Cátedra: Salud

Materia: Método I

Carrera:

Autor: Análisis Bivariado
García Fernando Cap. 7

53 por 100, con una desviación típica del 3 por 100. ¿Hasta qué punto se puede afirmar que el grado de apoyo de los electores al candidato en la campaña actual es diferente del manifestado en el pasado, al nivel de significación de a) 0,05 y b) 0,005?

6. De una muestra de 500 votantes elegidos aleatoriamente en una comunidad, el 55 por 100 de ellos son partidarios de un determinado candidato. Hallar los límites de confianza del a) 95 por 100 y b) 99 por 100 para la proporción de todos los votantes que son partidarios de dicho candidato.

7. ¿Qué tamaño de muestra debería tomarse en el caso del ejercicio anterior para que la confianza de que el candidato salga elegido (es decir, obtenga el 50 por 100 o más de votos) fuera del 95 por 100? ¿y del 99 por 100?

8. En un estudio sobre la ideología de los trabajadores de una gran empresa, se pidió a 100 trabajadores elegidos al azar que se posicionasen en una escala de preferencia política que tiene un recorrido del 1 al 10 (1=extrema izquierda; 10=extrema derecha). El valor medio de los posicionamientos de los trabajadores fue $\bar{X}=4{,}2$, con una desviación típica de 0,04. Estimar el valor medio del posicionamiento ideológico de todos los trabajadores, con un intervalo de confianza de a) 95 por 100 y b) 99 por 100.

BIBLIOGRAFIA

BLALOCK, Hubert M., *Social Statistics*, New York, McGraw-Hill, 1960.

COCHRAN, W. G.: «Some methods for strengthening the common χ^2 test», *Biometrics*, 10, 1954, pág. 417-451.

DOMÉNECH I MASSONS, J. M.: *Bioestadística. Métodos Estadísticos para Investigadores*, Barcelona, Herder, 1977.

LABOVITZ, Sanford: «The Nonutility of Significance Tests: The Significance of Test of Significance Reconsidered», *Pacific Sociological Research*, vol. 13, núm. 3, 1970, págs. 141-147.

LOETHER, N. J., y D. G. MCTAVISH: *Inferential Statistics for Sociologists*, Boston, Ally & Bacon, 1974.

MORRISON, D. E., y R. E. HENKEL: *The Significance Test Controversy*, Chicago, Aldine, 1970.

SÁNCHEZ-CRESPO, J. L.: *Principios elementales del muestreo y estimación de proporciones*, Madrid, INE, 1971.

SIEGEL, Sidney: *Nonparametric Statistics for the Behavioral sciences*, New York, McGraw-Hill 1956.

Capítulo 7

ESTADISTICA DESCRIPTIVA BIVARIABLE: CARACTERISTICAS DE UNA ASOCIACION BIVARIABLE

Nuestro objetivo en el presente capítulo es el estudio de las características de las distribuciones bivariables o clasificaciones cruzadas de dos variables. Con ello adelantamos un nuevo paso en nuestro recorrido por el camino que nos va mostrando, en su creciente complejidad, la diversidad de las técnicas estadísticas utilizadas en la investigación sociológica. En la estadística descriptiva univariable comenzamos con una distribución de frecuencias y, a partir de ella, estudiamos una serie de medidas resumen que nos permitieron lograr números índices, de gran utilidad para la descripción de los datos sociológicos. Al mismo tiempo, se desarrollaron una serie de medidas para cada uno de los rasgos definitorios de una distribución; esto es, la tendencia central, la forma y la variabilidad o dispersión. En el presente capítulo, nuestro objetivo es similar, pero, si cabe, más interesante, pues nos vamos acercando más al tipo de tarea que con mayor frecuencia se realiza en la realidad de la investigación sociológica, esto es, el estudio de las condiciones que influyen en la distribución de una variable.

7.1. DISTRIBUCIONES BIVARIABLES: UN EJEMPLO

Los sociólogos que estudian las actitudes políticas de la población vienen utilizando, desde hace años, en las encuestas de opinión una escala de preferencia ideológica izquierda-derecha. En la entrevista se pide al entrevistado que se sitúe en una de las casillas que componen una escala que, según el tipo de estudio, va del 1 al 7 o del 1 al 10, correspondiendo el extremo 1 a la preferencia por la extrema izquierda y el extremo 7, ó 10, a la preferencia por la extrema derecha. Como señalan los autores del estudio *La Conciencia Regional en España* (J. Jiménez Blanco *et al.*, 1977, pág. 88), «se trata de un planteamiento de extrema simplicidad, en el que se traspone la dimensión ideológica del espectro político a una dimensión geométrica en el plano». Pues bien, los porcentajes de autoubicación obtenidos de la población española en el referido estudio fueron los siguientes.

TABLA 7.1

Porcentajes de autoubicación en un espacio político abstracto

Escala		f	%
Izquierda.	1	127	2
	2	190	3
	3	381	6
	4	444	7
	5	1.522	24
	6	888	14
	7	381	6
	8	444	7
	9	254	4
Derecha.	10	317	5
	NS/NC	1.334	21
		6.342	100

FUENTE: J. JIMÉNEZ BLANCO, et. al.: *La Conciencia Regional en España*, Madrid, CIS, 1977, pág. 88.

La interpretación de estos datos puede ser muy diversa, según el número de casillas que se asignen a cada postura ideológica. Haciendo una «lectura desde el centro», como ensayan los autores, cabe asignar las primeras cuatro casillas a la izquierda y las últimas cuatro a la derecha, quedando para el centro las casillas cinco y seis. Ahora bien, los autores del estudio no desean tan sólo conocer la distribución global de la población española a lo largo de dicha escala, sino que, además, indagan su distribución en relación a otras variables relevantes. Así, a partir de la consideración de la influencia que tiene en las actitudes políticas el grado de religiosidad (o, en general, la religión), estudian la distribución de las preferencias ideológicas izquierda-derecha entre la población clasificada según su nivel de religiosidad.

En la tabla 7.2 aparece el conjunto de tales distribuciones. Se trata de cinco distribuciones univariantes de la escala izquierda-derecha, entre muy buenos católicos, católicos practicantes, católicos no muy practicantes, católicos no practicantes e indiferentes en materia religiosa.

En un primer análisis de estos datos se observan las diferencias que se producen en la distribución de la población entre las diversas posiciones ideológicas, para los distintos grupos considerados. Una parte importante de los que se declaran indiferentes en materia religiosa, o que se consideran a sí mismos católicos no practicantes, se autoubican en posturas de izquierda, mientras que todo lo contrario ocurre entre los muy buenos católicos, que tienden a situarse en el centro o, sobre todo, en la derecha. Los católicos practicantes o no muy practicantes se sitúan preferentemente en las casillas 5 y 6, correspondientes al centro.

TABLA 7.2

Escala izquierda-derecha entre la población clasificada según nivel de religiosidad

Escala izqda.-dcha. entre muy buenos católicos			*Escala izqda.-dcha. entre católicos practicantes*			*Escala izqda.-dcha. entre católicos no muy practicantes*		
Escala	*f*	*%*	*Escala*	*f*	*%*	*Escala*	*f*	*%*
Izqda. (1-4) . .	45	5	Izqda. (1-4) . .	187	8	Izqda. (1-4) . .	235	16
Centro (5-6) .	271	30	Centro (5-6) .	1.004	43	Centro (5-6) .	513	35
Dcha. (7-10) .	352	39	Dcha. (7-10) .	607	26	Dcha. (7-10) .	381	26
NS/NC	235	26	NS/NC	538	23	NS/NC	337	23
	903	100		2.336	100		1.466	100

Escala izqda.-dcha entre católicos no practicantes			*Escala izqda.-dcha. entre indiferentes*		
Escala	*f*	*%*	*Escala*	*f*	*%*
Izqda. (1-4) . .	292	34	Izqda. (1-4) .	352	52
Centro (5-6) .	317	37	Centro (5-6) .	190	28
Dcha (7-10) .	945	11	Dcha. (7-10) .	41	6
NS/NC	155	18	NS/NC	94	14
	859	100		677	100

FUENTE: J. JIMÉNEZ BLANCO, et. al.: *op. cit.* Elaboración propia.

