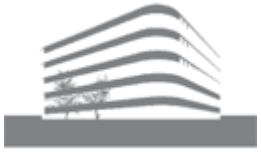


Análisis de segmentación : concepto y aplicaciones

Author(s): Escobar, Modesto
Date 1992
Type Working Paper
Series Estudios = Working papers / Instituto Juan March de Estudios e Investigaciones, Centro de Estudios Avanzados en Ciencias Sociales 1992/31
City: Madrid
Publisher: Centro de Estudios Avanzados en Ciencias Sociales

EL ANÁLISIS DE SEGMENTACIÓN: CONCEPTO Y APLICACIONES

Modesto Escobar

Estudio/Working Paper 1992/31
Enero 1992

Modesto Escobar es profesor titular de Sociología en la Universidad Complutense de Madrid y profesor de Metodología en el *Centro de Estudios Avanzados en Ciencias Sociales* del Instituto Juan March de Estudios e Investigaciones en Madrid.

1. Introducción*

Cuando en un cuestionario se desea explicar por qué los entrevistados dan contestaciones distintas a las preguntas, se recurre a una serie de cruces que permitan ver la asociación existente entre unas y otras variables. No se trata de cruzar cada pregunta con el resto, sino de seleccionar una serie de hipótesis plausibles con el conocimiento previo, teórico o empírico, de la realidad que se está investigando, y, de acuerdo con ellas, realizar los cruces que pongan a prueba las conjeturas. Una manera de facilitar la tarea de selección de variables relevantes en la explicación de la contestación a una pregunta dada es la técnica del análisis de segmentación, que proporciona además una descripción de las diferencias que los distintos grupos de una muestra pueden presentar en un determinado rasgo. Es esta una técnica de dependencia entre variables. En su uso se distinguen, por un lado, una o varias variables cuya distribución se desea explicar y, por el otro, un conjunto de variables, nominales u ordinales, con *status* de independientes. Estas reciben el nombre de predictores y tienen la finalidad de formar grupos que sean muy distintos entre sí en la variable o variables dependientes.[1]

Póngase como ejemplo claro que se desee describir en un pueblo pequeño quién lleva un determinado tipo de ropa. Para simplificar, tómese una prenda muy fácil de segmentar como es la falda. Entre las posibles variables que mejor pueden explicar quién la lleva y quién no, no es difícil reconocer que es el sexo el mejor predictor, pues prácticamente ningún hombre usa este tipo de prenda. La ejecución de la segmentación implicaría no contentarse con una sola variable y buscar otras que ayuden a distinguir mejor a los distintos usuarios de estas ropas. Es evidente que si ningún hombre la usa, este grupo es totalmente homogéneo en esta variable y, por tanto, no procede seguir con la segmentación. Pero en el caso de las mujeres, sí se pueden encontrar nuevas variables que nos distingan grupos diferentes en uso de ropa. Parece claro que la edad juega un papel importante: es bastante difícil ver a mujeres mayores con pantalones, mientras que entre las jóvenes el uso de éstos es muy habitual. Por tanto, si no se introducen nuevas variables, la población del

* Francisco Alvira, Daniel Peña, Javier Sánchez Carrión y Juan Carlos Rodríguez leyeron con interés una versión previa de estas páginas. Aunque muchos de sus valiosos comentarios estén recogidos, la responsabilidad de los errores y defectos del texto es exclusiva del autor.

[1] A menudo se confunde esta técnica con el análisis de conglomerados. Aunque las funciones clasificadoras son muy similares, se distinguen fundamentalmente en dos aspectos: a) La segmentación trabaja para la clasificación con *grupos de sujetos* (hombres, mujeres, jóvenes, personas de izquierda, practicantes de una determinada religión, solteros, casados…), seleccionando a aquellos que presentan características significativamente muy distintas en una o varias variables dependientes; el análisis de conglomerados trabaja con *individuos,* agrupando o distinguiendo a éstos en función de sus valores en un conjunto de variables, b) En al análisis de conglomerados no hay distinción entre variables dependientes e independientes, sino que todas ellas, con mayor o menor peso, sirven para clasificar a los sujetos; en el análisis de segmentación es necesario distinguir entre la variable dependiente que se desea explicar y las posibles variables independientes que puedan dar cuenta de ella.

pueblo quedaría segmentada en tres grupos: el de los hombres, donde nadie usa faldas; el de las mujeres jóvenes, con un porcentaje medio de portadoras de esta prenda, y el de mujeres mayores, cuya probabilidad de verlas con falda es muy alta.

El propósito de este artículo es doble: Primero, explicar sin demasiados ambages estadísticos y a través de distintos ejemplos, unos reales[2] y otros simulados, la lógica de esta técnica de análisis multivariado. Con este fin, se expondrá el análisis de segmentación a través de uno de sus algoritmos basado en el estadístico χ^2 y especialmente indicado cuando la variable dependiente es de tipo nominal. Se procederá a explicar los pasos lógicos de esta técnica: reducción de categorías, selección de predictores y detención de la segmentación. A continuación, en la segunda parte, se comparará este algoritmo con otros que operan con variable nominal y con el más conocido *AID, Automatic Interaction Detection* (Detección automática de la interacción), basado en el estadístico η^2. El segundo objetivo consiste en ofrecer una visión amplia de las aplicaciones de este análisis en el campo de las ciencias sociales. Aunque la opinión mayoritaria sobre esta técnica la concibe como una herramienta para la descripción o la exploración de datos, al final de esta presentación se plantean pistas para el análisis causal. Siguiendo esta clasificación de fines, en la tercera parte se ofrecen varios ejemplos de cómo se interpretan los resultados de la técnica de segmentación.

2. La lógica del análisis de segmentación. El algoritmo CHAID

Tradicionalmente, el análisis de segmentación se ha reducido al estudio de variables dependientes cuantitativas, utilizando el algoritmo presentado por Morgan y Sonquist (1963). Aquí, sin embargo, se centrará la atención en una derivación de esta técnica que se distingue por utilizar, en lugar de la suma cuadrática intergrupos, el estadístico χ^2 para la selección de los mejores predictores. De esta forma, la variable dependiente a utilizar debe estar medida en escala nominal.

Los pasos lógicos que deben seguirse para realizar esta tarea son los siguientes:

a) Preparación de las variables. Tarea del analista, que debe seleccionar una variable dependiente que sea de interés para el análisis y elegir un conjunto de posibles predictores relevantes (variables nominales u ordinales con pocas categorías, preferiblemente menos de diez) que permitan realizar una descripción y explicación óptima de la primera variable.

b) Agrupación de las categorías de los predictores, en el caso de que éstas tengan un perfil similar de la variable dependiente.

