentajes son necesarios para realizar la interpretación de los datos. El criterio para esta decisión está condicionada por la formulación de la hipótesis utilizada, fruto de la *pregunta* que da lugar a la investigación o al marco teórico elegido. En este ejemplo, utilizado para conocer las actividades de ocio que caracterizan a los hombres y las mujeres, consideramos que resulta más adecuada la utilización de los porcentajes de columna en la medida que *elimina* el diferente número de hombres y mujeres seleccionadas. Aunque los resultados son muy similares, utilizar el porcentaje de filas puede generar algunas apreciaciones incorrectas[121], como tendremos ocasión de demostrar en el siguiente apartado.

121. Nos referimos, concretamente, a la actividad "viajar" que ha sido resaltada en el comentario de la tabla de los porcentajes de filas, pero no así en la de columnas.

Algunos investigadores (entre otros Cea D'Ancona, 2004: 407) recomiendan utilizar los porcentajes de columna[122] considerando su mayor facilidad de lectura; basados en el hecho que –en la cultura occidental– la lectura se realiza en sentido horizontal. Los porcentajes verticales, calculados respecto a la variable situada en la columna, se comparan entre ellos horizontalmente (de la misma forma en que se realiza la lectura). Otros preferimos los porcentajes de fila por la facilidad de comparar los porcentajes *de arriba a abajo* (en sentido vertical), así como por la *ausencia de limitación* del número de categorías presentes en la variable independiente. Ciertos investigadores colocan en columnas la variable con menos categorías, indicando en cada tabla el sentido de los porcentajes. A nuestro juicio esta opción puede desconcertar al lector al tener que cambiar –en cada tabla– el criterio de lectura. Se opte por uno u otro, siempre será necesario indicar claramente en la tabla el sentido de los porcentajes.

Lo que no recomendamos, bajo ningún concepto, es solicitar todos los porcentajes en una sola tabla, aún cuando tal solicitud se realice con el fin de decidir después cual utilizar. Solicitar los porcentajes de columna, fila y total dificulta tremendamente la interpretación de la tabla; como puede apreciarse en el análisis de la parte última de la tabla 9.8.

Antes de proceder con nuevos conocimientos recomendamos a los lectores interpretar el *interior* de las tablas elaboradas de los ejercicios realizados al final de la sección 9.3, esto es, relación entre el grado de felicidad (a54) y el estado civil (e12); grado de felicidad y sexo (e9); y relación entre clase social de pertenencia (e26) y nivel de equipamiento del hogar.

4.2. Interpretación de los valores de la tabla utilizando los residuos

El análisis de los *residuo*s supone una excelente opción para la interpretación de *tablas de contingencia*, y llama la atención la escasa utilización de esta estrategia en las investigación actual. Comenzaremos la exposición olvidándonos –por un momento– de los valores de las celdillas, considerando únicamente los *marginales* de la tabla. Nuestro interés es conocer cuales hubieran sido los valores de cada celdilla si únicamente conociéramos los valores marginales o, dicho de otro modo, qué valores podrían *esperarse* al considerar las filas y las columnas de la tabla. En esta situación podríamos calcular los valores de la celdilla multiplicando la frecuencia de la fila por la frecuencia de la columna, para dividirlo todo entre el total de la tabla; tal y como

122. Siempre que esta variable no presente un gran número de categorías de respuesta.

se expuso en la tabla 9.3. Recuérdese que estos valores se conocen con el nombre de *frecuencias esperadas* (o frecuencias teóricas), y –como ya indicamos– son las frecuencias que tendrían las celdillas si no existiera relación entre variables. Estas frecuencias esperadas se solicitan en el SPSS marcando la opción *Frecuencia esperada* en la parte superior izquierda del cuadro de diálogo mostrado en la figura 9.3.

Se ha afirmado que si no existiera ninguna relación entre las variables que forman la tabla las celdillas tendrían el valor de las frecuencias esperadas, es decir la diferencia entre frecuencias esperadas y observadas sería cero. Esto significa que cuanto mayor sea la diferencia entre las frecuencias esperadas y las obtenidas la relación entre las variables será mayor. La diferencia entre estas magnitudes recibe varios nombres, aunque en el ámbito de las tablas de contingencia se le conoce como *residuos*. El *residuo* es la diferencia entre la frecuencia esperada y la observada, tal y como se aprecia en los cálculos del cuadro 9.5 y en los resultados mostrados en la tercera fila de cada una de las celdillas de la tabla 9.9. Un *residuo* positivo significa que la frecuencia observada es mayor que la esperada, mientras que cuando es negativo indica lo contrario.

El problema al que nos enfrentamos ahora es cuantificar el valor 8,1 del *residual* de la celdilla de la primera fila y primera columna de la tabla 9.9 (marcado en negrilla) que, como se ha dicho, es la diferencia entre la frecuencia observada y esperada de la citada celdilla (26 – 17,9). Se trata de comparar los valores de cada *residuo* puesto que –como se aprecia en esta tabla– existe una gran variabilidad en el tamaño de cada uno. Supongamos que tuviéramos otra celdilla con un *residual* similar, pero que fuera obtenido de la resta 100008,1 - 100000. ¿Pueden interpretarse igual ambas magnitudes? En ambos casos la diferencia es la misma, en el primer caso la diferencia de 8,1 entre 26 y 18 es importante, pero en el segundo la diferencia de 8 entre 100008 y 100000 es ridícula. Es por ello por lo que es aconsejable utilizar los *residuo*s eliminando el efecto que puedan tener los marginales sobre su valor, residuos *estandarizados* según su frecuencia esperada. Son denominados como residuos *tipificados* o *estandarizados*. El análisis de la cuarta magnitud de cada una de las celdillas de la tabla 9.1 muestra un *alisamiento* de los valores de los residuos al ajustarlos, alisamiento que varía según el número de casos en los que se fundamenta cada residuo. Al eliminar el influjo del tamaño muestral ya es posible realizar comparaciones entre ellos: la mayor diferencia se produce entre las opciones bailar (residuo tipificado de hombres –2,2 y de mujeres 1,7) e ir al cine con valores *residual*es tipificados en los hombres de –2,0 y de mujeres de 1,6. Estas actividades son las que más diferencian y más definen las actividades de ocio (fuera del hogar) de hombres y mujeres.