Nótese que estamos llegando a estas conclusiones a través de la comparación de porcentajes, en lugar de frecuencias absolutas, ya que el número de miembros que corresponde a cada uno de los cinco grupos considerados es diferente entre sí. Por eso, el uso de porcentajes está aconsejado para realizar comparaciones válidas, tal como se estudió en el capítulo 2.

Lo que hemos hecho hasta ahora ha sido comparar la distribución de los valores de la escala izquierda-derecha entre diversas distribuciones univariables. La variable dependiente es común a cada una de las distribuciones, y cada tabla se diferencia porque agrupa a la población que manifiesta determinado nivel de religiosidad. Pero una forma más eficaz y rápida de obtener conclusiones válidas bajo estas condiciones sería la de combinar las cinco tablas separadas en una sola tabla, tal como se hace a continuación.

TABLA 7.3

Distribución porcentual de la escala izquierda-derecha según el nivel de religiosidad de la población española

Escala izquierda-derecha	*Total*	*Muy buenos católicos*	*Católicos practicantes*	*Católicos no muy practicantes*	*Católicos no practicantes*	*Indiferentes*
Izqda. (1-4)	18	5	8	16	34	52
Centro (5-6)	38	30	43	35	37	28
Dcha. (7-10)	22	39	26	18	11	6
NS/NC	21	26	23	21	18	14
Total	100	100	100	100	100	100
N =	(6.342)	(903)	(2.366)	(1.466)	(859)	(677)

FUENTE: Ver tablas 7.1 y 7.2

Esta tabla nos permite comparar los diversos grupos entre sí, y cada uno de ellos con la media nacional, de una forma más rápida y eficaz, a la vez que ofrece un excelente resumen de la información que se contiene en las dos tablas anteriores, en forma del tipo de relación que se produce entre dos variables. Es el tipo de tabla que se conoce con el nombre de *distribución porcentual bivariable*, porque permite examinar la distribución porcentual de una variable (la variable dependiente) dentro de las diferentes categorías de otra variable (la variable independiente). Precisamente las ideas que subyacen debajo de tales clasificaciones cruzadas constituyen las bases del análisis empírico en la sociología, ya que es a través de dicho análisis como se trata de formular y contrastar el tipo de relación existente entre las variables, así como las condiciones en que se produce.

Una distribución bivariable, tal como la que se presenta en la tabla 7.3, permite no sólo el examen directo de la distribución global de una variable dependiente, sino también las condiciones que se supone influyen en la manera en que se distribuye dicha variable. Por lo que se refiere al caso de la ideología política, la teoría sugiere que, bajo ciertas condiciones, las posturas ideológicas de izquierda, centro y derecha se distribuirán de forma diferente que bajo otras condiciones. En el caso concreto de los datos que se incluyen en la tabla 7.3, tales condiciones corresponden a los diferentes niveles de religiosidad, aunque, como resulta obvio, el sociólogo puede pensar en otras condiciones que también pueden influir en la preferencia ideológica, tales como la edad, el sexo, la clase social, el lugar de residencia, etc. Como se ha dicho otras veces, se puede imaginar que una de las tareas de toda disciplina científica es la de buscar los tipos de condiciones que contribuyen a

mejor predecir y explicar el nivel de algún tipo de fenómeno (Loether y McTavish, 1974, pág. 174). Una tabla bivariable como la anterior, en realidad, pone en relación una serie de distribuciones condicionales con una distribución global de una variable dependiente.

En lo que resta de capítulo vamos a estudiar las principales características de las distribuciones bivariables o clasificaciones cruzadas de dos variables, y comenzaremos dicho estudio estableciendo algunas reglas prácticas sobre la forma de presentar y leer correctamente, desde un punto de vista tanto teórico como metodológico, una tabla bivariable.

7.2. Presentación y análisis de una tabla bivariable

La tabulación cruzada y comparativa de dos variables da lugar a una tabla compuesta de filas y columnas, utilizándose las categorías de cada variable para designar, respectivamente, las filas y las columnas. Se suele seguir la convención de situar la variable dependiente, cuando la hubiere, en las filas, y la variable independiente en las columnas.

Con el fin de ilustrar la forma en que se construye una tabla, supongamos que disponemos de 15 puntuaciones correspondientes a los

TABLA 7.4

Ejemplo de una tabla de frecuencias bivariable

Puntuaciones	*Sexo*	*Escala izquierda-derecha*	*Puntuaciones (cont.)*	*Sexo*	*Escala izquierda-derecha*
1	v	I	8	m	C
2	m	D	9	m	D
3	v	C	10	v	D
4	v	C	11	v	I
5	m	D	12	m	I
6	v	I	13	v	C
7	m	D	14	m	C
			15	m	I

Tabla de preferencia ideológica por sexo

v — varón
m — mujer
I — izquierda
C — centro
D — derecha

Escala izqda.-dcha	*Sexo* v	*Sexo* m	*Total filas*
I	3	2	5
C	3	2	5
D	1	4	5
Total columnas	7	8	15

valores obtenidos en la escala de preferencia ideológica izquierda-derecha para 15 personas entrevistadas en una encuesta. La distribución bivariable se establecerá a partir de la consideración del sexo de los entrevistados y de los valores que cada uno de ellos ha obtenido en la escala izquierda-derecha (que, por simplificación, categorizaremos como izquierda *I*, centro *C* y derecha *D*). Para este caso se considera como variable independiente el sexo, por lo que las dos categorías que componen dicha variable, varón y mujer, se situarán encabezando las columnas. La variable dependiente, preferencia ideológica, vendrá compuesta por tres categorías, izquierda, centro y derecha, situándose su encabezamiento en las filas. En la tabla 7.4 aparecen los datos correspondientes al ejemplo utilizado.

Establecida una tabla 2×3 (dos columnas×tres filas), se forman seis celdillas en las que se escriben las correspondientes frecuencias que aparecen en la distribución global de puntuaciones y datos. De esta forma no sólo se distribuyen las frecuencias en cada celdilla, sino que, además, se calculan dos subtotales (el subtotal de las filas y el subtotal de las columnas) y un total global.

Para una tabla 2×2, se pueden simbolizar las operaciones que hemos realizado en el ejemplo anterior del siguiente modo:

	Columna 1	*Columna 2*	*Total filas*
Fila 1	n_{11}	n_{12}	$\sum_j n_{1j}$
Fila 2	n_{21}	n_{22}	$\sum_j n_{2j}$
Total columnas	$\sum_i n_{i1}$	$\sum_i n_{i2}$	N

El valor de la frecuencia correspondiente a cada celdilla se simboliza mediante n_{ij}, en donde el subfijo i indica el número de la fila y el subfijo j indica el número de la columna. Esto es, que la frecuencia de cada celdilla indica el número de casos que comparten dos características simultáneamente, y el conjunto de las frecuencias de todas las celdillas constituyen las distribuciones condicionales. Los subtotales correspondientes a las filas y columnas reflejan los marginales o distribuciones univariables.

7.2.a. *Cálculo de los porcentajes en una tabla*

Habitualmente, las tablas bivariables se presentan con números que reflejan porcentajes en lugar de frecuencias absolutas. Con ello se facilita la realización de comparaciones numéricas entre las distribuciones

condicionales que se desean comparar. En tal caso surge la pregunta: ¿qué comparación debe facilitarse, la de las cifras de las distintas columnas o las que aparecen en las diferentes filas?

Existe una regla sencilla, universalmente aceptada, que se utiliza como guía para responder a la anterior pregunta. Dicha regla puede utilizarse cuando uno de los factores se considera como la causa de la distribución del otro factor. Entonces, *los porcentajes deben calcularse en el sentido del factor «causal» o de la variable independiente.*

Conviene tener en cuenta, al aplicar esta regla de «causa y efecto», que no se trata en realidad de resolver un problema de causalidad real, sino de qué factor consideramos que afecta a la distribución porcentual del segundo factor. Por esa razón, algunos autores prefieren hablar de variables independientes y dependientes, en cuyo caso se dice que los porcentajes deben calcularse en la dirección de la variable independiente. Todo esto quiere decir que los porcentajes deben sumar el 100 por 100 para cada categoría del factor causal o de la variable independiente, es decir, que los porcentajes de cada columna deben sumar el 100 por 100. Algunas veces, por motivos de presentación de la tabla, las variables independientes se escriben en las filas, en cuyo caso los porcentajes tendrán que calcularse en el sentido horizontal.