[2] Los ejemplos reales proceden de las investigaciones realizadas por cl CIRES (Centro de Investigaciones sobre la Realidad Social). El ejemplo cuya variable dependiente es el aborto se realizó en octubre de 1990 y el que utiliza la opinión sobre la titularidad de la sanidad se aplicó a los entrevistados en noviembre del mismo año. Ambos estudios poseen sendas muestras de 1200 individuos extraídas de la población española con más de 18 años.

c) Primera segmentación, que consiste en la selección del mejor predictor de la variable dependiente.

d) Segunda segmentación. Para cada segmento formado en el paso anterior, se busca el mejor de los predictores, cuyos valores han sido previamente agrupados de la misma forma que en el paso b).

e) Sucesivas segmentaciones. Se procede de forma similar al paso anterior en cada grupo formado por la segmentación previa.

Supóngase que se quieren formar grupos homogéneos, llamados en adelante segmentos, respecto de la aprobación del aborto en el supuesto de que un matrimonio no desee tener más hijos. Esta será la variable dependiente, con tres posibles valores: "Lo aprueba", "lo desaprueba" y "no sabe/no contesta". Para formar grupos homogéneos con esta técnica, se ha de elegir una serie de características medidas nominal u ordinalmente. En este caso, por ejemplo, sexo ("hombre","mujer"), edad ("menos de 46 años", "más de 45"), e ideología ("izquierda", "centro", "derecha").

Tabla 1

CRUCE DE OPINION SOBRE EL ABORTO SEGUN SEXO, EDAD E IDEOLOGIA

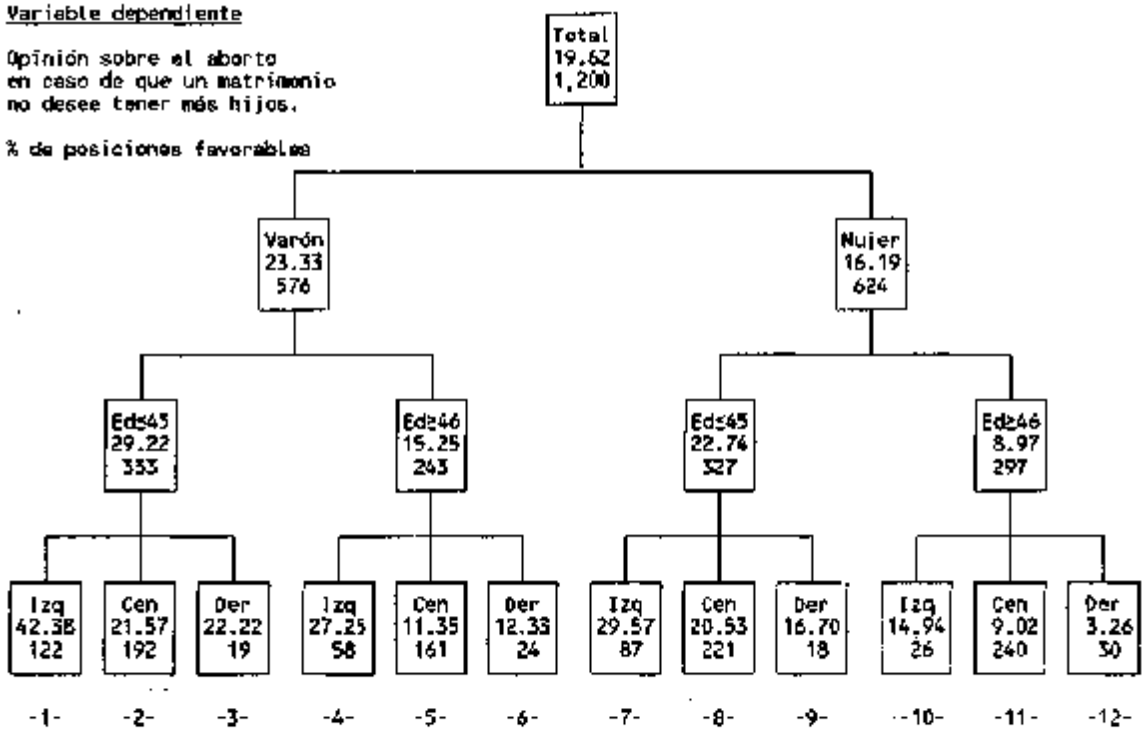
	SEXO											
	Varón						Mujer					
	EDAD						EDAD					
	18-45			46 y más			18-45			46 y más		
	IDEOLOGIA POLITICA			IDEOLOGIA POLITICA			IDEOLOGIA POLITICA			IDEOLOGIA POLITICA		
	Izq.	Cent.	Der.	Izq.	Cent.	Der.	Izq.	Cent.	Der.	Izq.	Cent.	Der.
ABORTO SI SE DESEA												
Si...	42.4%	21.6%	22.2%	27.3%	11.3%	12.3%	29.6%	20.5%	16.7%	14.9%	9.0%	3.3%
No..	55.0%	72.6%	72.0%	72.7%	81.7%	79.2%	62.5%	75.0%	77.7%	85.1%	84.1%	96.7%
NS/NC	2.6%	5.8%	5.7%		6.9%	8.4%	7.9%	4.5%	5.6%		6.9%	
TOTAL	122	192	19	58	161	24	87	221	18	26	240	30

En la Tabla 1 se pueden contemplar 12 segmentos (columnas) distintos formados por el cruce de las categorías de las tres variables predictoras (2 de sexo por 2 de edad por 3 de ideología). Cada uno de ellos está caracterizado por un tamaño (fila correspondiente al total) y tres porcentajes relativos a cada uno de los valores de la variable dependiente, en este caso, posición ante el aborto.

El segmento más numeroso es el correspondiente a las mujeres de centro con más de 46 años (n=240), seguido por el de las mujeres jóvenes de la misma ideología (221). Los

grupos de hombres con más componentes son el de los jóvenes de centro (192) y el de los hombres con más de 46 años de la misma ideología (161). Por el contrario, los grupos menos numerosos son el de los hombres y el de las mujeres jóvenes de derecha (19 y 18 sujetos respectivamente). Al observar la variable dependiente (el porcentaje de los que aprueban el aborto) se obtiene un perfil distinto para cada uno de los 12 segmentos formados por las tres variables predictoras: Los más favorables al aborto (42.4%) son los jóvenes varones de ideología de izquierdas y el grupo con menor porcentaje de sujetos que aprueban esta práctica (3.3%) es el de la mujeres con más de 46 años e ideología de derechas.

Tabla 2



La Tabla 1 es en realidad una tabla de contingencia o cruce formado por cuatro variables dispuestas en cuatro dimensiones. La técnica de la segmentación tiene una estructura similar. En la Tabla 2 se muestra una pseudo-segmentación, basada en los datos de la Tabla 1. En cada rectángulo se incluye el *valor* de la variable predictora que conforma el segmento determinado, el *porcentaje* de sujetos del grupo que poseen un determinado valor de la variable dependiente, generalmente el primero, en este caso el porcentaje de favorables al aborto, y el *número de casos* que forman el grupo. Las cifras incluidas en los 12 rectángulos numerados en la base inferior son idénticas a la de los primeros porcentajes y a la de los totales de las 12 columnas de la Tabla 1. Para que este resultado fuese una verdadera segmentación, las divisiones se tendrían que haber realizado de una forma

automática y jerárquica, y sólo en el supuesto de que hubiera existido significación estadística.