Una mejor solución propone Haberman (1973) al dividir el residuo tipificado entre la raíz cuadrada de la varianza del residuo, calculada tal y como se muestra en la última parte del cuadro 9.5. Estos residuos se interpretan como cualquier

1.- **Frecuencias esperadas (o teóricas)** (FE)**:**

Columna 1:

70 * 46 / 180 = 17,9
70 * 12 / 180 = 4,7
70 * 20 / 180 = 7,8
70 * 12 / 180 = 4,7
70 * 40 / 180 = 15,6
70 * 10 / 180 = 3,9
70 * 20 / 180 = 7,8
70 * 20 / 180 = 7,8

2.- **Residuos** (FO-FT):

Columna 1:

26 - 17,9 = **8,1**
0 - 4,7 = -4,7
12 - 7,8 = 4,2
6 - 4,7 = 1,3
12 - 15,6 = -3,6
0 - 3,9 = -3,9
12 - 7,8 = 4,2
8 - 7,8 = -5,8

3.- **Residuos tipificados** (**estandarizados**): Res / √ FT

Columna 1:

8,1 / √ 17,9 = 1,9
-4,7 / √ 4,7 = -2,2
4,2 / √ 7,8 = 1,5
1,3 / √ 4,7 = 0,6
-3,6 / √ 15,6 = -0,9
-3,9 / √ 4,7 = -2,0
4,2 / √ 7,8 = 1,5
-5,8 / √ 7,8 = -2,1

4.- **Residuos tipificados ajustados o corregidos**:

Std Res / √ V_{ij} (Haberman, 1973: 215).

V_{ij} es una estimación de la varianza de e_{ij}, calculada con la expresión

V_{ij}: [1- (FO_i / n)] * [1- (FO_j / n)]

Donde:

FO_i: total de fila, frecuencia observada de fila
FO_j: total de columna, frec. observada de columna
n : tamaño muestral.

Cálculo:

(1 - [70/180]) * (1- [46/180]) = 0,4549
(1 - [70/180]) * (1- [12/180]) = 0,5704
(1 - [70/180]) * (1- [20/180]) = 0,5432
(1 - [70/180]) * (1- [12/180]) = 0,5703
(1 - [70/180]) * (1- [40/180]) = 0,4753
(1 - [70/180]) * (1- [10/180]) = 0,5772
(1 - [70/180]) * (1- [20/180]) = 0,5432
(1 - [70/180]) * (1- [20/180]) = 0,5432

Cálculo de residuos ajustados:

1,9 / √ 0,4549 = 2,81
-2,2 / √ 0,5704 = -2,91
1,5 / √ 0,5432 = 2,09
0,6 / √ 0,5703 = 0,79
-0,9 / √ 0,4753 = 1,30
-2,0 / √ 0,5772 = -2,63
1,5 / √ 0,5432 = 2,09
2,1 / √ 0,5432 = 2,85

Nota: a fin de simplificar los cálculos se ha calculado únicamente en la columna de los hombres. El proceso es elmismo para el resto de las celdillas.

Cuadro 9.5. Cálculo de los componentes de una tabla de contingencia (ejemplo con la tabla 9.1).

Tabla de contingencia (v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género

			(v49) Genero		
			Hombre	**Mujer**	**Total**
(v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre	Beber, ir de copas	Recuento	26	20	46
		Frec. esperada	17,9	28,1	46,0
		Residuo	**8,1**	-8,1	
		Residuo tipificado	1,9	-1,5	
		Residuo corregido	2,8	-2,8	
	Bailar	Recuento	0	12	12
		Frec. esperada	4,7	7,3	12,0
		Residuo	-4,7	4,7	
		Residuo tipificado	-2,2	1,7	
		Residuo corregido	-2,9	2,9	
	Hacer deporte	Recuento	12	8	20
		Frec. esperada	7,8	12,2	20,0
		Residuo	4,2	-4,2	
		Residuo tipificado	1,5	-1,2	
		Residuo corregido	2,1	-2,1	
	Ir de excursión y al monte	Recuento	6	6	12
		Frec. esperada	4,7	7,3	12,0
		Residuo	1,3	-1,3	
		Residuo tipificado	,6	-,5	
		Residuo corregido	,8	-,8	
	Viajar	Recuento	12	28	40
		Frec. esperada	15,6	24,4	40,0
		Residuo	-3,6	3,6	
		Residuo tipificado	-,9	,7	
		Residuo corregido	-1,3	1,3	
	Ir al cine	Recuento	0	10	10
		Frec. esperada	3,9	6,1	10,0
		Residuo	-3,9	3,9	
		Residuo tipificado	-2,0	1,6	
		Residuo corregido	-2,6	2,6	
	Practicar alguna afición o hobby	Recuento	12	8	20
		Frec. esperada	7,8	12,2	20,0
		Residuo	4,2	-4,2	
		Residuo tipificado	1,5	-1,2	
		Residuo corregido	2,1	-2,1	
	Otras	Recuento	2	18	20
		Frec. esperada	7,8	12,2	20,0
		Residuo	-5,8	5,8	
		Residuo tipificado	-2,1	1,7	
		Residuo corregido	-2,8	2,8	
Total	Recuento	70	110	180	
	Frec. esperada	70,0	110,0	180,0	

Tabla 9.9. Ejemplo de tabla con todas las opciones disponibles en la opción *Casillas*.

valor de una variable tipificada (estandarizada) con una distribución normal: un valor superior a +1,96, o inferior a –1,96, indica que hay relación entre ambas categorías a un *nivel de confianza* del 95%, y el +/-2,58[123] indica que existe relación a un nivel de confianza del 99% (Haberman 1973: 216-218). De este modo cuanto mayor es el valor del residuo mayor es la diferencia, indicando el signo la dirección de la relación.

De la tabla 9.9 se extrae la 9.10 donde se muestra en cada celdilla únicamente el número de casos y los residuos tipificados corregidos. Basta con un breve vistazo a esta tabla para detectar las mayores relaciones, aquellas celdillas donde se encuentran los residuos con mayor valor (siempre que sean superiores a 1,96 y menores de –1,96, que son los umbrales de *significación* al 95%). En la tabla 9.10 los residuos significativos se han destacado en negrilla para facilitar la interpretación.

Los dos primeras filas presentan las magnitudes más altas, lo que indica una gran relación entre variables, relación que alcanza su punto álgido entre los entrevistados que eligen bailar como la actividad fuera de casa que más hacen en su tiempo libre; con un valor –2,9 para los hombres y de 2,9 para las mujeres[124]. El valor de estos *residual*es está indicando que existe una gran relación (positiva) entre las mujeres y bailar, y negativa entre los hombres y bailar. Dicho de otro modo, las mujeres destacan por emplear su tiempo de ocio bailando, mientras que los hombres destacan por las pocas elecciones en esta actividad. El alto valor positivo del residuo correspondiente a la celdilla hombres y beber está indicando que los hombres destacan por emplear el ocio en esta actividad, caso contrario al de las mujeres.

Similar interpretación cabe hacer de la práctica del deporte y de practicar alguna actividad o hobby, actividades elegidas fundamentalmente por el colectivo masculino. Las mujeres, además de bailar, destacan también por acudir al cine. En definitiva, una interpretación idéntica a la realizada en el comentario a la tabla 9.8 (tabla con porcentajes de columna y fila), si bien –desde mi punto de vista– con una mayor sencillez puesto que no hay que decidir el *tipo* de porcentajes a utilizar, como compararlos, etc. Es importante destacar la nula referencia a viajar, puesto que presenta unos residuos corregidos no significativos. Recuérdese, como se señaló en la nota a pie número 20, que esta actividad únicamente apareció en el comentario de los porcentajes de filas, y no en los porcentajes verticales (de columnas).