Tres son las formas en que se pueden calcular los porcentajes en una tabla, a partir de la distribución de las frecuencias absolutas. Los porcentajes se pueden calcular bien utilizando el total de las columnas como base de los porcentajes, bien tomando como base el total de las filas o bien tomando el total global como base porcentual. Considerando los datos hipotéticos que se han presentado en la tabla 7.4 se pueden calcular los tres tipos de porcentajes, tal como se observa en la siguiente tabla 7.5. Dado que la variable independiente en el ejemplo que venimos utilizando es el sexo, la forma correcta de examinar las diferencias que existen entre las dos categorías de la variable sexo, por lo que a sus preferencias políticas se refiere, es a través de las distribuciones porcentuales calculadas en base a los subtotales de las columnas, es decir, en base a los totales de varones y a los totales de mujeres (tabla 7.5.a).

Pero si lo que se desea es comparar la distribución de cada categoría de la variable dependiente entre las categorías de la variable independiente, entonces hay que calcular los porcentajes en la otra dirección, es decir, en sentido horizontal. De este modo, por ejemplo, podemos comprobar que mientras el 60 por 100 de los que se identifican como de izquierdas son varones y el 40 por 100 son mujeres, de los que se identifican como de derechas el 20 por 100 son varones, mientras que el 80 por 100 son mujeres (tabla 7.5.b).

Las comparaciones en una tabla bivariable porcentual, pues, se realizan mediante el examen de las diferencias que aparecen entre los porcentajes. Así, por ejemplo, en la tabla 7.5.a, la diferencia entre los

TABLA 7.5

Ilustración de las diferentes formas en que se pueden calcular los porcentajes en una tabla bivariable

Distribución de las frecuencias de la tabla 7.4.

Escala izqda.-dcha.	Sexo		
	Varón	*Mujer*	*Total*
I	3	2	5
C	3	2	5
D	1	4	5
Total	7	8	15

7.5.a. *Cálculo de porcentajes tomando como base los totales de las columnas.*

Escala izqda.-dcha.	Sexo		
	Varón %	*Mujer* %	*Total* %
I	42,9	25,0	33,3
C	42,9	25,0	33,3
D	14,2	50,0	33,3
Total	100	100	100

7.5.b. *Cálculo de porcentajes tomando como base los totales de las filas.*

Escala izqda.-dcha.	Sexo		
	Varón %	*Mujer* %	*Total* %
I	60,0	40,0	100
C	60,0	40,0	100
D	20,0	80,0	100
Total	46,7	53,3	100

7.5.c *Cálculo de porcentajes tomando como base la frecuencia total N.*

Escala izqda.-dcha.	Sexo		
	Varón %	*Mujer* %	*Total* %
I	20,0	13,3	33,3
C	20,0	13,3	33,3
D	6,7	26,7	33,3
Total	46,7	53,3	100

porcentajes de «izquierda» entre hombres y mujeres es 17,9 por 100 (42,9−25,0=17,9 por 100). Este valor se llama *epsilón*, y se simboliza mediante la letra griega ε (Loether y McTavish, *op. cit.*, pág. 178). En tablas con más de dos columnas se pueden calcular porcentajes de contraste o epsilones para cada par de columnas y entre las correspondientes categorías. Más adelante veremos con más detalle el uso que se puede hacer de tales valores.

Como destaca Zeisel en su clásico libro *Digalo con números* (1974, pág. 38), el contenido estadístico de las tres formas de calcular los porcentajes en una tabla es el mismo, pero al calcular los porcentajes en diferentes sentidos se acentúan distintas distribuciones y se ofrecen diferentes comparaciones. Así, del examen de la tabla 7.5.a se puede concluir que los varones se identifican preferentemente con posiciones ideológicas de izquierda y centro —el 42,9 por 100 se identifica con la izquierda y otro tanto lo hace con el centro, mientras que sólo un 14,2 por 100 lo hace con la derecha—, y las mujeres se identifican en mayor proporción con las posiciones de derecha y, en menor grado, con el centro e izquierda —en concreto, el 50 por 100 se identifica con la derecha, y sólo el 25 por 100 lo hace con la izquierda y centro, respectivamente.

Conclusiones de otro tipo se alcanzarán si examinamos la tabla 7.5.b. Así, si nos fijamos en la categoría ideológica del centro, se observa que hay más varones, el 60 por 100, que mujeres, el 40 por 100, entre los que se identifican con tales posiciones. De igual modo se pueden analizar las categorías de izquierda y derecha, comprobando para cada una de ellas la composición de varones y mujeres que con ellas se identifican.

Por último, cabe observar la distribución de los datos que se presentan en la tabla 7.5.c. Aquí, los números de cada celdilla representan los porcentajes de casos en relación al total *N*. La lectura de esta tabla nos permite, por ejemplo, concluir que la categoría más numerosa es la de mujeres de derechas, el 26,7 por 100, mientras que la categoría menos numerosa es la de varones de derechas, el 6,7 por 100; o también se puede observar que existe el mismo porcentaje de varones de izquierdas que de derechas, el 20 por 100, y lo mismo ocurre con las mujeres, entre las que un 13,3 por 100 se manifiesta de izquierdas y otro tanto como de centro.

No siempre es posible determinar qué variable es la independiente (o «causa») y qué variable es la dependiente (o «efecto»). En tal caso no es aplicable la regla causa y efecto o variable independiente a variable dependiente, ya que las tablas que se obtengan al calcular los porcentajes en un sentido u otro ofrecerán diferentes interpretaciones. Como ejemplo vamos a considerar los siguientes resultados, obtenidos en una encuesta sobre temas de actualidad realizada por el Centro de Investigaciones Sociológicas entre la población española en febrero de 1981. Una de las preguntas incluidas en el cuestionario hacía referencia a la per-

TABLA 7.6

Percepción de la situación económica general del país según las preferencias ideológicas de la población

7.6.a. *Frecuencias absolutas*

Escala ideológica	Situación económica			Total
	Buena	Regular	Mala	
Izquierda	12	99	198	309
Centro	8	93	115	216
Derecha	3	39	69	110
Total	22	231	382	635

7.6.b. *Porcentajes calculados en el sentido de las columnas*

Escala ideológica	Situación económica			Total
	Buena	Regular	Mala	
Izquierda	55	53	52	49
Centro	36	40	30	34
Derecha	9	17	18	17
Total	100	100	100	100

7.6.c *Porcentajes calculados en el sentido de las filas.*

Escala ideológica	Situación económica			Total
	Buena	Regular	Mala	
Izquierda	4	32	64	100
Centro	4	43	53	100
Derecha	2	36	62	100
Total	4	36	60	100

FUENTE: Resultados parciales de la Tabla 6 correspondiente al Barómetro de Opinión Pública, febrero 1981, *REIS*, núm. 15, 1981, pág. 189.

cepción de la situación económica general del país. Al cruzar las respuestas obtenidas por la preferencia ideológica se obtuvieron los resultados parciales que se presentan en la tabla 7.6.

En la tabla 7.6.a se presentan las frecuencias absolutas que se obtienen al cruzar la variable «percepción de la situación económica» por la variable «preferencia ideológica». A partir de tales resultados se han calculado los porcentajes en el sentido de las columnas (tabla 7.6.b) y en el sentido de las filas (tabla 7.6.c). En este ejemplo resulta difícil

destacar qué variable es la causa de la otra. Por eso, al calcular de una forma u otra los porcentajes, podemos responder a diferentes preguntas. Por medio de la tabla 7.6.b se puede responder a la siguiente pregunta: ¿cuál es la distribución de las preferencias ideológicas entre los que consideran como buena (o regular, o mala) la situación económica?; mientras que a través de los resultados que se presentan en la tabla 7.6.c se puede responder a la siguiente pregunta: ¿cómo perciben la situación económica los grupos de personas que se identifican con cada una de las posturas ideológicas?