Hay variados procedimientos para llevar a cabo la segmentación. A continuación se presenta con mayor detalle el algoritmo llamado CHAID *(Chi-squared Automatic Interaction Detection).* Esta técnica, desarrollada por Cellard *et al.* (1967), Bourouche y Tennenhaus (1972), Kass (1980) y Magidson (1989), quien la ha adaptado para el SPSS, tiene como principal característica el que la variable dependiente es de naturaleza nominal. Al igual que otras prácticas de segmentación, las operaciones elementales que ésta realiza son: a) la agrupación de las categorías de las variables predictoras; b) la comparación de efectos entre distintas variables, y c) la finalización del proceso de segmentación.

2.1. *Reducción de las categorías más discriminantes de cada predictor*

Este primer paso consiste en seleccionar las categorías de una variable predictora que realmente discriminan a los sujetos en la variable dependiente. Suponiendo que una determinada variable tuviera c valores, se trata de convertirlos a un número $k \leq c$ que reduzca la complejidad de la segmentación sin pérdida sustancial de información.

Se puede optar por tres modalidades de reducción según sean las características de las variables predictoras:

1) Opción *sin restricciones:* Cada valor de la variable predictora puede ser agregado a cualquier otro valor de la misma variable. Sea, por ejemplo, la variable situación ocupacional con los valores "ocupado", "parado", "jubilado", "ama de casa", "estudiante" y "otros". De cara a la formación de grupos, la categoría "ocupado" podría formar grupo con "parados" y/o "estudiantes". La primera categoría es contigua, pero la segunda no lo es. Este procedimiento se aplica por regla general a variables de tipo nominal.

2) Opción *monótona:* Un valor de la variable sólo puede ser agregado a otro si es contiguo en la escala. En el procedimiento anterior, la categoría "ocupado" sólo podría unirse en un primer momento con la categoría "parado". Los "jubilados" podrían agregarse con los "parados" o las "amas de casa"; pero no con "ocupados", "estudiantes" u "otros". Es fácil deducir que este procedimiento sólo se puede aplicar con propiedad a variables ordinales[3]. Como este no es el caso de la variable del ejemplo anterior, para su tratamiento tendría

[3] Sin embargo, este método no debería aplicarse automáticamente a toda variable ordinal. Si no existe una pauta de relación lineal del predictor con la variable dependiente, las variables ordinales han de tratarse con el procedimiento sin restricciones. Por ejemplo, si se espera que mayores ingresos no conlleve opinión más favorable al aborto, sino que lo más probable es que los sujetos con altos y bajos ingresos sean los menos (o más) inclinados a aprobar esta práctica y los más (o menos) favorables sean los de ingresos medios, entonces el predictor ordinal no debería ser considerado monótono.

que usarse la opción sin restricciones, Un ejemplo de predictor monótono adecuado es el nivel de estudios. Si esta variable tuviera como valores "primarios", "secundarios" y "universitarios", el procedimiento permitiría la fusión de las categorías primera y segunda o segunda y tercera, y descartaría la posibilidad de formar un grupo compuesto por sujetos con estudios primarios y universitarios.

3) Opción mixta: Es idéntica a la opción monótona; pero permite un mayor grado de libertad, por cuanto un valor, generalmente el "No sabe, no contesta", puede agregarse libremente a cualquier grupo. Si la variable nivel de estudios tuviera el valor "Ns. Nc", con este procedimiento, los sujetos que no contestasen podrían agruparse con cualquiera de las tres categorías establecidas.

El funcionamiento de formación de grupos de categorías homogéneas se basa en el estadístico χ^2- Los pasos son los siguientes:

1) Se forman todos los pares posibles de categorías. Esto dependerá de la opción que se haya preferido dar a un determinado predictor. Así, en la variable situación ocupacional, que presentaba seis valores, el número posible de pares sería de 15 (combinaciones de 6 elementos tomados de dos en dos). Si se hubiese optado por la opción monótona, los pares posibles serían 5 (número de categorías menos una). Y si se escogiese la opción mixta, las posibilidades serían 9 (las 5 anteriores más 2 unidades menos que el número total de categorías). Véase Tabla 3.

Tabla 3

PARES POSIBLES DE LA VARIABLE SITUACIÓN OCUPACIONAL SEGÚN LAS TRES OPCIONES POSIBLES

Opción sin restricciones		Opción monótona		Opción mixta	
Ocupados	- Parados	Ocupados	- Parados	Ocupados	Parados
Ocupados	- Jubilados	Parados	- Jubilados	Ocupados	Otros
Ocupados	- Amas de casa	Jubilados	- Amas de casa	Parados	Jubilados
Ocupados	- Estudiantes	Amas de casa	- Estudiantes	Parados	Otros
Ocupados	- Otros	Estudiantes	- Otros	Jubilados	Amas de casa
Parados	- Jubilados			Jubilados	Otros
Parados	- Amas de casa			Amas de casa -	Estudiantes
Parados	- Estudiantes			Amas de casa -	Otros
Parados	- Otros			Estudiantes -	Otros
Jubilados	- Anas de casa				
Jubilados	- Estudiantes				
Jubilados	- Otros				
Amas de	- Estudiantes				
Amas de	- Otros				
Estudiantes	- Otros				

2) Para cada posible par se calcula el χ^2 correspondiente a su cruce con la variable dependiente. El par con más bajo χ^2. siempre que no sea significativo[4], formará una nueva categoría de dos valores fusionados. La condición de que no sea significativo es muy importante porque, caso de que lo fuese, indicaría que las dos categorías que se pretenden fusionar no lo pueden hacer, ya que son heterogéneas entre sí en los valores de la variable dependiente, y el objetivo es justo lo contrario, asimilar categorías con comportamiento semejante.

3) Si se ha fusionado un determinado par de categorías, se procede a realizar nuevas fusiones de los valores del predictor, pero esta vez con una categoría menos, pues dos de las antiguas han sido reducidas a una sola.

4) El proceso se acaba cuando ya no pueden realizarse más fusiones porque los χ^2 ofrecen resultados significativos.

De esta forma, como casos extremos, podría suceder que un predictor con c categorías siguiera con sus c grupos, en el supuesto de que todos ellos sean diferentes entre sí; o bien, que las categorías tengan valores tan parecidos en la variable dependiente que se queden reducidos a uno solo, con lo que el poder discriminador del predictor es nulo.