---

123. Son los *valores críticos* estadístico que indican las zonas de significación de la curva normal, con un nivel de confianza del 95 y del 99% respectivamente (Everitt y Wykes, 2001: 211).
124. Téngase en cuenta que en el caso de una tabla con dos columnas, como ocurre en este ejemplo, los residuos se *contraponen* con el fin de sumar 0. El análisis de los residuos corregidos es más interesante con tablas mayores, en aquellas donde el análisis de los porcentajes resulta más complicado.

Tabla de contingencia (v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género

			(v49) Genero		
			Hombre	**Mujer**	**Total**
(v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre	Beber, ir de copas	Recuento	26	20	46
		Residuo corregido	**2,8**	**-2,8**	
	Bailar	Recuento	0	12	12
		Residuo corregido	**-2,9**	**2,9**	
	Hacer deporte	Recuento	12	8	20
		Residuo corregido	**2,1**	**-2,1**	
	Ir de excursión y al monte	Recuento	6	6	12
		Residuo corregido	,8	-,8	
	Viajar	Recuento	12	28	40
		Residuo corregido	-1,3	1,3	
	Ir al cine	Recuento	0	10	10
		Residuo corregido	**-2,6**	**2,6**	
	Practicar alguna afición o hobby	Recuento	12	8	20
		Residuo corregido	**2,1**	**-2,1**	
	Otras	Recuento	2	18	20
		Residuo corregido	**-2,8**	**2,8**	
Total	Recuento	70	110	180	

Tabla 9.10. Ejemplo de tabla con residuos tipificados corregidos.

Antes de terminar con la explicación de los *residual*es señalar que el procedimiento propuesto por Haberman ha sido utilizado en la investigación con encuesta en nuestro país en los trabajos de Felman y otros (1988), González (1994), Ayerdi (1995) y Díaz de Rada (2004).

Del cuadro de diálogo principal mostrado en la figura 9.1 tan sólo resta explicar el botón *Exactas...*, que supera ampliamente los objetivos aquí planteados, y *Formato...*, que está referido al orden de presentación de la tabla. El cuadro de diálogo que surge tras pulsar este botones se muestra en la figura 9.4 y permite presentar las filas en orden ascendente o descendente; siempre considerando la codificación de la variable colocada en filas (no el número de casos contados). Por defecto aparece la presentación en orden ascendente (figura 9.4); esto es del valor inferior al superior, que es como se han mostrado las tablas presentadas hasta el momento: recuérdese que la opción *beber ir de copas* estaba codificada con el valor 1, *bailar* con el 2, etc. Por último, y vol-

viendo a la figura 9.1, obsérvese que en la parte inferior izquierda aparecen dos opciones que permiten *suprimir tablas*, y *mostrar los gráficos de barras agrupadas*. Disponer de gráficos supone una notable ayuda en la interpretación de la tabla de contingencia; si bien consideramos que son más interesantes los gráficos mostrados desde el menú gráficos. Por ello recomendamos no ejecutar los gráficos desde aquí puesto que se realizan en base a las frecuencias observadas (números absolutos), y no permite la utilización de porcentajes. Estos motivos nos llevan a recomendar el menú Gráficos (mostrado en la figura 4.9); que además presenta un mayor número de dispositivos gráficos.

Fijar los conocimientos aprendidos es el objetivo de toda actividad docente. Con el fin de consolidar lo aprendido recomendamos interpretar el interior de la tabla de los ejercicios planteados al final de la sección anterior. Se trataba, concretamente, de:

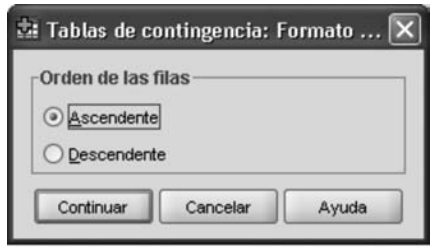

Figura 9.4. Botón *Formato* dentro de cuadro de diálogo de la Tablas de contingencia. Mostrar interior de la tabla.

- Relación entre grado de felicidad (a54) y sexo (e9); entre grado de felicidad (a54) y estado civil (e12).
- Relación entre clase social de pertenencia (e26) y "nivel de equipamiento del hogar" (variable creada en el apartado 8.7 en base a la pregunta 20).

5. Utilización de test estadísticos para conocer la relación entre variables ordinales

Todos los ejemplos presentados hasta ahora se han referido a variables nominales, buscando satisfacer los objetivos de la investigación planteada; que recordemos se fundamentaba en conocer las actividades de ocio (fuera del hogar) que caracterizan a los hombres y a las mujeres. Resuelta esa cuestión, en este momento plantea-

mos si existe relación entre el número de libros leídos en los últimos cinco meses no relacionados con los estudios (v08) y los libros leídos relacionados con los estudios (v06). Es decir, se trata de determinar si los entrevistados que presentan un mayor índice de lectura de libros (en general) son también los que más leen libros relacionados con tus estudios, o si más bien se trata del efecto contrario: si los que leen todo tipo de libros presentan un menor nivel de lectura de libros relacionados con los estudios. De modo que la *pregunta de investigación* que guiará la exposición en este apartado propone lo siguiente: ¿podríamos decir que los entrevistados que más leen todo tipo de libros[125] son también los que leen más libros relacionados con sus estudios?; o –más bien– al leer otros libros reducen la lectura de textos relacionados con sus estudios. (Dicho de otro modo, ¿existe relación entre el número de libros *relacionados* con los estudios y la lectura de libros *generalistas*?).

Las frecuencias de ambas variables, mostradas en la tabla 9.11, desvelan que un 20% de los entrevistados no leen ningún libro, considerando tanto los relacionados con los estudios como los no relacionados. Las frecuencias de la izquierda dan cuenta de los libros leídos relacionados con los estudios, y muestra que uno de cada cuatro entrevistados lee un libro, un 15% dos libros, y el 14% tres libros. Es interesante utilizar la columna porcentaje acumulado para conocer el número de entrevistados que leen tres y menos libros, estrategia que requiere tener en cuenta que la categoría *ninguno* aparece en la primera fila de la tabla; de modo que será necesario restar al porcentaje acumulado (73,8%) el porcentaje de personas que no han leído ningún libro (20,4%). De este modo obtenemos que el 53,4% de los entrevistados (73,8 – 20,4) han leído menos de cuatro libros.

Es posible emplear una estrategia que evite realizar esta resta, colocando la categoría *ninguno* en la parte superior de la distribución, esto es, codificándola con un valor mayor al valor más alto de la distribución. En el caso de v06, por ejemplo, podríamos codificar esta categoría con el valor 50[126], de modo que este categoría quedará situada por encima del 15, facilitando la lectura del porcentaje acumulado. Sin embargo, no recomendamos esta práctica puesto –como indicamos en la sección 3.4– "la medición ordinal debe respetar las relaciones observadas en la asignación del sistema de medición, ordenando los números según su orden serial". Codificar la categoría el ninguno con el valor 50, además de *romper* el orden de la distribución (al pasar del 15 al 50), altera su orden serial que –recordemos– implica que los valores más altos de la distribución son –en este caso– los que más libros leen.