De este modo, consultando la tabla 7.6.b se observa que, entre los que perciben la situación económica como buena, el grupo de los que se identifican con la izquierda está sobrerrepresentado en relación a la media, mientras que el grupo de los que se identifican con la derecha está subrepresentado. Lo contrario ocurre cuando se considera la columna de los que perciben la situación económica como mala, ya que entre ellos la proporción de los de derecha es más elevada que en relación a la media, mientras que baja ligeramente la proporción de los que se consideran de izquierdas. Al realizar estas comparaciones hay que tener presente que las bases numéricas de los que consideran buena o mala la situación económica es muy diferente, ya que los primeros sólo representan 22 casos, mientras que los segundos ascienden a 382. Por ello, los porcentajes calculados sobre esta última cifra son más fiables que los calculados en base a 22, ya que en este último caso se está operando con números absolutos muy pequeños.

La lectura de la tabla 7.6.c ofrece resultados diferentes a los anteriores. Así, por ejemplo, se puede observar que, tanto entre los que se identifican con la izquierda como entre los que se identifican con la derecha, las distribuciones de los que consideran la situación económica como buena, regular o mala son muy semejantes, mientras que entre los que se identifican con el centro político es mayor, relativamente, la proporción de los que consideran la situación como regular que en el resto, mientras que disminuye la proporción de los que consideran la situación económica como mala.

Otras consideraciones se podrían realizar al hilo de la comparación de las distribuciones condicionales que se incluyen en la tabla 7.6, sobre todo si situáramos tales resultados en el contexto de la estructura social de la sociedad española. Pero basten las breves consideraciones aquí formuladas para poner de manifiesto las posibilidades de este tipo simple, pero eficaz, de análisis estadístico.

7.2.b. *Distribuciones condicionales más complejas*

Con frecuencia, los sociólogos preparan tablas que, sin dejar de contener relaciones bivariables, son más complejas que las que hemos tenido ocasión de ver en las páginas anteriores. Así, se pueden confeccio-

nar tablas en las que las distribuciones condicionales hacen referencia a más de una variable independiente, y en las que la atención se concentra en la observación de la categoría o categorías más relevantes de la variable dependiente. La tabla 7.7 ilustra con claridad lo que venimos diciendo.

TABLA 7.7

Perfil del español interesado y desinteresado por la política
(Porcentajes de encuestas nacionales de 1976 y 1980)

Variables socieconómicas	*Con mayor interés*		*Con menor interés*	
	1976	*1980*	*1976*	*1980*
Sexo:				
% de hombres	43	43	33	27
% de mujeres	23	23	54	49
Edad:				
% menos de 25 años	43	47	29	27
% más de 64 años	19	13	62	57
Ocupación:				
Más cualificados	51	61	28	16
Menos cualificados	16	14	65	55
Municipio:				
Más de 500.000 habitantes	41	43	35	38
Menos de 2.000 habitantes	30	19	44	52

FUENTE: LÓPEZ PINTOR, Rafael: «El estado de la opinión pública», *REIS*, 13, 1981, pág. 30.

En un estudio sobre el interés por la política entre la población española, Rafael López Pintor trata de delimitar el perfil de los españoles interesados por la política y el de aquellos que no lo están. Para ello toma datos de dos encuestas sobre actitudes políticas de los españoles realizadas, respectivamente, en 1976 y 1980, y con ellos prepara la tabla 7.7. En esta tabla sólo se contienen datos referentes a los porcentajes de los individuos que manifiestan tener un interés alto y un interés bajo por la política, según el sexo, edad, ocupación y municipio de residencia de los entrevistados. A su vez, de estas cuatro variables independientes sólo se incluyen en la tabla las dos categorías que aparecen más relevantes para el análisis buscado. Naturalmente, el sexo aparece dicotomizado en «hombres» y «mujeres», pero de la variable edad sólo se destacan los «menores de 25 años» y los «mayores de 64 años»; de la variable ocupación, los «más cualificados» y los «menos cualificados»; mientras que, en lo referente al municipio de residencia, sólo se incluyen los que residen en municipios de más de 500.000 habitantes y los que lo hacen en municipios menores de 2.000 habitantes.

De este modo, se destacan tan sólo los valores extremos o perfiles más acusados de los interesados o no por la política. Así, de la observancia de la tabla anterior cabe concluir que el perfil tipo del español más interesado por la política es el de un hombre, de edad menor de veinticuatro años, con ocupación cualificada y residente en una gran ciudad. El perfil de la persona más desinteresada por la política sería el de una mujer, de edad superior a los sesenta y cuatro años, de baja cualificación ocupacional y residente en una zona rural. En una sola tabla, pues, se ha podido condensar un gran volumen de información, que ha permitido obtener conclusiones muy generales. Ni que decir tiene que este tipo de tabla es de gran interés analítico, sobre todo en aquellos estudios, como los realizados a través de encuesta, que permiten obtener un gran volumen de información que, necesariamente, se ha de resumir para poder alcanzar resultados globales. Obsérvese igualmente que los porcentajes no suman 100, ya que sólo se han incluido unas pocas categorías, y éstas, en consecuencia, no son exhaustivas.

7.3. CARACTERÍSTICAS DE UNA ASOCIACIÓN DE DOS VARIABLES

Cuando establecemos una clasificación cruzada de dos variables, nuestro interés se centra sobre todo en el conocimiento de la forma en que se distribuye la variable dependiente para las diferentes categorías de la variable independiente o causal. La forma en que se relacionan dos variables se denomina *asociación entre dos variables*. Volviendo a la tabla 7.3, se observa que a medida que disminuye el nivel de religiosidad se incrementa la proporción de personas que se identifican con las posiciones ideológicas de izquierda y, viceversa, a medida que aumenta el nivel de religiosidad se incrementa la proporción de personas identificadas con posiciones de derechas. Ese es, pues, el tipo de asociación que cabe observar entre las variables religiosidad e ideología.

Tal como señalan Loether y McTavish (*op. cit.*, 185), de igual modo que, al estudiar las distribuciones univariables, éstas quedaban caracterizadas mediante el estudio de su tendencia central, variación o dispersión y forma, asimismo, se puede caracterizar la relación entre dos variables mediante el estudio de las siguientes características: 1) *existencia* o no de una asociación; 2) la *fuerza* de la asociación; 3) la *dirección* de la asociación, y 4) la *naturaleza* de la asociación.

A continuación vamos a estudiar con cierto detalle las cuatro características y, más adelante, se desarrollarán algunos índices que se pueden utilizar para medirlas.

Ya hemos dicho anteriormente que existe una asociación entre dos

variables cuando la distribución de una variable difiere de algún modo entre las diversas categorías de la segunda variable. En su forma más general, y una vez calculados los porcentajes de la forma apropiada, se puede decir que existe una asociación entre dos variables cuando las correspondientes distribuciones condicionales porcentuales difieren en mayor o menor grado entre sí. Así, por ejemplo, en la siguiente tabla, que relaciona la variable «sexo» con la variable «interés por la política», se puede afirmar que existe una asociación entre ambas:

Interés por la política	*Varones*	*Mujeres*	*Total*
• Mucho	11	4	7
• Regular	27	19	23
• Poco	29	19	24
• Nada	32	54	43
• NS/NC	2	4	3
Total	100	100	100
	(576)	(616)	(1.193)

FUENTE: Barómetro de Opinión Pública, Sep. 1981, *REIS*, núm. 16, 1981, pág. 224.

En efecto, las distribuciones condicionales porcentuales correspondientes a varones y mujeres difieren claramente entre sí, ya que la proporción de los no interesados por la política es mucho mayor entre las mujeres, el 54 por 100, que entre los hombres, el 32 por 100. E, inversamente, las proporciones de los «mucho» o «regular» interesados por la política son superiores entre los varones que entre las mujeres. Por ello, la conclusión es de que existe una asociación entre la variable sexo y la variable interés por la política.

Anteriormente vimos también que las comparaciones se pueden realizar mediante el cálculo de los epsilón, ε_i, que son las diferencias porcentuales calculadas en la dirección en que se han realizado los porcentajes. Así, para la categoría «poco» interés por la política, $\varepsilon = 29 - 19 = 10$ por 100. Pues bien, cuando hay asociación entre dos variables, la mayor parte de los epsilón calculados al comparar las diferentes categorías son diferentes de cero, mientras que cuando todos los epsilón son cero no existe asociación alguna entre las variables. La idea de ausencia de asociación conviene ampliarla, por su interés para el cálculo de medidas de asociación.