Véase un ejemplo práctico con la variable ideología como predictora de la posición ante el aborto (Tabla 4). En este caso la ideología tiene tres valores ("izquierda", "centro" y "derecha"). En el primer paso, se halla el χ^2 de los siguientes contrastes: "izquierda" *versus* "centro", "centro" *versus* "derecha" e "izquierda" *versus* "derecha", aunque éste último no se aplicaría si la ideología se considerara monótona. De estos tres χ^2, el menor (1.11) corresponde al contraste "centro" *versus* "derecha"; lo que indica que estas categorías son las más parecidas entre sí en lo que se refiere a opinión sobre el aborto. Por eso y, especialmente, porque ambas categorías no presentan diferencias significativas, se pueden reagrupar para explicar la actitud en cuestión. El siguiente paso, sería comprobar si el contraste ("centro", "derecha") *versus* "izquierda" presenta un χ^2 significativo; en cuyo caso estos dos serían los grupos de este predictor que posteriormente habría que contrastar con otros predictores. Si, por el contrario, no hubiese resultado significativo, habría sido consecuente juntar los dos grupos de valores y, en este caso, ya sólo hubiese quedado un único grupo de categorías, y, por tanto, la variable ideología no habría sido útil para discriminar la opinión estudiada.

[4] Todos los cruces tienen el mismo número de grados de libertad porque la variable dependiente es la misma para todos los contrastes y la variable independiente sólo tiene dos valores, pues recuérdese que se está trabajando con pares de categorías.

Tabla 4

PRIMER PASO:

CRUCE DE OPINION ANTE EL ABORTO SEGUN PARES DE VALORES DE IDEOLOGIA.

	Izquierda / Centro			Izquierda / Derecha			Centro / Derecha		
IDEOLOGIA-> / ABORTO	Izq.	Cent.		Izq.	Der.		Cent.	Der.	
Si	97 33.1	127 15.6	224 20.2	97 33.1	11 12.3	109 28.2	127 15.6	11 12.3	138 15.2
No	187 63.5	639 78.4	826 74.5	187 63.5	76 83.2	262 68.1	639 78.4	76 83.2	715 78.9
NS/NC	10 3.4	49 6.0	59 5.3	10 3.4	4 4.5	14 3.7	49 6.0	4 4.5	53 5.8
Total	294 26.5	815 73.5	1109 100.0	294 76.3	91 23.7	385 100.0	875 89.9	91 10.1	906 100.0

χ^2	G.L.	Significación	χ^2	G.L.	Significación	χ^2	G.L.	Significación
42.03	2	.0000	14.91	2	.0006	1.11	2	.5734

SEGUNDO PASO:

CRUCE DE OPINION ANTE EL ABORTO SEGUN VALORES AGRUPADOS EN EL PASO ANTERIOR

IDEOLOGIA-> / ABORTO		Izq.	Cent-Der	
Si	1	97 33.1	138 15.2	235 19.6
No	2	187 63.5	715 78.9	902 75.1
NS/NC	9	10 3.4	53 5.8	63 5.2
Total		294 24.5	906 75.5	1200 100.0

χ^2	G.L.	Significación
45.63	2	.0000

Existe un procedimiento que ahorra gran cantidad de cálculos y posee una razonable base lógica. Se trata de limitarse a la obtención de segmentaciones binarias. Esto implica que, sea cual sea el número de categorías de los predictores, se busque la mejor combinación de ellas que genere sólo dos grupos (k=2). En consecuencia, habría que formar todas las posibles combinaciones de dos grupos con las c categorías y seleccionar aquél con un χ^2 mayor. Es evidente que si se considera la opción monótona, el número de posibilidades de agrupación se reduce. En el caso de una variable ordinal con valores "mucho", "bastante", "poco" y "nada", el número de contrastes binarios sería de 7 con la opción sin restricciones y de 3 con la monótona. (Véase Tabla 5).

Tabla 5

CONTRASTES BINARIOS POSIBLES CON UNA VARIABLE DE CUATRO VALORES SEGUN OPCION

Opción monótona	Opción sin restricciones
(Mucho) vs. (Bastante, Poco, Nada)	(Mucho) vs. (Bastante, Poco, Nada)
(Mucho, Bastante) vs. (Poco, Nada)	(Bastante) vs. (Mucho, Poco, Nada)
(Mucho, Bastante, Nada) vs. (Nada)	(Poco) vs. (Mucho, Bastante, Nada)
	(Nada) vs. (Bastante, Mucho, Poco)
	(Mucho, Bastante) vs. (Poco, Nada)
	(Mucho, Poco) vs. (Bastante, Nada)
	(Mucho, Nada) vs. (Bastante, Poco)

2.2 *Selección de los mejores predictores.*

Una vez que para cada predictor se ha realizado la combinación de categorías más conveniente, el siguiente paso sería la selección de los mejores predictores. Para hacerlo, hay que calcular para cada predictor reducido su χ^2 correspondiente y comparar las significaciones obtenidas; sin embargo, es conveniente en este proceso modificar la significación de cada predictor con el ajuste de Bonferroni[5], porque la probabilidad de obtención de un resultado significativo aumenta artificialmente con la proliferación de pruebas estadísticas que implica este análisis.

[5]El ajuste de Bonferroni consiste en la aplicación de la desigualdad establecida por el mismo autor. Esta dice que en el caso de que se hagan B pruebas de significación, la significación total (P_T) debe ser menor o igual que la suma de cada una de las significaciones (P_i).

$$P_T \leq \sum_{i=1}^{B} P_i$$

El número posible de pruebas de significación se puede calcular a través de fórmulas combinatorias a partir del número de categorías iniciales del predictor (c) y del número de grupos formados tras la agrupación de categorías (k). Es obvio que el cálculo será distinto según la opción de reducción de categorías que se utilice.

Así, si se escoge la opción sin restricciones la fórmula es la siguiente:

$$B=\sum_{i=0}^{k-1} (-1)^i \frac{(k-i)^c}{i!(k-i)!}$$

Si se utiliza la monótona:

$$B = \binom{c-1}{k-1}$$

Y para la opción mixta:

$$B = \frac{k-1+k(c-k)}{c-1}\binom{c-1}{k-1}$$

En la práctica, hay que multiplicar la significación del χ^2 por el resultado de B, con lo que se evita el riesgo de rechazo inadecuado de hipótesis por realizar múltiples ensayos.

En la Tabla 27 hay un ejemplo donde se puede aplicar este proceso. En las anteriores segmentaciones no se ven las implicaciones de este ajuste porque sólo hay variables con dos categorías y, en estos casos, B es siempre igual a 1. La última variable de la citada tabla, ingresos del entrevistado, tiene un χ^2 de 62.9, al que con 6 grados de libertad debería corresponderle una significación de 1.2E-11. Sin embargo, en este caso, como B, el número de comparaciones posibles, es igual a 5, después de aplicar la fórmula para predictores mixtos con los parámetros c=4 y k=3, la significación real es menor o igual a 5.9E-11. Para más detalle véase Kass (1980) y Hawkins y Kass (1982).

Tabla 6

ANALISIS DE LA MUESTRA COMPLETA (Grupo 1).