125. A partir de ahora, y con el fin de simplificar la exposición, nos referiremos a éstos como *libros de todo tipo* o libros *generalistas*.

126. Indicando, en el procedimiento *Recodificar en distintas variables*", que el valor 0 es igual a 50. Posteriormente habría que *etiquetar* el valor 50 con la opción *ninguno*.

V06 Libros relacionados con tus estudios leídos en los últimos cinco meses				v08 libros leídos en los últimos cinco meses no relacionados con tus estudios (libros generalistas)			
	Frec.	Porcent	Porcent acum		Frec.	Porcent	Porcent acum
Ninguno	39	20,4	20,4	Ninguno	38	19,9	19,9
1	48	25,1	45,5	1	34	17,8	37,7
2	28	14,7	60,2	2	32	16,8	54,5
3	26	13,6	73,8	3	34	17,8	72,3
4	12	6,3	80,1	4	10	5,2	77,5
5	8	4,2	84,3	5	6	3,1	80,6
6	12	6,3	90,6	6	4	2,1	82,7
7	4	2,1	92,7	10	10	5,2	88,0
8	4	2,1	94,8	11	2	1,0	89,0
10	6	3,1	97,9	12	2	1,0	90,1
11	2	1,0	99,0	14	2	1,0	91,1
15	2	1,0	100,0	15	6	3,1	94,2
			20	2	1,0	95,3	
			20	2	1,0	96,3	
				No responde	7	25,1	100,0
Total	191	100,0		Total	191	100,0	

Tabla 9.11. Frecuencias de v06 y v08.

El análisis de v08, donde se recoge el número de libros leídos no relacionados con los estudios (libros generalistas), presenta 7 personas que no responden, de modo que debe definirse ese valor como *perdido* con el fin de poder interpretar adecuadamente el porcentaje de respuestas obtenidas. En la tabla 9.12 se presenta la tabla de frecuencias sin considerar las no respuestas, que muestran unas cifras ligeramente superiores a las mostradas en la variable v06: un 18,5% de los entrevistados ha leído un libro generalista, el 17% dos, y el 18% tres. El análisis del porcentaje acumulado (restando el porcentaje de la categoría *ninguno*) desvela que un 54% de los entrevistados han leído menos de cuatro libros generalistas. Esta similitud en los porcentajes nos lleva a sospechar que los entrevistados que presentan un mayor índice de lectura de libros (en general) son

(v08) Libros leídos en los últimos cinco meses no relacionados con los estudios

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Ninguno	38	19,9	20,7	20,7
	1	34	17,8	18,5	39,1
	2	32	16,8	17,4	56,5
	3	34	17,8	18,5	75,0
	4	10	5,2	5,4	80,4
	5	6	3,1	3,3	83,7
	6	4	2,1	2,2	85,9
	10	10	5,2	5,4	91,3
	11	2	1,0	1,1	92,4
	12	2	1,0	1,1	93,5
	14	2	1,0	1,1	94,6
	15	6	3,1	3,3	97,8
	20	2	1,0	1,1	98,9
	24	2	1,0	1,1	100,0
	Total	184	96,3	100,0	
Perdidos	No responde	7	3,7		
Total	191	100,0			

Tabla 9.12. Frecuencias de v08 considerando las no respuestas (valor 99) como valor perdido.

también los que más leen libros relacionados con tus estudios. No obstante, será necesario utilizar el cruce de tablas para dar respuesta a esta hipótesis.

Buscando complicar aún más nuestra exposición –y con el fin de repasar procedimientos vistos anteriormente– supondremos que el demandante de la investigación ha indicado que desea conocer la relación de v06 con v08 únicamente entre los entrevistados que menos leen, es decir, en aquellos que leen 0, 1, 2 y 3 libros.

Para realizar los análisis en este colectivo utilizaremos la selección de casos mediante criterios condicionales que fue explicada en la sección 8.8. Tras pulsar *Datos⇒Seleccionar Casos* se marca la opción *Si se satisface la condición* (figura 8.20) y aparece el cuadro de diálogo que se muestra en la figura 9.5. Seleccionada la variable v06, se desplaza a la ventana central para añadir los símbolos que indican *menor o igual*

a tres. Posteriormente se incluye el operando y, se desplaza la variable v08 a la ventana central, y se indica –en este caso– que seleccione los valores menores al cuatro[127]. Pulsando el botón *Continuar* y *Aceptar* se lleva a cabo la selección de los entrevistados que leen menos de cuatro libros.

Figura 9.5. Cuadro de diálogo *Seleccionar casos*, condición lógica con dos términos.

Será necesario solicitar las frecuencias de ambas variables para conocer cómo la selección efectuada afecta a ambas distribuciones. En la tabla 9.12 puede apreciarse una reducción en el tamaño muestral de 191 a 112 casos, así como los escasos cambios (en ambas variables) en la categoría *ninguno*. Comparando la distribución obtenida con la mostrada en la tabla 9.11 podemos apreciar –en v06– una disminución en todas las categorías: los entrevistados que no leen ningún libro descienden de 39 a 36, aquellos que leen un libro se reducen de 48 a 42, los que leen dos libros de 28 a 18, y los lectores de tres libros de 26 a 16. Menores cambios se producen en v08 puesto que el número de entrevistados que no lee ningún libro se mantiene estable, y se reduce ligeramente aquellos que leen un libro (de 34 a 32). En el resto de categorías de v08 se produce una reducción similar a la experimentada en v06: el número de los entrevistados que leen dos libros disminuye de 32 a 20, y los que leen tres libros de 34 a 22.

Hemos prestado atención al cambio entre ambas distribuciones para que el lector reflexione sobre las personas que no han sido seleccionadas, que son otros sino aquellos que leen muchos libros *generales* (altos valores en v08) y –a la vez– leen muchos libros relacionados con los estudios (altos valores en v06). Se trata de un colectivo de 77 entrevistados (191 – 112), un 40,3% (77 / 191 * 100) de la muestra original.

El demandante de la investigación indica que su interés de centra en conocer los que no leen ningún libro, los que leen uno, y –de forma agregada– los que leen dos

127. Seleccionar *menor o igual a tres* y *menor que cuatro* proporciona los mismos resultados puesto que el valor cuatro no está incluido en ninguna de las dos instrucciones. Se ha seleccionado *menor que cuatro* en v08 para mostrar el mayor número de recursos susceptibles de ser utilizados.