Otra forma de decir si existe o no asociación entre dos variables consiste en comparar las *frecuencias observadas* en la tabla con las frecuencias que cabría esperar si no existiera asociación, o *frecuencias esperadas*. Si al comparar la tabla de datos reales con la tabla de no asociación no se observa diferencia alguna, cabe hablar entonces de que no

existe asociación alguna entre las dos tablas cuyas categorías se cruzaron en la tabla.

Por su interés para ulteriores cálculos, vamos a ver a continuación cómo se calcula una tabla o modelo de no asociación a partir de una tabla bivariable cualquiera. El problema consiste en calcular las frecuencias de cada celdilla a partir de los datos totales, de forma que las distribuciones condicionales no ofrezcan asociación alguna. Supongamos que partimos de la siguiente distribución de frecuencias absolutas entre dos variables dicotómicas:

	(*X*)		
(*Y*)	I	II	Total
I	*a*	*b*	52
II	*c*	*d*	40
Total	30	62	92

Para obtener los valores de *a*, *b*, *c* y *d* de forma que no exista relación entre las variables (*X*) e (*Y*) hay que partir del siguiente razonamiento. Para que no exista asociación entre ambas variables, tanto las filas como las columnas que componen el cuerpo de la tabla deben distribuirse, en términos proporcionales, de la misma forma que lo hacen los valores globales de las dos variables, esto es, de la forma que lo hacen los «totales».

En general, la frecuencia esperada f_e para una determinada celdilla se calcula multiplicando el total de la fila correspondiente a la celdilla por el total de la columna que corresponde igualmente a dicha celdilla, y dividiendo el producto obtenido por el número global *N* de casos. Esto es:

$$f_{eij} = \frac{n_i \cdot n_j}{N} \qquad [7.1]$$

en donde f_{eij} representa la frecuencia esperada de la celdilla correspondiente a la fila *i* y a la columna *j* de la tabla; n_i es el total para la fila *i*, y n_j es el total para la columna *j*, siendo *N* el número total de casos.

Aplicando la fórmula [7.1] a los datos que se contienen en la tabla anterior, se calculan de inmediato los valores de *a*, *b*, *c* y *d*:

$$a = \frac{52 \cdot 30}{92} = 16{,}9 \qquad b = \frac{52 \cdot 62}{92} = 35{,}0$$

$$c = \frac{40 \cdot 30}{92} = 13{,}0 \qquad d = \frac{40 \cdot 62}{92} = 27{,}0$$

Una vez calculados los valores esperados en cada casilla, se pueden comparar los valores observados f_o, o reales, de la tabla con los valores esperados f_e. La comparación se realiza restando el valor esperado de cada celdilla del valor observado de la celdilla correspondiente. Este valor, denominado *delta* $\Delta = f_o - f_e$, se calcula para cada celdilla de la tabla. Mientras algunos de los valores Δ obtenidos sean diferentes de cero, se puede hablar de la existencia de algún tipo de asociación entre las dos variables. Si todos los deltas son cero, entonces se puede afirmar que no existe asociación alguna entre las variables, o, dicho con otras palabras, existe *independencia estadística* entre las dos variables.

Ahora bien, no es lo mismo, desde el punto de vista de la asociación entre dos variables, que los valores epsilón o delta sean altos o bajos. Aquí conviene introducir la noción del *grado o fuerza de la asociación* entre dos variables. Cuando los valores epsilón o delta son elevados cabe hablar de un alto grado de asociación o de una fuerte asociación entre las variables, mientras que si tales valores son pequeños se trata de una débil asociación o de un bajo grado de asociación.

Existe un problema con la utilización de los valores de epsilón y delta, y es que resulta difícil determinar con precisión el significado de un valor determinado, aparte de revelar la existencia o no de una asociación, ya que no existe una escala con un valor mínimo y un valor máximo entre los que puedan variar los valores obtenidos de epsilón y delta. Por dicha razón se utilizan con mayor frecuencia otro tipo de índices «estandarizados» que varían, de una forma fija, predeterminada e interpretable, entre un valor mínimo de no asociación y un valor máximo de mayor asociación. Más adelante estudiaremos los índices o coeficientes estandarizados de mayor uso en la investigación empírica en sociología.

Por lo que se refiere a la tercera de las características enunciadas, la *dirección de la asociación*, sólo cabe hablar de ella cuando las variables se han medido, como mínimo, al nivel ordinal. Con variables nominales o clasificatorias no cabe hablar de dirección de la asociación. Cuando, en una tabla, la tendencia de variación conjunta de las dos variables es a que los valores altos de una variable se correspondan con los valores altos de la segunda variable (y los valores bajos se corresponden igualmente), cabe hablar de la existencia de una *asociación positiva*. Así, en el siguiente ejemplo, con datos ficticios entre el nivel de ingresos y el nivel de satisfacción general, la dirección de la asociación entre ambas variables es positiva, ya que, a mayor nivel de ingresos, más elevado es el nivel de satisfacción:

Nivel de satisfacción	*Nivel de ingresos*		
	Bajo	*Medio*	*Alto*
— Bajo	60	40	30
— Medio	30	40	45
— Alto	10	20	25

Por el contrario, cuando los valores superiores de una variable se corresponden con los valores bajos de la segunda, y los valores altos de ésta se corresponden con los valores bajos de aquélla, se dice entonces que la dirección de la asociación es negativa. Así, por ejemplo, al estudiar la relación existente entre el nivel de ingresos de los individuos y el grado de anomia que padecen se observa una asociación negativa, ya que los individuos de ingresos altos tienden a tener un grado menor de anomia que los individuos de ingresos bajos, que padecen un mayor grado de anomia, como se puede observar en el siguiente cuadro:

Grado de anomia	*Nivel de ingresos*		
	Bajo	*Medio*	*Alto*
— Bajo	20	40	55
— Medio	30	25	45
— Alto	50	35	20

Finalmente, nos queda por analizar la cuarta característica de la asociación entre dos variables. La *naturaleza* de una asociación se refiere a la forma general en que se distribuyen los datos en la tabla. Habitualmente, dicha forma general o modelo se describe mediante el examen de las distribuciones de los porcentajes. En unos casos la distribución es irregular, distribuyéndose las diferencias elevadas o las aproximaciones entre cada par de porcentajes de una manera desigual, mientras que en otros casos se produce una progresión uniforme de las diferencias porcentuales desde las categorías bajas a las altas de las variables. Cuando, al pasar de una categoría a otra de una variable, el número de casos tiende a incrementarse (o disminuir) de una forma bastante homogénea entre las correspondientes categorías de la otra variable, se produce una asociación «lineal», esto es, que los casos se concentran en la variable dependiente siguiendo una línea recta. Las asociaciones lineales simples tienen un gran valor en la estadística en general, y en la investigación sociológica en particular, como modelos de asociaciones simples, aunque con frecuencia los datos sociológicos se distribuyen siguiendo formas

curvilíneas o de otra naturaleza. Más adelante volveremos con mayor detalle a tratar este tema.

7.4. La obtención de medidas de asociación entre dos variables: Independencia estadística y asociación perfecta

Tal como se ha indicado anteriormente, el investigador necesita disponer de medidas que en un solo índice indiquen la existencia, grado y dirección de la asociación entre dos variables. Habitualmente, lo que se busca es una medida cuyos valores puedan variar a lo largo de una escala desde un valor mínimo, que indique una relación negativa, hasta un valor máximo, que indique una asociación positiva, pasando por el cero, que indicará la ausencia de asociación. Idealmente, tales valores límites se hacen coincidir con el -1 al $+1$, en el que el valor $+1$ indica una asociación perfecta en el sentido positivo, y el -1 una asociación perfecta en el sentido negativo, y el cero indica la existencia de independencia estadística.