PREDICTOR	FAVORABLE AL ABORTO			
Sexo	Sí	No	NS/NC	(100%)
Varón	23.33	71.71	4.96	576
Mujer	16.19	78.31	5.51	624
Total	19.62	75.14	5.25	1200

χ^2: 9.7 (g.l.=2 ; p ≤ 0.008)

PREDICTOR	FAVORABLE AL ABORTO			
Edad	Sí	No	NS/NC	(100%)
Menos de 46	26.01	68.94	5.04	660
46 o más	11.79	82.71	5.49	540
Total	19.62	75.14	5.25	1200

χ^2: 38.5 (g.l.=2 ; p ≤ 4.3e-9)

PREDICTOR	FAVORABLE AL ABORTO			
Ideología	Sí	No	NS/NC	(100%)
Izquierda	33.17	63.46	3.43	294
Centro y derecha	15.24	78.93	5.84	906
Total	19.62	75.14	5.25	1200

χ^2: 45.3 (g.l.=2 ; p ≤ 2.9e-10)

El ejemplo de predicción de la opinión ante el aborto en función de sexo, edad e ideología ayuda a entender este proceso. En la Tabla 6 se analiza el conjunto de la muestra por los distintos predictores, a los que ya se ha aplicado el proceso de agregación de categorías. Del conjunto de 1200 sujetos que han sido entrevistados, 576 son varones y 624 mujeres. El 23.3% de los primeros es favorable al aborto y el 16.2% de las mujeres sostiene la misma posición. El χ^2 (9.7) tiene una significación[6] ajustada de 0.008. Existe, pues, relación significativa; pero antes de proceder a seleccionar este predictor, es necesario analizar el resto de predictores incluidos en el análisis. La muestra está repartida entre 660 sujetos con menos de 46 años y 540 de mayor edad. También entre estos dos segmentos de la muestra hay diferencias en la opinión, incluso mayores. El 26.0% de los jóvenes autorizan esta práctica; pero sólo un 11.8% de los de más avanzada edad mantienen esta actitud. El χ^2 lógicamente presenta un valor mayor (38.5) y, de igual forma, su significación es muy baja (4.3E-9). Sin embargo, el mejor predictor es la ideología. Un tercio aproximadamente de los sujetos de izquierda aprueba el aborto en el supuesto de que un matrimonio no desee tener más hijos, mientras que la probabilidad de que los entrevistados de centro y los de derecha mantengan esta actitud es sólo de un 15.2%. El χ^2 es el mayor de los tres (45.3) y la significación, la más baja (2.9E-10). Por tanto, este es el mejor predictor de los tres y es el que se utilizará para realizar la primera segmentación de la muestra. De este modo,

[6] Se recuerda que por significación se entiende la probabilidad de cometer un error de tipo I (rechazo de una hipótesis verdadera). Por tanto, mientras más baja sea, las posibilidades de error son menores y la relación entre las variables es más fuerte.

quedarán formados dos grupos: uno de 294 sujetos (los autoubicados en la izquierda) y otro de 906 individuos que han declarado ser de centro o de derecha.

Tabla 7

ANALISIS DE LA MUESTRA DE SUJETOS DE IZQUIERDA (Grupo 2).

PREDICTOR	FAVORABLE AL ABORTO			
Sexo	Si	No	NS/NC	(100%)
Varón	37.50	60.74	1.77	180
Mujer	26.17	67.77	6.06	114
Total	33.11	63.46	3.43	294

χ^2: 6.8 (g.l.=2 ; $p \leq 0.03$)

PREDICTOR	FAVORABLE AL ABORTO			
Edad	Si	No	NS/NC	(100%)
Menos de 46	37.03	58.15	4.81	209
46 o más	23.40	76.60	0.00	85
Total	33.11	63.46	3.43	294

χ^2: 10.7 (g.l.=2 ; $p \leq 0.005$)

Una vez realizada la primera segmentación, se procede a la ejecución de sucesivas segmentaciones para cada uno de los grupos formados por la primera. Prosiguiendo con el ejemplo, habría que averiguar si entre los individuos de izquierda existen diferencias considerables de sexo o edad. Así, en la Tabla 7 se observa que los varones de izquierda son significativamente más favorables al aborto que las mujeres de esta ideología (37.5% vs. 26.2%). No obstante, sobre los entrevistados de izquierda, el predictor edad tiene mayor poder de discriminación. Los jóvenes apoyan este tipo de aborto en un 37.0% de casos, mientras que los mayores de 45 años sólo lo hacen en un 23.4% ($p \leq 0.005$).

También hay que realizar el proceso con los individuos de centro y derecha. Pero en este caso, además de probar el efecto de sexo y edad, hay que analizar si las personas de centro y derecha son diferentes entre sí. Esto no se aplicaba al otro grupo porque era un grupo homogéneo en ideología: estaba formado exclusivamente por sujetos de izquierdas. Tras el cálculo de los χ^2, (véase Tabla 8) la única variable discriminadora es la edad. Los sujetos de centro y los de derecha mantienen posiciones similares (el 15.6% de los de centro son favorables al aborto, y el 12.3% de los de derechas: $p \leq 0.60$). Como contrapartida, de los 906 entrevistados de centro y derecha, un 20.9% de los 451 jóvenes mantiene una actitud favorable al aborto; pero entre los 455 de mayor edad sólo un 9.6% tiene la misma opinión ($p \leq 1.2E\text{-}5$). Por tanto, al igual que ocurría entre los individuos de izquierda, el segundo paso de la segmentación realizado con los entrevistados de centro y derecha, divide a estos sujetos según su edad.

Tabla 8

ANALISIS DE LA MUESTRA DE SUJETOS DE CENTRO Y DERECHA (Grupo 3).

PREDICTOR	FAVORABLE AL ABORTO			
Sexo	Sí	No	NS/NC	(100%)
Varón	16.89	76.69	6.42	396
Mujer	13.95	80.66	5.38	510
Total	15.24	78.93	5.84	906

χ^2: 2.1 (g.l.=2 ; p ≤ 0.3)

PREDICTOR	FAVORABLE AL ABORTO			
Edad	Sí	No	NS/NC	(100%)
Menos de 46	20.90	73.95	5.15	451
46 o más	9.64	83.85	6.51	455
Total	15.24	78.93	5.84	906

χ^2: 22.6 (g.l.=2 ; p ≤ 1.2e-5)

PREDICTOR	FAVORABLE AL ABORTO			
Ideología	Sí	No	NS/NC	(100%)
Centro	15.57	78.45	5.98	815
Derecha	12.29	85.19	4.52	91
Total	15.24	78.93	5.84	906

χ^2: 1.1 (g.l.=2 ; p ≤ 0.6)

Hasta aquí se han sido realizado tres *segmentaciones* en dos *niveles* y en este proceso se han conformado cuatro *grupos*:

a) Sujetos de izquierda jóvenes: (n= 209; p_a= 37.0%).

b) Sujetos de izquierda mayores: (n= 85; p_a= 23.4%).

c) Sujetos de centro y derecha jóvenes: (n= 451; p_a= 20.9%).

d) Sujetos de centro y derecha mayores: (n= 455, p_a= 9.6%).