V06 Libros relacionados con tus estudios leídos en los últimos cinco meses				v08 libros leídos en los últimos cinco meses no relacionados con tus estudios			
	Frec.	Porcent.	Porcent acum		Frec.	Porcent valid	Porcent acum
Ninguno	36	32,1	32,1	Ninguno	38	33,9	33,9
1	42	37,5	69,6	1	32	28,6	62,5
2	18	16,1	85,7	2	20	17,9	80,4
3	16	14,3	100,0	3	22	16,9	100,0
Total	112	100,0		Total	112	100,096,3	
				No responde	7	25,1	100,0
Total	191	100,0		Total	191	100,0	

Tabla 9.13. Frecuencias de v06 y v08, seleccionados los entrevistados que leen tres y menos libros.

y tres libros. Aunque la experiencia investigadora recomienda siempre presentar los resultados de la forma más desagregada posible, siguiendo las indicaciones del demandante de la investigación se han agrupado las personas que leen dos y tres libros en (tanto en v06 como en v08). No reproducimos la tabla por motivos de espacio, y porque el lector puede construir fácilmente la tabla uniendo las categorías 2 y 3 de la tabla 9.13.

Con el fin de resolver el objetivo planteado[128] será necesario realizar una *tabla de contingencia* entre ambas variables siguiendo las instrucciones presentadas en el apartado 9.2, pero esta vez colocando v08 en columnas y v06 en filas. Solicitaremos también los porcentajes de columnas y los residuos corregidos. Posteriormente será necesario seleccionar las medidas de asociación precisas para conocer la relación entre ambas variables.

Antes de elegir las medidas de asociación para conocer la relación entre las variables será preciso preguntarnos por la escala de medida de éstas, que como se ha expuesto anteriormente (capítulo III) se trata de variables ordinales. Hay varias medidas que permiten analizar la relación entre variables ordinales, si bien aquí únicamente expondremos los más utilizados: *Gamma*, *d de Somer*, *Tau-b de Kendall* y *Tau-c de Kendall*. Basta con observar de nuevo el cuadro de diálogo *Estadísticos...*, presentado

128. Descubrir si existe relación entre el número de libros relacionados con los estudios, y la lectura de libros generalistas.

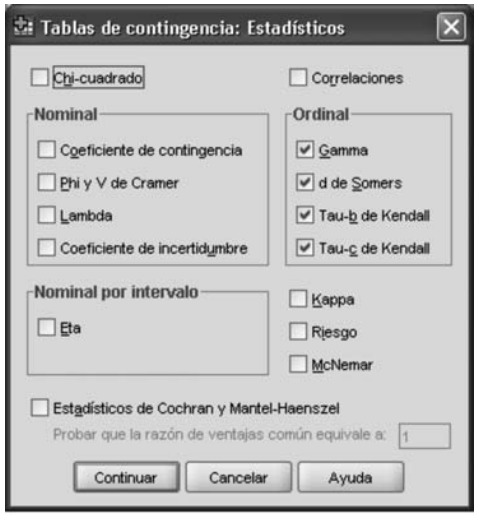

Figura 9.6. *Estadísticos* de tablas de contingencia para variables ordinales.

en la figura 9.6, para constatar que éstos se utilizan con variables ordinales. La tabla 9.14 muestra los resultados obtenidos para el cruce de v06 y v08.

El objetivo de la utilización de estadísticos para variables ordinales es analizar la relación entre la clasificación de cada individuo en cada una de las variables utilizadas; de modo que existirá relación cuando la distribución de los casos de la primera variable permita predecir la ordenación de los casos en la segunda variable (García Ferrando, 1985: 244). Nuestro objetivo se centra más en la interpretación de los coeficientes obtenidos que el cálculo de los mismos, de modo que –como en anteriores ocasiones–centraremos la exposición en la interpretación de los coeficientes, recomendando al lector interesado en los cálculos la consulta del anexo 1 donde se expone el cálculo de pares. La formulación de cada coeficiente se muestra en el cuadro 9.6.

El primero de los estadísticos solicitados, el coeficiente *Gamma* de Goodman y Kruskal es una medida simétrica que va desde –1 a 1 y se calcula restando al número de pares concordantes (N_c) los pares discordantes (N_d), y dividiendo este resultados entre la suma de ambos. Del cálculo de pares[129] se desprende que cuando todos los pares son concordantes existirá una asociación positiva entre las variables, puesto que los entrevistados que leen más libros relacionados con tus estudios serán tam-

129. Ver el anexo 1 donde se realiza una introducción al cálculo de pares.

A) Gamma de Goodman y Kruskal:

$$\text{Gamma} = \frac{N_c - N_d}{N_c + N_d} = \frac{1.828 - 1.064}{1.828 + 1.064} = 0{,}264$$

B) Tau-b de Kendall:

$$T_b = \frac{N_s - N_d}{\sqrt{(N_s + N_d + T_y)\,(N_s + N_d + T_x)}} =$$

$$T_b = \frac{1.828 - 1.064}{\sqrt{(1.828 + 1.064 + 1.264)\,(1.828 + 1.064 + 1.272)}} = 0{,}1836$$

donde T_x = son los pares empatados en la variable x, y
T_y = son los pares empatados en la variable Y.

C) Tau-c de Kendall:

$$T_c = \frac{2 * m * (N_s - N_d)}{N^2 * (m-1)} = \frac{2 * 3 * (1.828 - 1.064)}{112^2 * (3-1)} = 0{,}1827$$

donde m = mínimo número de filas o columnas en la tabla de contingencia
N = tamaño de la tabla

D) D de Somer:

$$d_{yx} = \frac{N_s - N_d}{N_s + N_d + T_y} = \frac{1.828 - 1.064}{1.828 + 1.064 + 1.264} = 0{,}1838$$

$$d_{xy} = \frac{N_s - N_d}{N_s + N_d + T_x} = \frac{1.828 - 1.064}{1.828 + 1.064 + 1.272} = 0{,}1834$$

Nota: el cálculo de pares se ha presentado en el anexo 1.

Cuadro 9.6. Medidas para conocer la relación entre variables ordinales (ejemplo con la tabla 9.15).

bién los que leen más libros no relacionados, mientras que un mayor número de pares discordantes provocan ausencia de relación, es decir que una persona que lee muchos libros (de todo tipo) leerá pocos libros relacionados con sus estudios. De este modo *Gamma* es el exceso de pares concordantes en relación al número de pares concordantes y discordantes, o dicho de otra forma es el número de predicciones correctas (número de entrevistados con el mismo orden en las dos variables) menos las incorrectas, dividido entre el total de predicciones. Tras constatar que se trata de un valor significativo, se procede con su interpretación. Un valor *Gamma* de 0,264 indica un 26% más de pares concordantes que discordantes, una mejora en la predic-

Medidas simétricas

		Valor	Error típ. sint.(a)	T aproximada (b)	Sig. aproximada
Ordinal por ordinal	Tau-b de Kendall	,184	,087	2,118	,034
	Tau-c de Kendall	,183	,086	2,118	,034
	Gamma	,264	,122	2,118	,034
N de casos válidos		112			

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

Medidas direccionales

			Valor	Error típ. sint.(a)	T aproximada (b)	Sig. aproximada
Ordinal por ordinal	d de Somer	Simétrica	,184	,087	2,118	,034
		(v06) Libros relacionados con tus estudios leídos en los últimos cinco meses dependiente	,184	,087	2,118	,034
		(v08) Libros leídos en los últimos cinco meses no relacionados con los estudios dependiente	,183	,087	2,118	,034

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

Tabla 9.14. *Estadísticos* para variables ordinales: D de Somer, Gamma, Tabu-b de Kendall y tau-c de kendall.

ción, una reducción del 26% del error al predecir los casos de una variable conociendo la ordenación de los casos en otra variable.