Una medida de asociación del tipo que acabamos de describir es una medida de asociación estandarizada o tipificada, ya que los valores respectivos obtenidos mediante tal índice en diferentes tablas se pueden comparar entre sí. Así, por ejemplo, si al cruzar la variable «interés por la política» con la edad se obtiene una asociación de $+0{,}52$, y al hacer lo propio con la variable nivel de ingresos se obtiene una asociación de $+0{,}35$, podremos afirmar que la asociación entre edad e interés por la política es más fuerte que la asociación entre esta última variable y el nivel de ingresos. Dado que la comparación es la base de toda investigación científica, y la interpretación de los resultados obtenidos en las comparaciones es decisiva para lograr conclusiones relevantes, se entiende que el uso y desarrollo de medidas estandarizadas o tipificadas es muy importante en la investigación sociológica.

Las medidas estandarizadas o tipificadas de asociación suelen ser simples proporciones o cocientes *(ratios)* que son sensibles a los cambios que se producen en el grado de asociación y, en algunos casos, en la dirección y naturaleza de la misma. De lo que se trata es de conseguir índices que reflejen realmente la variación de los aspectos relevantes de las variables cruzadas y que sean más insensibles a las variaciones de características no relevantes para la asociación, como pueden ser el número de filas y columnas o el número total de casos en que se basan los porcentajes.

La medida más sencilla que se puede crear a partir de los valores delta Δ, o diferencias entre los valores observados y esperados, es la propia suma de todos los deltas que se pueden calcular al comparar las diferentes columnas y categorías que componen una tabla. De este modo se obtendría un sumatorio de todos los deltas, $\Sigma\Delta = \Sigma(f_o - f_e)$, que

es igual a la suma de todas las diferencias que se pueden establecer entre los valores observados y esperados. Ahora bien, este índice es muy deficiente, ya que depende, en primer lugar, del tamaño de los valores esperados y, además, los valores deltas particulares se pueden neutralizar entre sí al sumar cantidades afectadas de signo contrario.

Estas limitaciones se superan en buena medida si, en lugar de sumar los diferentes deltas, se suman sus cuadrados (con lo que desaparece el problema de los signos) y se divide cada delta al cuadrado por el valor esperado para cada celdilla, con lo que se controlan en cierto modo los efectos distorsionantes que se producen al considerar diferentes números de casos. Los valores así obtenidos dan lugar a una medida de asociación de las denominadas «de distribución libre», ya que no depende de condiciones especiales que deban cumplir los datos. En el capítulo anterior ya estudiamos la prueba del chi-cuadrado para una sola variable, que se basa en los cálculos que acabamos de describir, de igual manera que encontraremos el chi-cuadrado cuando estudiemos las pruebas de decisión estadística con dos o más variables. Pero ahora volvamos a la medida de asociación que acabamos de estudiar. La medida resultante de sumar todos los cocientes anteriores para cada celdilla se denomina *chi-cuadrado* (x^2):

$$x^2 = \sum\left(\frac{\Delta^2}{f_e}\right) \qquad [7.2]$$

Este índice se utiliza más en la estadística inferencial, para la prueba de hipótesis, que en la estadística descriptiva, para medir el grado de asociación entre dos variables, ya que presenta ciertos problemas al tratar de estandarizar sus valores. Sin embargo, al tener una «distribución libre», se convierte en una prueba muy útil para variables nominales y ordinales.

El coeficiente chi-cuadrado es siempre un número positivo, y se hace cero en las tablas en las que no hay asociación entre las variables. Sin embargo, el límite superior del coeficiente x^2 no es fijo, sino que vale $N(K-1)$, en donde N es el tamaño de la muestra y K es el número de filas o columnas en la tabla, según sean unas u otras las que representan el número más pequeño. Para una tabla 2×2, el límite superior de la magnitud de x^2 es N. Por tanto, dadas dos tablas que tengan una asociación idéntica en su forma porcentual, si una de ellas se basa en un número doble de casos que en la otra, su valor de x^2 será el doble que en la tabla basada en el menor número de casos.

Por todo ello, el coeficiente de chi-cuadrado no se utiliza como medida de asociación, aunque, como se ha dicho anteriormente, sí se utiliza, y ampliamente, en la estadística inferencial. Otros coeficientes basados en chi-cuadrado tratan de aprovecharse de las ventajas que ofrece dicho coeficiente, a la vez que tratan de superar, mediante de-

terminadas correcciones, sus deficiencias o limitaciones. Así, se puede utilizar el coeficiente de «contingencia cuadrática media» o *fi-cuadrado*, ϕ^2, que se define simplemente como el valor de chi-cuadrado dividido por N:

fi-cuadrado

$$\phi^2=\frac{x^2}{N} \qquad o \qquad \phi=\sqrt{\frac{x^2}{N}} \qquad [7.3]$$

El valor de fi varía entre 0 —para el caso de independencia estadística— a un máximo de +1 —cuando existe una asociación perfecta—, en cualquier tabla de tamaño $2\times K$, pudiéndose interpretar su magnitud como una medida del grado de asociación. Sin embargo, presenta el inconveniente de que, en tablas que contengan más de dos categorías en cada variable, el valor máximo de fi sobrepasa la unidad, dado que el límite superior de x^2, $N(K-1)$, se convierte en tal caso en un valor superior a N. El valor máximo de $\phi^2=K-1$, en donde K representa el número más pequeño, bien de las filas o bien de las columnas.

El propio inventor del chi-cuadrado, el inglés Karl Pearson (1857-1936), considerado por muchos como el auténtico fundador de la estadística moderna, suministró una solución parcial a las anteriores limitaciones, mediante el desarrollo del *«coeficiente de contingencia» o coeficiente C de Pearson*. La fórmula para C es la siguiente:

Coeficiente de contingencia

$$C=\sqrt{\frac{x^2}{x^2+N}} \qquad [7.4]$$

El coeficiente C no puede ser superior a la unidad, con independencia del tamaño de la tabla, ya que el coeficiente x^2 aparece tanto en el denominador como en el numerador, y aquél es siempre mayor que éste, ya que contiene la suma x^2+N, que será siempre superior a x^2. En su valor mínimo, el coeficiente C puede llegar a ser cero cuando, en los casos de ausencia de asociación, el valor de x^2 sea también cero, pero nunca alcanza exactamente la unidad, aunque hubiera asociación perfecta, por la razón anteriormente apuntada de que el denominador es siempre superior al numerador en la expresión [7.4]. Para una tabla cuadrada, es decir, una tabla en la que el número de filas sea igual al número de columnas, el valor máximo de C se puede calcular a partir de la expresión siguiente:

Para tabla cuadrada (= nº de filas y columnas)

$$C \text{ máximo}=\sqrt{\frac{K-1}{K}}$$

en donde K es el número de filas (o de columnas) en una tabla cuadrada. Así, por ejemplo, para una tabla 2×2, el valor máximo de C es 0,707; para una tabla 4×4, el valor máximo de C es 0,87, y para una tabla 5×5, el C máximo es 0,89. Así, pues, utilizando el coeficiente C no se pueden realizar comparaciones con esta medida de asociación entre tablas de diferentes tamaños.

Otros autores han tratado de mejorar la obtención de un coeficiente de asociación que pueda utilizarse para comparar tablas de diferente tamaño, es decir, que se pueda disponer de un coeficiente suficientemente estandarizado o normatizado. El *coeficiente T de Tschruprow* corrige el problema del límite superior de C mediante una ligera modificación del denominador de la expresión [7.4], de tal modo que incluya un valor que refleje el número de celdillas de la tabla. En otras palabras, se trata de introducir el concepto de los grados de libertad en la fórmula del coeficiente de asociación. Parece ser que el propio Pearson nunca llegó a comprender el concepto de grados de libertad en relación tanto con el chi-cuadrado como en relación con el cálculo de los errores de probabilidad (H. M. Walker, 1978, pág. 695). Por esa razón han tenido que ser otros autores los que se preocuparon de obtener medidas de asociación mejor normatizadas. Recordemos que, en una tabla de n filas y m columnas, los *grados de libertad** $df=(n-1)\cdot(m-1)$, es decir, es igual al número de filas menos uno multiplicado por el número de columnas menos uno.