Aún se podría proseguir la segmentación en su tercer nivel para cada uno de estos cuatro grupos. Véase cada uno de ellos:

Tabla 9

ANALISIS DE LA MUESTRA DE SUJETOS JOVENES DE IZQUIERDA (Grupo 4).

PREDICTOR	FAVORABLE AL ABORTO			
Sexo	Sí	No	NS/NC	(100%)
Varón	42.38	55.01	2.61	122
Mujer	29.57	62.54	7.89	87
Total	37.03	58.15	4.81	209

χ^2: 5.6 (g.l.=2 ; p ≤ 0.06)

Dado que se han introducido sólo tres predictores, el grupo de jóvenes de izquierda únicamente puede ser subsegmentado con el predictor restante, el sexo. ¿Existen

diferencias en la posición ante el aborto entre los hombres y las mujeres de este segmento? Los 122 varones que forman este grupo son favorables en un 42.4%; las 87 mujeres sólo en un 29.6%. Estas diferencias parecen importantes; sin embargo, (Tabla 9) los tamaños de estas muestras no son suficientemente grandes para que esta desigualdad sea estadísticamente significativa. Por tanto, el análisis automático no subsegmentaría a este grupo de jóvenes de izquierda y de esta forma quedaría considerado como *grupo terminal*.

Tabla 10

ANALISIS DE LA MUESTRA DE SUJETOS MAYORES DE IZQUIERDA (Grupo 5).

PREDICTOR	FAVORABLE AL ABORTO			
Sexo	Si	No	NS/NC	(100%)
Varón	27.25	72.75	0.00	58
Mujer	14.94	85.06	0.00	26
Total	23.40	76.60	0.00	85

χ^2: 1.6 (g.l.=2 ; p ≤ 0.5)

La muestra de 85 individuos de izquierda mayores de 45 años es muy pequeña para que al subdividirla presente diferencias significativas entre los dos sexos. Efectivamente, en la Tabla 10, aunque los 58 varones mayores de izquierda son más favorables que las 26 mujeres de similares características, las diferencias no son estadísticamente significativas.

Tabla 11

ANALISIS DE LA MUESTRA DE SUJETOS JOVENES DE CENTRO Y DERECHA (Grupo 6).

PREDICTOR	FAVORABLE AL ABORTO			
Sexo	Si	No	NS/NC	(100%)
Varón	21.63	72.56	5.81	211
Mujer	20.25	75.18	4.57	239
Total	20.90	73.95	5.15	451

χ^2: 0.5 (g.l.=2 ; p ≤ 0.8)

PREDICTOR	FAVORABLE AL ABORTO			
Ideología	Si	No	NS/NC	(100%)
Centro	21.02	73.88	5.10	414
Derecha	19.56	74.75	5.69	37
Total	20.90	73.95	5.15	451

χ^2: 0.1 (g.l.=2 ; p ≤ 1.0)

En el grupo de los 451 jóvenes de centro y derecha, son posibles dos segmentaciones, bien con el predictor sexo, bien con la ideología, separando a los de centro de los de derecha. En la Tabla 11, también se detecta que ninguna de estas segmentaciones es significativa; sin embargo, en esta ocasión, no tanto por el bajo tamaño de las muestras, como por la pequeña diferencia de porcentajes (21.6% *vs.* 20.2% tomando en cuenta el sexo y 21.0% *vs.* 19.6% haciendo uso de la ideología).

Tabla 12

ANALISIS DE LA MUESTRA DE SUJETOS MAYORES DE CENTRO Y DERECHA (Grupo 7).

PREDICTOR	FAVORABLE AL ABORTO			
Sexo	Si	No	NS/NC	(100%)
Varón	11.47	81.42	7.11	185
Mujer	8.38	85.51	6.11	270
Total	9.64	83.85	6.51	455

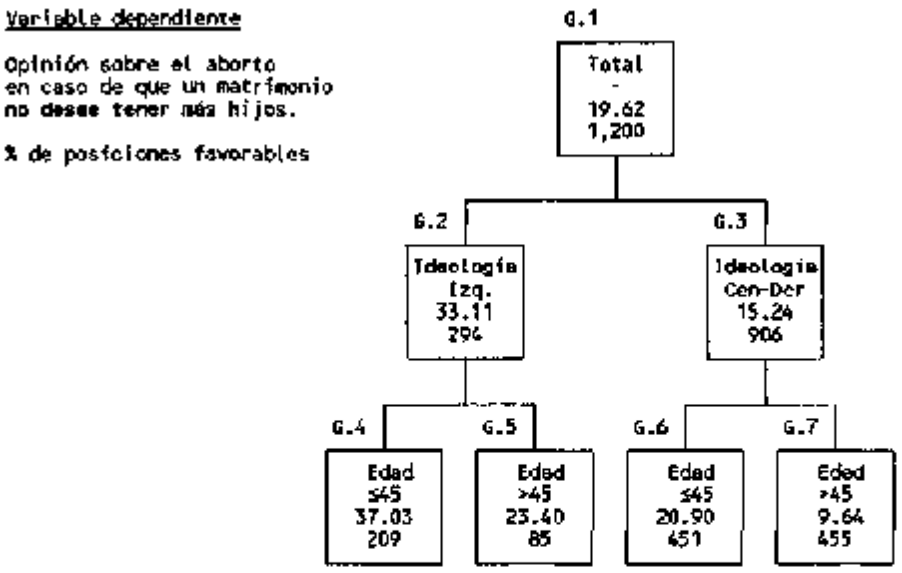
χ^2: 1.5 (g.l.=2 ; $p \leq 0.5$)

PREDICTOR	FAVORABLE AL ABORTO			
Ideología	Si	No	NS/NC	(100%)
Centro	9.96	83.15	6.89	401
Derecha	7.26	89.02	3.72	54
Total	9.64	83.85	6.51	455

χ^2: 1.3 (g.l.=2 ; $p \leq 0.5$)

Por último, el grupo de mayores de centro y derecha está compuesto por 455 sujetos, de los que sólo 9.6% aprueban el aborto en el supuesto de que un matrimonio no desee tener más hijos. El sexo no los discrimina, pues los varones favorables son el 11.5% y las mujeres 8,4% (diferencias no significativas); ni existe distinción significativa en la opinión entre los de centro y los de derecha (respectivamente 10.0% y 7.3%) (Tabla 12). En definitiva, tampoco este grupo es susceptible de posterior segmentación.

Tabla 13



En consecuencia, el análisis de segmentación subdivide a la muestra en los cuatro grupos descritos en la página 12 y representados en la Tabla 13. Destaca la diferencia en opinión entre los grupos terminales G.4 y G.7: por un lado, los jóvenes de ideología de izquierda, con un 37.0% de favorables al aborto en las circunstancias examinadas, y en el

lado opuesto, los mayores de centro-derecha con un 9.6% de la misma opinión. Entre estas dos posiciones los dos grupos restantes presentan porcentajes muy similares entre sí, posiblemente no significativos y, por tanto, no heterogéneos en los valores de la variable dependiente, aunque sí en los de las independientes o predictores. Estos grupos son el G.5, formado por mayores de ideología de izquierda, y el G.6, compuesto por jóvenes de ideólogía de centro-derecha. Por su desigual composición se justifica que, aun semejantes en su posición ante el aborto, se sigan considerando como segmentos distintos.