La mayoría de los expertos señalan que pese a que *Gamma* es medida muy versátil que es utilizada muy frecuentemente tiene el problema que suele sobreestimar la relación existente entre las variables analizadas, fundamentalmente si el número de pares empatados en una variable es muy elevado (Dometrius 1992: 309). Debido a estas

críticas algunos investigadores utilizan la *Tau-b* de Kendall que oscila entre –1 a 1 y su magnitud indica el grado de asociación entre dos variables: en que medida el cambio en una variable provocará cambios en la otra.

Esta medida tiene el inconveniente que únicamente puede utilizarse con tablas cuadradas (2 x 2, 3 x 3, 4 x 4, etc.) puesto que si las tablas no son cuadradas no puede llegar a uno, ya que cuando hay un número diferente de filas que de columnas existen más pares empatados en una variable que en la otra. Es por ello por lo que en tablas rectangulares se sustituye por *Tau-c*, que oscila entre –1 a 1 indicando su magnitud el grado de asociación entre las variables; esto es, cómo la variación de una variable produce variaciones en la segunda.

El siguiente estadístico, la *D de Somer* es una medida asimétrica que se interpreta de modo similar a *Gamma*, si bien presenta la ventaja que no sobreestima la relación entre variables al eliminar la influencia de los pares empatados en la variable dependiente. García Ferrando (1985: 253) explica la relación entre *D* y las *Taus* y afirma que la *tau-b* es un promedio de los dos coeficientes *D* de Somer que pueden calcularse en una misma tabla. Así la *D* es la reducción proporcional en el error cometido al predecir el ordenamiento de los casos en una variable mediante el conocimiento de la ordenación de los casos en otra variable, es decir que tiene una interpretación similar a la Gamma.

¿Que estadístico elegir?

Tras esta exposición, y considerando que todas las medidas explicadas son adecuadas para conocer la relación entre variables ordinales nos preguntamos qué estadístico es el mejor para conocer la relación entre dos variables. Ruiz Maya (1990: 287) considera que *Gamma* de Goodman y Kruskall es la más utilizada debido fundamentalmente a la facilidad de su interpretación, aunque cuando el número de empates es muy alto aconseja utilizar otras medidas. Dometrius (1992: 313) llega a una conclusión similar cuando expone que *Gamma* es más simple de calcular y de interpretar, aunque también afirma que la elección de la medida es un criterio personal y aconseja que cada investigador utilice la medida que mejor comprenda, opinión compartida también por Manzano (1995: 255). Dometrius, por otra parte, aconseja utilizar la *Tau* de Kendall porque es una medida más conservadora al eliminar el efecto de los empates, aunque señala que cuando el número de empates es pequeño el valor de la *Tau* será similar a *Gamm*a (992: 314).

Desde nuestro punto de vista creemos que la mejor medida es la *D de Somer* porque recoge la facilidad de interpretación de *Gamm*a, al tiempo que elimina el principal defecto de ésta (la sobreestimación de la relación) al excluir el efecto de los pares empatados. Además de ser la única medida asimétrica. No obstante nuestro consejo es que cada uno utilice la medida que mejor comprenda.

Tabla de contingencia (v08) Libros leídos en los últimos cinco meses no relacionados con los estudios * (v06) Libros relacionados con tus estudios leídos en los últimos cinco meses

			(v08) Libros leídos NO relacionados con los estudios (libros generalistas)			
			Ninguno	**Uno**	**Dos y tres**	**Total**
(v06) Libros relacionados con...	Ninguno	Recuento	22	2	12	36
		% de (v08) Libros NO relacionados con...	57,9%	6,3%	28,6%	32,1%
		Residuos corregidos	4,2	-3,7	-,6	
	Uno	Recuento	10	14	18	42
		% de (v08) Libros NO relacionados con...	26,3%	43,8%	42,9%	37,5%
		Residuos corregidos	-1,8	,9	,9	
	Dos y tres	Recuento	6	16	12	34
		% de (v08) Libros NO relacionados con...	15,8%	50,0%	28,6%	30,4%
		Residuos corregidos	-2,4	2,9	-,3	
Total		Recuento	38	32	42	112
		% de (v08) Libros NO relacionados con...	100,0%	100,0%	100,0%	100,0%

Tabla 9.15. Tabla de contingencia v08 y v06.

Otra cuestión a plantear es como interpretar la magnitud de cada estadístico. Dometrius (1992: 314) considera que magnitudes superiores al 0,3 ya indican niveles de asociación importantes. No obstante, desde su punto de vista una relación puede ser considerada fuerte o débil no tanto por ella misma sino en relación con el marco teórico previo y otras investigaciones similares. Este autor pone el ejemplo de un grado de asociación entre el *Nivel de Estudios* y el *Nivel de Ingresos* de 0,25. Éste será un importante resultado para analizar que elementos están detrás de este bajo nivel de asociación en unas variables que en nuestro (supuesto) marco teórico aparecían muy relacionadas.

En cualquier caso, se opte por una u otra medida, al observar el ejemplo utilizado se detecta que existe una relación *directa* entre el número de libros leídos en los últimos cinco meses y el número de libros leídos relacionados con tus estudios. Una relación directa implica que a medida que aumenta una variable se incrementa también los valores de la otra; es decir, que las personas que más leen todo tipo de libros son también las que leen más libros relacionados con sus estudios.

Resumen del procesamiento de los casos

	Casos					
	Válidos		Pérdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
(v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género	180	94,2%	11	5,8%	191	100,0%

Esta relación se observa con precisión en el interior de la tabla 9.15, bien utilizando los porcentajes de columna o los residuos corregidos. En este caso se han utilizado los porcentajes de columna, que indican que el 57,9% de los entrevistados que no leen libros generalistas tampoco leen otro tipo de libros. Sin embargo también es verdad que la mitad de los que han leído un libro generalista han leído dos o tres libros relacionados con los estudios, y un 44% un libro. De los que han leído *dos y tres* libros no relacionados con sus estudios (libros generalistas), el 43% ha leído un libro relacionado, y el 29% dos o tres libros.

Finalizaremos esta sección indicando que cuando se considera la relación entre una variable nominal y otra ordinal, como sucede en los ejercicios planteados dos párrafos más arriba, deben utilizarse los estadísticos y medidas de asociación propias de las variables nominales. Debe tenerse en cuenta que el programa calcula todo, absolutamente todo, y que el investigador es quién debe elegir qué interpretar en función de su hipótesis de trabajo y del tipo de variables que utiliza. En la tabla 9.16 se han solicitado todas las medidas de asociación vistas a lo largo de la exposición, medidas para variables nominales, para variables ordinales, e incluso para variables de intervalo[130]. El programa lo calcula absolutamente todo, y es el investigador el que debe discernir lo que debe interpretar.