Pues bien, el coeficiente T de Tschruprow se define del siguiente modo:

Coeficiente T.

$$T=\sqrt{\frac{x^2}{N(df)}} \qquad [7.5]$$

El coeficiente T representa un avance en la búsqueda de una medida de asociación que esté adecuadamente estandarizada o normatizada para cualquier tipo de tabla. En efecto, el límite superior de T vale la unidad, con independencia del tamaño de la tabla, en tanto que ésta sea cuadrada, es decir, que el número de filas sea igual al número de columnas. Ahora bien, para tablas que no son cuadradas, el valor de T no puede alcanzar la unidad, aunque su valor máximo sea constante para tablas con idénticos grados de libertad.

Otro coeficiente, la *V de Cramer*, trata de resolver el problema de la estandarización o normatización mediante la sustitución en la ex-

* El concepto de grados de libertad lo hemos estudiado en el capítulo introductorio a la estadística inferencial. De una forma intuitiva, su concepto se puede entender en el estudio de las tablas bivariales, señalando que en una tabla 2×2 en donde $df=(2-1)(2-1)=1$, se puede conocer una frecuencia esperada conociendo una frecuencia observada en una celdilla. El resto se puede calcular por substracción, ya que los marginales son fijos y, por tanto, conocidos. Se tiene, pues, un grado de libertad en elegir la frecuencia de una celdilla antes de que se puedan determinar las restantes frecuencias. En una tabla 3×3, se han de elegir cuatro frecuencias de celdillas antes de determinar el resto, esto es, tiene cuatro grados de libertad, y así para otros tamaños de la tabla.

presión de T de los grados de libertad df por un valor t que representa el número más pequeño de las dos cantidades, $n-1$ o $m-1$, siendo n y m el número de filas y columnas, respectivamente. Así, pues, la fórmula de la V de Cramer es como sigue:

$$V=\sqrt{\frac{\chi^2}{N \cdot t}} \qquad [7.6]$$

El coeficiente V de Cramer siempre puede alcanzar el límite superior de la unidad, con independencia del tamaño de la tabla, y vale cero cuando no existe asociación alguna. Por todo ello se trata de un coeficiente mejor normatizado que los vistos con anterioridad.

Pero, tal como señalan Loether y McTavish (*op. cit.*, pág. 197), el problema de la estandarización o normatización representa tan sólo una parte del problema que surge al desarrollar un número índice que sea útil para resumir la asociación en una tabla. Al menos se pueden citar otras dos características que debe reunir una buena medida de tal tipo. En primer lugar, el número que se obtenga debe poder interpretarse de una forma intuitiva y, en segundo lugar, el significado de 1,0, la norma de la «asociación perfecta», debe poder ser definible. Por lo que se refiere a los coeficientes basados en el chi-cuadrado, la magnitud del coeficiente sólo puede ser entendida dentro de una escala entre el 0 y el 1, de tal modo que cuanto más cerca de 1 se encuentra la magnitud obtenida del coeficiente, mayor será el grado de asociación. Sin embargo, no se pueden interpretar tales tipos de coeficientes en el sentido de, por ejemplo, la variación porcentual de una variable que es explicada por otra, o como la proporción de error predictivo que se puede reducir a través del conocimiento previo de una de las variables. Precisamente en el próximo capítulo estudiaremos tales tipos de medidas, que son de mayor interés para el sociólogo, que busca la «explicación» de la relación entre variables.

Por lo que se refiere al significado de los valores extremos del coeficiente que mide el grado de asociación de dos variables, ya hemos señalado anteriormente que el 0 representa la ausencia de asociación, o la independencia estadística entre las dos variables. En cuanto al límite superior 1, anteriormente hemos introducido el concepto de «asociación perfecta» para su interpretación, aunque no hemos explicado lo que tal concepto significa. Veámoslo ahora con mayor detalle.

Se dice que una tabla bivariable refleja una *asociación perfecta* cuando todos los casos de la tabla se concentran en una diagonal, lo que significa que cada valor de una variable se encuentra asociada con un solo valor de la segunda variable, de tal modo que para cualquier categoría de la variable independiente sólo será diferente de cero una celdilla de la variable dependiente, mientras que el resto de las celdillas

serán ceros. En una tabla 2×2, esto significa que las celdillas de una de las diagonales contienen valores y las de la segunda diagonal contienen cero, como se observa a continuación:

a)

	(X)	
(Y)	I	II
I	*A*	0
II	0	*B*

b)

	(X)	
(Y)	I	II
I	0	*A*
II	*B*	0

La primera tabla *a)* refleja una «asociación perfecta positiva», ya que se corresponden las mismas categorías de ambas variables, mientras que la tabla *b)* refleja una «asociación perfecta negativa», pues la correspondencia se produce entre las categorías opuestas de ambas variables.

Veamos a través de un ejemplo hipotético el funcionamiento de estos conceptos. Supongamos que tratamos de contrastar la teoría de que los delitos por consumo de drogas son más elevados en las grandes ciudades en relación a las pequeñas ciudades y zonas rurales. Pues bien, el modelo de la asociación perfecta significaría que, al distribuir los datos en una tabla que cruzase la variable «frecuencia de delitos por consumo de drogas» por la variable «tamaño del lugar de residencia», todos los delitos de tal tipo se concentrarían en las grandes ciudades, mientras que en las ciudades pequeñas no se produciría ninguno de tales delitos. Cualquier desviación en relación a esta forma de distribuirse los datos correspondientes a las dos variables significa una asociación no perfecta.

Existe todavía una definición menos restrictiva de asociación perfecta, y que consiste en una distribución tal que, por lo que se refiere a una tabla 2×2, sólo sea cero la frecuencia de una de las celdillas, mientras que las otras tres celdillas tienen una frecuencia diferente de cero. El *coeficiente Q de Yule* se basa en esta consideración menos restrictiva de la asociación perfecta para establecer otra medida de asociación para tablas 2×2. Supongamos que las frecuencias de las cuatro celdillas de una tabla 2×2 son *a*, *b*, *c* y *d*, como se observa en la figura:

	(X)	
(Y)	I	II
I	*a*	*b*
II	*c*	*d*

Pues bien, el coeficiente Q de Yule se calcula a partir de los productos cruzados de las celdillas de una de las diagonales *ad* y de las celdillas de la segunda diagonal *bc*. El coeficiente Q de Yule se calcula mediante una fórmula como sigue:

$$Q=\frac{ad-bc}{ad+bc} \qquad [7.7]$$

Cuando la frecuencia de una de las celdillas sea cero, entonces el valor de Q es $+1{,}0$ o $-1{,}0$, según la dirección de la asociación. El coeficiente Q se puede utilizar con variables nominales y, cuando alcanza el valor de la unidad en una tabla 2×2, refleja la existencia de una asociación perfecta.

7.4.1. *Medidas simétricas y asimétricas de asociación*

Finalmente, vamos a señalar una última distribución de las medidas de asociación que tiene interés para la investigación sociológica. Hay medidas de asociación que distinguen entre la variable independiente (o «causa») y la variable dependiente («efecto»), mientras que otras medidas de asociación no realizan tal distinción.

Pues bien, a las medidas de asociación que no distinguen entre variables independientes o dependientes se les denomina *medidas simétricas*. Tales medidas reflejan tan sólo la fuerza (y dirección) de la relación entre dos variables, y no distinguen entre los papeles asignados a cada variable. Los coeficientes vistos con anterioridad, tales como la Q de Yule, el coeficiente fi, la C de Pearson, la V de Cramer o la T de Tschruprow son ejemplos de medidas simétricas de asociación.

Por otro lado, hay medidas de asociación que requieren para su cálculo que se distinga previamente entre la variable independiente y la variable dependiente. Se trata de *medidas asimétricas de asociación*, que están orientadas, en general, a medir la capacidad e influencia de una variable independiente en la predicción de los valores de la variable dependiente. Buena parte de los coeficientes que vamos a estudiar en los próximos capítulos son de tipo asimétrico, aunque ya en este mismo capítulo hemos tenido ocasión de estudiar una de tales medidas. En efecto, el coeficiente epsilón, que, como se recordará, es una simple diferencia entre porcentajes, ofrece diferentes valores según sea el sentido en que se calculen los porcentajes, es decir, según sea una u otra la variable que se considera independiente. Naturalmente, al variar las bases sobre las que se calculan los porcentajes, así variarán los valores de epsilón. De todas maneras, este coeficiente apenas se utiliza en la práctica de la investigación sociológica, porque no se trata de una medida normatizada, como las que veremos en el próximo capítulo.