El proceso de segmentación debe ser examinado en sus distintas fases con el objeto de valorar el comportamiento de los predictores alternativos. El problema estriba en que el programa analiza varios predictores en cada paso de la segmentación y tiene que elegir entre ellos uno solo. Si en una determinada fase existen varios predictores de similar poder de segmentación, el análisis de la elección efectuada puede conducir a interpretaciones precipitadas. Para descubrir la posible existencia de este problema, habrá que prestar atención en cada segmentación a la significación ajustada del χ^2 de los predictores alternativos.

Tabla 14

SIGNIFICACIONES DEL χ^2 PARA CADA GRUPO Y PREDICTOR

Predictor	G.1	G.2(1)	G.3(1)	G.4(2)	G.5(2)	G.6(3)	G.7(3)
Sexo	0.008	0.03	0.3	0.06	0.5	0.8	0.5
Edad	4.3e-9	**0.005**	**1.2e-5**	.	.	.	.
Ideología	**2.9e-10**	.	0.6	.	.	1.0	0.5
n:	1200	290	910	205	85	446	464

N.B.: Los grupos en negrita son los grupos terminales. Entre paréntesis, el grupo del que procede. Los coeficientes en negrita indican la variable por la que se efectúa la segmentación en un determinado grupo.

Una posible presentación de la contribución de los predictores en las sucesivas fases de la segmentación sería la exposición en columnas de los distintos grupos, dispuestos por su orden de formación y representando en cada uno de ellos la significación mencionada de los distintos predictores, siempre y cuando éstos tengan después de la reducción más de una categoría con sujetos. Así en el ejemplo que antecede, se observan las siguientes peculiaridades (Tabla 14). La variable ideología, que fragmentó al grupo total, tiene como rival a la variable edad, cuya diferencia de significación es ínfima. El grupo segundo también podría haber sufrido la segmentación en lugar de la edad con el sexo, porque también este predictor es significativo. En el grupo tercero, sin embargo, la variable edad no tiene competidora en la segmentación.

Tabla 15

OPINION DEL ABORTO SEGUN LOS CUATRO GRUPOS TERMINALES FORMADOS POR LA SEGMENTACION

SEGMENTOS->		Jov-Izq.	May-Izq.	Jov-C.d.	May-C.d.	
ABORTO						
Si	1	77 37.0	20 23.4	94 20.9	44 9.6	235 19.6
No	2	122 58.2	65 76.6	333 74.0	382 83.8	902 75.1
NS/NC	9	10 4.8		23 5.2	30 6.5	63 5.2
	Total	209 17.4	85 7.1	451 37.6	455 37.9	1200 100.0

χ^2	G.L.	Significación	V de Cramer
75.13	6	.0000	0.17696

Por último, para determinar la capacidad predictora de la segmentación en su conjunto, resulta muy útil cruzar la variable dependiente con una nueva variable compuesta, cuyos valores sean las características de cada uno de los grupos terminales formados por la segmentación (Tabla 15). Un coeficiente de asociación[7], como puede ser la V de Cramer, resume el poder de predicción de los segmentos en su explicación de la variable dependiente. En este caso, el coeficiente, cuyo rango va de 0 a 1, no es tan alto como sería de desear, lo que indica la escasa capacidad de predicción que tienen la ideología y la edad para explicar la actitud de los individuos ante el aborto en el supuesto de que los padres no deseen tener más hijos.

2.3 *La finalización del proceso de segmentación.*

Si no se pusieran límites al proceso de segmentación, este análisis podría producir una gran cantidad de grupos terminales de tamaño muy pequeño que serían difíciles de interpretar. En un caso extremo, con un número elevado de variables y sin restricción alguna, este análisis produciría tantos grupos como individuos tuviese la muestra. En la situación común de una muestra de 1000 sujetos con 5 predictores de tres categorías cada uno, el número posible de grupos terminales sería de 243 (3^5) con un tamaño medio de cuatro personas (1000/243). Es conveniente, por tanto, poner límites al proceso de segmentación. Existen cuatro tipo de filtros que evitan la continuación de la segmentación: los de significación, los de asociación, los de tamaño y los de nivel.

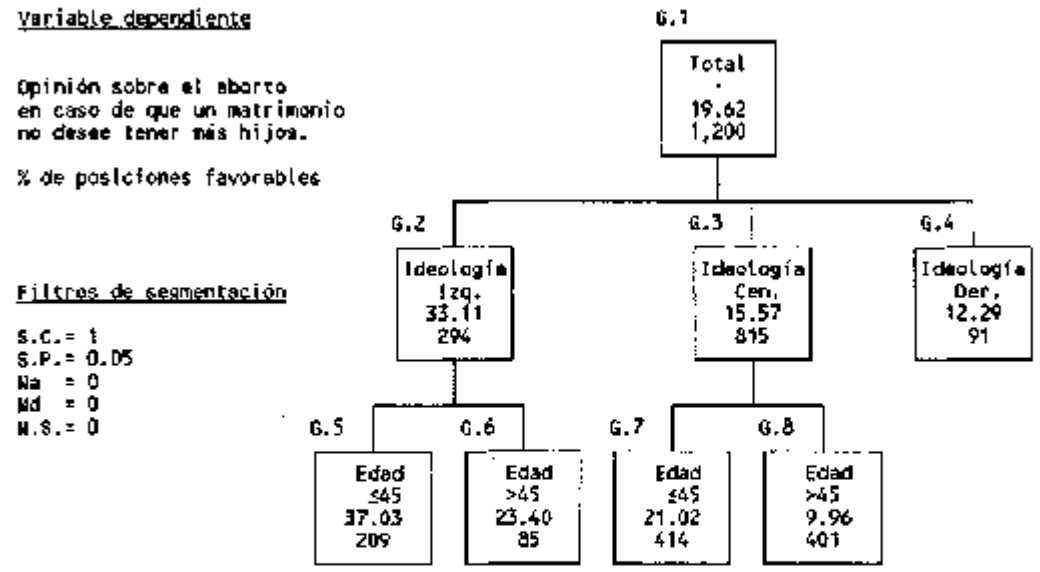
[7]Sobre coeficientes de asociación entre dos variables nominales, véase entre otros Ruiz-Maya *et al.* (1990), especialmente los capítulos 10 y 11.

2.3.1 *Filtros de significación.*

Son los más utilizados en la técnica CHAID de segmentación. Su criterio consiste básicamente en no permitir segmentaciones que no sean estadísticamente significativas. Por omisión, se sobrentiende que los límites de significación se sitúan en el nivel 0.05, que se corresponde con un nivel de confianza del 95%. Estos filtros pueden ser aplicados en dos de los procesos explicados anteriormente: bien en la agrupación de categorías de una variable, bien en la selección del mejor predictor.