Con el fin de fijar los conocimientos aprendidos en esta sección, antes de considerar nuevos contenidos, proponemos unos ejercicios utilizando el archivo de datos sobre *Vida Cotidiana*. ¿Hasta que punto el número de amigos de verdad (pregunta 34b, variable b68) está relacionado con la edad (e10) de los entrevistados (edad en cuatro

130. Obsérvese que en la tabla 9.16 aparecen dos medidas no tratadas, la *Correlación de Pearson* (R de Pearson) y *Correlación de Spearman*, no explicadas en este capítulo al alejarse de nuestros propósitos.

Tabla de contingencia (v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género

			(v49) Genero		
			Hombre	**Mujer**	**Total**
(v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre	Beber, ir de copas	Recuento	26	20	46
		% de (v49) Género	37,1%	18,2%	25,6%
		Residuos corregidos	2,8	-2,8	
	Bailar	Recuento	0	12	12
		% de (v49) Género	,0%	10,9%	6,7%
		Residuos corregidos	-2,9	2,9	
	Hacer deporte	Recuento	12	8	20
		% de (v49) Género	17,1%	7,3%	11,1%
		Residuos corregidos	2,1	-2,1	
	Viajar	Recuento	12	28	40
		% de (v49) Género	17,1%	25,5%	22,2%
		Residuos corregidos	-1,3	1,3	
	Ir al cine	Recuento	0	10	10
		% de (v49) Género	,0%	9,1%	5,6%
		Residuos corregidos	-2,6	2,6	
	Practicar alguna afición o hobby	Recuento	12	8	20
		% de (v49) Género	17,1%	7,3%	11,1%
		Residuos corregidos	2,1	-2,1	
	Otras	Recuento	8	24	32
		% de (v49) Género	11,4%	21,8%	17,8%
		Residuos corregidos	-1,8	1,8	
Total	Recuento	70	110	180	
		% de (v49) Género	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	**gl**	**Sig. asintótica (bilateral)**
Chi-cuadrado de Pearson	31,447(a)	6	,000
Razón de verosimilitud	38,885	6	,000
Asociación lineal por lineal	1,249	1	,264
N de casos válidos	180		

a 2 casillas (14,3%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 3,89.

Medidas direccionales

			Valor	Error típ. sint.(a)	T aproximada (b)	Sig. aproximada
Ordinal por ordinal	d de Somer	Simétrica	,141	,063	2,236	,025
		(v01bis) Actividad, fuera de casa, que más te gusta hacer... dependiente	,193	,086	2,236	,025
		(v49) Género dependiente	,112	,050	2,236	,025

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

Medidas simétricas

		Valor	Error típ. sint.(a)	T aproximada (b)	Sig. aproximada
Nominal por nominal	Phi	,418			,000
	V de Cramer	,418			,000
	Coeficiente de contingencia	,386			,000
Ordinal por ordinal	Tau-b de Kendall	,147	,066	2,236	,025
	Tau-c de Kendall	,183	,082	2,236	,025
	Gamma	,230	,102	2,236	,025
	Correlación de Spearman	,166	,074	2,243	,026(c)
Intervalo por intervalo	R de Pearson	,084	,075	1,118	,265(c)
N de casos válidos		180			

a Asumiendo la hipótesis alternativa.
b Empleando el error típico asintótico basado en la hipótesis nula.
c Basada en la aproximación normal.

Tabla 9.16. Cruce de tabla con todos los estadísticos disponibles.

grupos: de 18 a 29 años, de 30 a 44, de 45 a 64, y 65 y más años[131]). ¿Y el número de amigos con el hábitat o tamaño del municipio donde se reside (e36)?

Al final de la sección 3 se interpretó una relación entre el nivel de equipamiento y la clase social. ¿Qué tipos de variables han sido utilizadas en esa relación? ¿Está bien realizado el ejercicio? Proponer la solución correcta.

6. Anexo 1: Introducción al cálculo de pares

Las medidas utilizadas para analizar la relación entre variables ordinales se fundamentan en el cálculo de pares que se expone en el presente apartado. Nosotros hemos tomado esta explicación de la obra de García Ferrando (1985), entre las páginas 245-250. El número total de pares, como se expone más abajo, es el tamaño de la muestra por el tamaño de la muestra menos uno, dividido entre dos; de modo que la tabla utilizada en el apartado 9.5 cuenta con 6.216 pares. El total de pares se divide en cinco tipos de pares: semejantes o concordantes (aquellos que se distribuyen en idéntico orden en ambas variables); desemejantes o discordantes (ordenados en orden opuesto); empatados en la variable independiente X (y no en la variable dependiente Y); empatados solo en la variable dependiente Y (y no en la variable independiente X); y pares empatados en ambas variables.

Tabla de contingencia (v08) Libros leídos en los últimos cinco meses no relacionados con los estudios * (v06) Libros relacionados con tus estudios leídos en los últimos cinco meses

		(v06) Libros relacionados con tus estudios leídos en los últimos cinco meses			
		Ninguno	**Uno**	**Dos y tres**	**Total**
(v08) Libros leídos NO relacionados...	Ninguno	**22**	10	6	38
	Uno	2	**14**	16	32
	Dos y tres	12	18	**12**	42
Total	36	42	34	112	

131. Respecto al criterio de agrupación de la edad, se ha realizado siguiendo la distribución realizada por Amando de Miguel (1997: 59) en sus estudios sobre la sociedad española. Obsérvese que se trata de la misma división por grupos de edad que se propuso al final de la sección 4 del capítulo VIII.

El cálculo de pares, siguiendo a García Ferrando (1985: 246), debe comenzar con la elección de la diagonal que une las celdillas que contienen los valores *alto-alto* y *bajo-bajo* en ambas variables, denominada por García Ferrando como *diagonal positiva*. En la tabla sobre este párrafo se aprecia ésta es la diagonal que une el extremo superior izquierdo con el extremo inferior derecho, la que presenta los valores 22–14–12, valores que han sido colocados en negrilla. Definida la diagonal positiva, procedemos con el cálculo de pares, tomado de la obra de García Ferrando (1985: 245-250).

a) *Número total de pares:*

$$\text{Pares} = \frac{N\,(N-1)}{2} = \frac{112 * 111}{2} = 6.216$$

b) *Pares semejantes o concordantes:* $\mathbf{N_s}$

22 * (14 + 16 + 18 + 12) = 1.320
10 * (16 + 12) = 280
2 * (18 + 12) = 60
14 * 12= 168
Total: 1.828

c) *Pares desemejantes o discordantes:* $\mathbf{N_d}$

12 * (10 + 6 + 14 + 16) = 552
18 * (16 + 6) = 396
2 * (10 + 6) = 32
14 * 6= 84
Total: 1.064

d) *Pares empatados solo en la variable independiente X:* $\mathbf{T_x}$

22 * (2 + 12) = 308
2 * 12 = 24
10 * (14 + 18) = 320
14 * 18= 252
6 * (16 + 12) = 168
16 * 12 = 192
Total: 1.264

e) *Pares empatados solo en la variable dependiente Y:* $\mathbf{T_y}$

22 * (10 + 6) = 352
10 * 6 = 60
2 * (14 + 16) = 60

14 * 16= 224
12 * (18 + 12) = 360
16 * 12 = 216
Total: 1.272

f) *Pares empatados simultáneamente en X e Y:* $\mathbf{T_{xy}}$
Se trata de aplicar la fórmula "f (f – 1) / 2" a cada celdilla, donde f es la frecuencia de cada celdilla