7.5. Terminología

Se recomienda la memorización y comprensión del significado de cada uno de los términos y conceptos siguientes:

- Distribución porcentual bivariable.
- Distribución condicional.
- Asociación entre dos variables:
  - Existencia de la asociación.
  - Fuerza o grado de la asociación.
  - Dirección de la asociación.
  - Naturaleza de la asociación.
- Frecuencias observadas.
- Frecuencias esperadas.
- Independencia estadística.
- Asociación perfecta.
- Asociación positiva.
- Asociación negativa.
- Coeficiente epsilón.
- Coeficiente delta.
- Coeficiente fi.
- Coeficiente chi-cuadrado.
- Coeficiente C de Pearson.
- Coeficiente T de Tschruprow.
- Coeficiente V de Cramer.
- Coeficiente Q de Yule.
- Grados de libertad.
- Medidas simétricas de asociación.
- Medidas asimétricas de asociación.

EJERCICIOS

1. De los siguientes pares de variables: ¿Cuáles están formados por variables independientes entre sí (es decir, no es posible *a priori* especificar una ordenación causal o temporal entre ellas)?; ¿cuáles están formados por variables que están relacionadas entre sí condicionalmente? Para estos últimos pares de variables, especificar para cada par qué variable, desde un punto de vista lógico, antecedente de la otra.

   1) Tamaño de familia y religiosidad de los cónyuges.
   2) Edad y región de nacimiento.
   3) Interés por la política y nivel de educación.

4) Fallecimientos por crisis cardiacas y ocupación.
5) Consumo de alcohol y lugar de residencia.
6) Absentismo laboral y *status* civil.
7) Preferencia política y sexo.
8) Religiosidad y nivel de ingresos.
9) Número de libros leídos al año y preferencia política.
10) Afiliación sindical y origen social.

2. En un estudio sobre las preferencias del público por lo que respecta a los programas de televisión, se obtuvieron los siguientes resultados al agrupar a la población encuestada en tres segmentos según el nivel de su *status* socioeconómico:

	Status socioeconómico		
Programa preferido	*Alto*	*Medio*	*Bajo*
Seriales	33	108	75
Informativos	44	90	32
Teatro	15	66	30
Películas cine	85	288	202

A partir de la distribución bivariable de frecuencias de la tabla anterior, calcular las distribuciones porcentuales en el sentido vertical, horizontal y en relación al total de casos. Explique en sus propias palabras lo que expresan los porcentajes en cada tabla acerca de la existencia, grado, dirección y naturaleza de la asociación de la tabla. ¿Qué se puede decir acerca del efecto del *status* socioeconómico de la población en relación a la preferencia de los programas de televisión? ¿Cómo atraen los diferentes programas a los tres grupos de población?

3. En una encuesta de la juventud, realizada en 1982, la identificación religiosa de los jóvenes se distribuyó del siguiente modo teniendo en cuenta su edad:

	Edad (años)		
Religiosidad	19-20	17-18	15-16
Católico practicante	343	431	468
Católico no practicante	657	576	411
Otras religiones	15	11	12
No creyente	91	71	28
Indiferente	205	133	90

De las dos variables, edad y religiosidad, ¿cuál se puede considerar antecedente de la otra?
Calcular los porcentajes en la dirección de la variable antecedente o causal y explique en sus propias palabras lo que expresan los porcentajes acerca de la existencia, grado, dirección y naturaleza de la asociación de la tabla.

4. En una encuesta de opinión pública sobre la policía, la evaluación del trabajo de la policía en el cumplimiento de su deber se distribuyó del siguiente modo, teniendo en cuenta las siguientes variables: sexo, nivel de estudios, ideología, partido político, región y hábitat.

	Evaluación del trabajo de la policía			
Variables	*Bueno*	*Regular*	*Malo*	*NS/NC*
— Edad:				
Menos de 21 años	26	41	24	9
Más de 60 años	49	22	6	23
— Sexo:				
Hombre	41	35	14	10
Mujer	43	30	10	17
— Nivel de estudios:				
Primarios	42	35	9	13
Universitarios	29	36	28	7
— Ideología:				
Izquierda	29	43	21	7
Derecha	60	26	7	7
— Partido que votó:				
P.C.E.	27	40	26	7
C.D.	64	24	4	7
— Región:				
Andalucía	51	24	12	13
Vascongadas	15	36	32	17
— Hábitat:				
Rural	53	28	6	13
Metropolitano	39	38	14	9

A la vista de las anteriores distribuciones porcentuales, ¿qué se puede decir sobre la evaluación que realizan los distintos grupos sociales sobre el trabajo de la policía?

5. En una encuesta sobre actitudes de la población hacia el aborto, las opiniones sobre la legalización o prohibición del aborto se distribuyeron del siguiente modo, teniendo en cuenta la identificación ideológica de la población:

	¿Debe permitirse el aborto?		
Ideología	*En ningún caso*	*Por necesidad*	*Por decisión libre*
Izquierda	100	280	370
Centro	250	410	90
Derecha	370	280	60

A la vista de la anterior distribución, ¿se puede afirmar que existe asociación entre ambas variables? ¿De qué tipo es? En caso afirmativo, calcular el grado de asociación mediante el coeficiento de contingencia C de Pearson. Comparar el valor obtenido con el valor máximo de C que se podría obtener para una tabla del tamaño como la presente.

BIBLIOGRAFIA

JIMÉNEZ BLANCO, J., et al.: *La Conciencia Regional en España*, Madrid, CIS, 1977.

LOETHER, H. J. y D. G. MCTAVISH: *Descriptive Statistics for Sociologists*, Boston, Ally & Bacon, 1974.

LÓPEZ PINTOR, R.: «El estado de la opinión pública y la transición a la democracia», *REIS*, núm. 13, 1981, págs. 7-47.

WALKER, Helen W.: «Karl Pearson», en W. H. Kruskal y J. M. Tanur (eds.) *International Encyclopedia of Statistics*, New York, Free Press, 1978, págs. 691-698.

ZEISEL, Hans: *Dígalo con números*, México, Fondo de Cultura Económica, 1974. (e. o. 1947).

Capítulo 8

MEDIDAS DE ASOCIACION PARA VARIABLES NOMINALES Y ORDINALES

Son muy variadas las medidas de asociación de que puede disponer un sociólogo interesado en el estudio de relaciones bivariables. En el capítulo anterior tuvimos ocasión de estudiar algunas de ellas basadas en el valor de delta, o diferencia entre la frecuencia observada y la frecuencia esperada. Pero algunos de los coeficientes estudiados en dicho capítulo no son de interés para el investigador social, ya que no están «normatizados» y, por lo tanto, no está recomendada su utilización comparativa entre diferentes tablas, y menos aún la interpretación del carácter de la asociación. En el presente capítulo vamos a estudiar las medidas de asociación basadas en el criterio de «reducción proporcional del error», por ser las más utilizadas por los sociólogos, y ello para las relaciones entre variables medidas a nivel nominal y a nivel ordinal. En el próximo capítulo continuaremos con el estudio de las medidas basadas en el mismo criterio de reducción del error, pero para el caso de variables de intervalo, con lo que abordaremos uno de los temas centrales de la estadística, el estudio de la regresión simple.

Dado el carácter introductorio del presente libro, no vamos a estudiar las medidas de asociación apropiadas para situaciones especiales, porque esperamos que, con el bagaje de técnicas estadísticas que se presentan aquí, el estudiante de sociología puede pasar a realizar por sí mismo una investigación empírica sólida. Por ello remitimos al lector interesado en medidas de asociación especiales a otros libros, tales como el de Freeman (1971), y algunos otros trabajos que se citan en la bibliografía, para que pueda estudiar y conocer las mismas.

8.1. MEDIDAS DE ASOCIACIÓN BASADAS EN EL CRITERIO DE «REDUCCIÓN PROPORCIONAL DEL ERROR» (RPE)

Un simple repaso al estudio de las diferentes medidas de asociación disponibles para el estudio de datos pone rápidamente de manifiesto la dificultad de encontrar un principio lógico consistente que sea