22 (22 – 1) / 2 = 231
6 (6 – 1) / 2 = 15
14 (14 – 1) / 2 = 91
12 (12 – 1) / 2 = 66
12 (12 – 1) / 2 = 66
10 (10 – 1) / 2 = 45
2 (2 – 1) / 2 = 1
16 (16 – 1) / 2 = 120
18 (18 – 1) / 2 = 153
Total: 788

7. Anexo 2: Lenguaje de sintaxis de los análisis realizados

En el apartado 7 del capítulo VIII se explicó el origen de cada uno de estos mandatos (pulsando el botón Pegar en el cuadro de diálogo correspondiente), así como el proceso de *ejecución* de cada uno.

Apartado 2: Elaboración de tabla de contingencia con dos variables

```
FREQUENCIES
    VARIABLES=v01
    /ORDER= ANALYSIS.

RECODE v01 (4=14) (7=14) (8=14) (9=14) (10=14) (15 thru 18=14) (98=99) into
v01bis.

FREQUENCIES
    VARIABLES=v01bis
    /ORDER= ANALYSIS.

CROSSTABS
    /TABLES=v01bis BY v49
```

```
/FORMAT= AVALUE TABLES
/CELLS= COUNT
/COUNT ROUND CELL.
```

Apartado 3.1: Relación entre variables nominales utilizando el Chi-cuadrado

```
CROSSTABS
  /TABLES=v01bis BY v49
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ
  /CELLS= COUNT
  /COUNT ROUND CELL.

CROSSTABS
  /TABLES=v01bis BY v49
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ
  /CELLS= COUNT EXPECTED
  /COUNT ROUND CELL.

CROSSTABS
  /TABLES=v01bis BY TITULAC
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ
  /CELLS= COUNT EXPECTED
  /COUNT ROUND CELL.
```

Apartado 3.2: Consideraciones a tener en cuenta en la utilización del Chi-cuadrado

```
CROSSTABS
  /TABLES=v01 BY v49
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ CC PHI
  /CELLS= COUNT
```

```
    /COUNT ROUND CELL.

CROSSTABS
    /TABLES=v028 BY v49
    /FORMAT= AVALUE TABLES
    /STATISTIC=CHISQ CC PHI
    /CELLS= COUNT
    /COUNT ROUND CELL.
```

Apartado 3.3: Estadísticos basados en el Chi-cuadrado

```
CROSSTABS
    /TABLES=v01bis BY v49
    /FORMAT= AVALUE TABLES
    /STATISTIC=CHISQ CC PHI
    /CELLS= COUNT
    /COUNT ROUND CELL.
```

Apartado 4.1: Análisis del interior de la tabla: cálculo y diferencia de porcentajes.

```
CROSSTABS
    /TABLES=v01bis BY v49
    /FORMAT= AVALUE TABLES
    /CELLS= COUNT ROW COLUMN TOTAL
    /COUNT ROUND CELL.
```

Apartado 4.2: Análisis del interior de la tabla: utilización de residuos

```
CROSSTABS
    /TABLES=v01bis BY v49
    /FORMAT= AVALUE TABLES
    /CELLS= COUNT RESID SRESID ASRESID
    /COUNT ROUND CELL.
```

```
CROSSTABS
    /TABLES=v01bis BY v49
    /FORMAT= AVALUE TABLES
    /CELLS= COUNT ASRESID
    /COUNT ROUND CELL.
```

Apartado 5: Utilización de test estadísticos para conocer la relación entre variables ordinales.

```
FREQUENCIES
    VARIABLES=v06 v08
    /ORDER= ANALYSIS.

MISSING VALUE V08 (99).

FREQUENCIES
    VARIABLES=v08
    /ORDER= ANALYSIS.

USE ALL.
COMPUTE filter_$=(v06 <= 3 & v08 < 4).
VARIABLE LABEL filter_$ 'v06 <= 3 & v08 < 4 (FILTER)'.
VALUE LABELS filter_$ 0 'No seleccionado' 1 'Seleccionado'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.

FREQUENCIES
    VARIABLES=v06 v08
    /ORDER= ANALYSIS.

CROSSTABS
    /TABLES=v08 BY v06
    /FORMAT= AVALUE TABLES
    /STATISTIC=GAMMA D BTAU CTAU
    /CELLS= COUNT SRE
    /COUNT ROUND CELL.
```

```
CROSSTABS
    /TABLES=v01bis BY v49
    /FORMAT= AVALUE TABLES
    /STATISTIC=CHISQ CC PHI CORR GAMMA D BTAU CTAU
    /CELLS= COUNT COLUMN ASRESID
    /COUNT ROUND CELL.
```

Apartado 11: introducción al cálculo de pares

```
CROSSTABS
    /TABLES=v08 BY v06
    /FORMAT= AVALUE TABLES
    /STATISTIC=GAMMA D BTAU CTAU
    /CELLS= COUNT SRE
    /COUNT ROUND CELL.
```

Apartado 12: presentación de un ejemplo.

```
FREQUENCIES
    VARIABLES=v44
    /ORDER= ANALYSIS.
RECODE v44 (4=10) (1 thru 3=20) (5 thru 23=20) INTO TITULAC.
VARIABLE LABELS TITULAC 'Titulación (Sociología/no sociología)'.
VALUE LABELS TITULAC 10"Sociología" 20"No Sociología".
FREQUENCIES
    VARIABLES=titulac
    /ORDER= ANALYSIS.

FREQUENCIES
    VARIABLES=v18
    /ORDER= ANALYSIS.
RECODE v18 (90 thru 99=SYSMIS).
FREQUENCIES
    VARIABLES=v18
    /ORDER= ANALYSIS.
RECODE v18 (3=2).
VALUE LABELS v18 1"Sábado" 2"Domingo y otro día festivo" 4"Otro día no festivo".
```

```
FREQUENCIES
    VARIABLES=v18
    /ORDER= ANALYSIS.

CROSSTABS
    /TABLES=v18 BY TITULAC
    /FORMAT= AVALUE TABLES
    /STATISTIC=CHISQ PHI
    /CELLS= COUNT COLUMN
    /COUNT ROUND CELL.

CROSSTABS
    /TABLES=v18 BY TITULAC
    /FORMAT= AVALUE TABLES
    /STATISTIC=CHISQ PHI
    /CELLS= COUNT ASRESID
    /COUNT ROUND CELL.
```