La aplicación en el primer proceso es en realidad un mecanismo indirecto de finalización de la segmentación. Su efecto opera fundamentalmente en la cantidad de categorías de una determinada variable que van a segmentarse. Consiste en determinar la significación mínima para que dos categorías de una variable queden englobadas en el mismo segmento. El valor (S.C., significación de las categorías) más comúnmente asumido para este parámetro es el de 0.05. Si la significación de la diferencia en la variable dependiente entre dos categorías de la variable independiente es menor que este valor, se permite rechazar la hipótesis nula con un 95% de confianza y, como consecuencia, las dos susodichas categorías quedan separadas y se puede proseguir la segmentación. En cambio, si el valor es superior a 0.05, las categorías se funden, y, si quedan agrupadas todas las categorías de todas las variables, la segmentación se detiene.

Tabla 16



Los valores extremos permiten comprender con mayor eficacia el efecto de este mecanismo. Si se escoge el mayor valor posible del parámetro (1.0), entonces, la agrupación o reducción de categorías de las variables se torna imposible y, siempre que haya

significación entre predictor y variable dependiente, la segmentación formará con una determinada variable tantos grupos como categorías tenga. Se puede extraer un buen ejemplo de este procedimiento a partir de la operación mostrada en la Tabla 4. En aquel caso las categorías centro y derecha quedaron unidas porque la significación de sus diferencias era de 0.57 (superior a 0.05). Si se hubiese establecido el criterio con un parámetro superior a dicha cifra, la segmentación hubiese sido más "frondosa", siguiendo la metáfora de la representación en forma arbórea. En concreto, la primera subdivisión de la muestra, en lugar de dar lugar a dos grupos, proporciona tres grupos. (Comparar Tabla 13 y Tabla 16).

Tabla 17

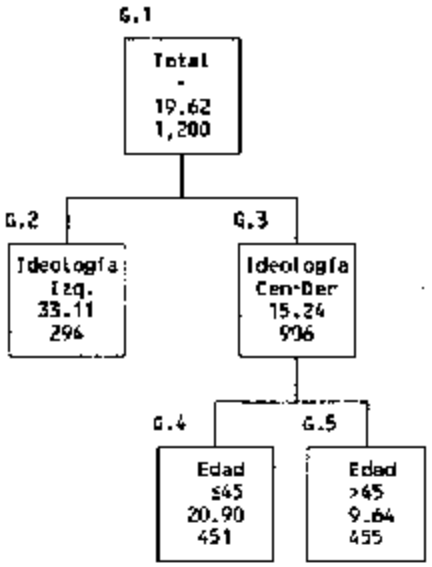
Variable dependiente

Opinión sobre el aborto en caso de que un matrimonio no desee tener más hijos.

% de posiciones favorables

Filtros de significación

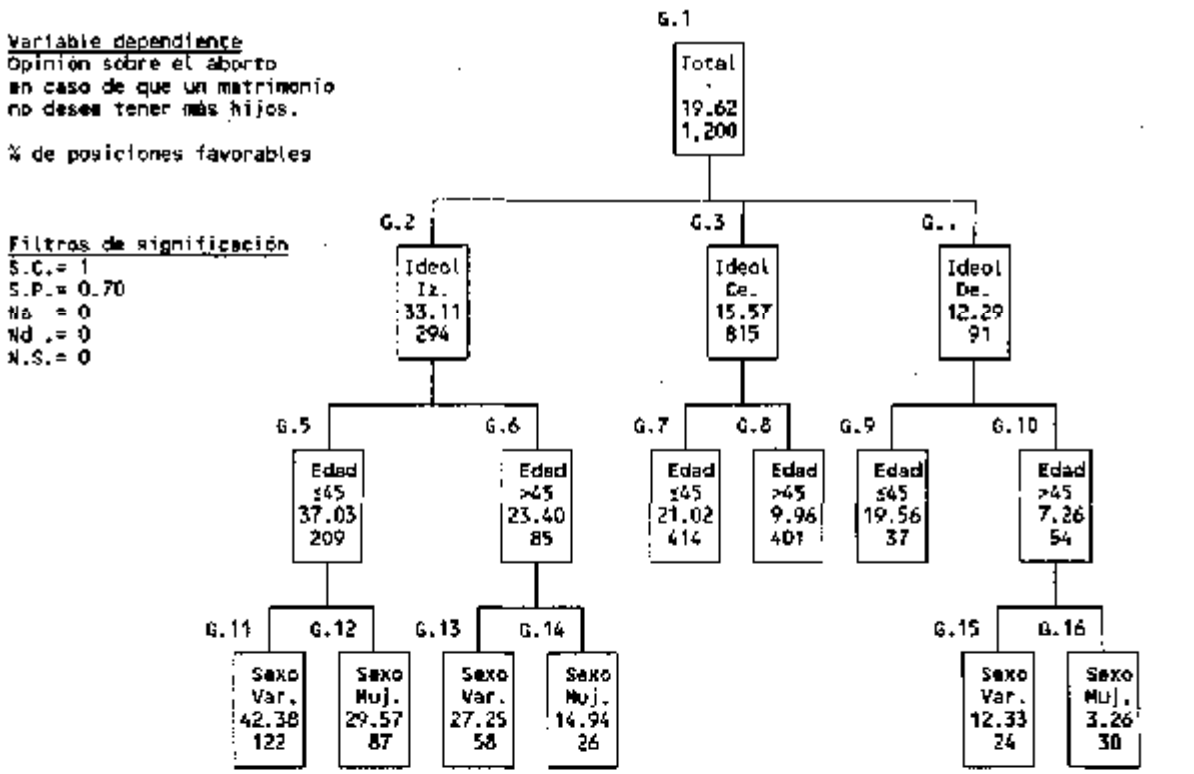
S.C.= 0,0005
S.P.= 0,05
Na = 0
Nd = 0
N.S.= 0



Si, en vez de poner el nivel de significación de la agrupación de las categorías en un valor alto, se situara en un valor bajo (por ejemplo, 5E-4), entonces, en lugar de producirse más subdivisiones entre los grupos, se generarían menos divisiones entre las categorías, con el riesgo añadido de que una determinada variable no funcione como un buen predictor. Esto es lo que sucede en el ejemplo de la Tabla 17, que se diferencia de la Tabla 13 en que no se produce segmentación por edad entre los individuos de izquierda. Y ocurre de esta manera porque la diferencia de porcentajes de las categorías de jóvenes y mayores no proporciona una significación menor de 0.0005. No siempre sucede esto de forma que implique la detención de la segmentación de un grupo. Lo lógico es esperar que una subdivisión de c categorías se reduzca a un número k, inferior al producido por un nivel de significación superior. En este caso, como el número inicial de categorías es igual a 2, la reducción implica la obtención de una sola categoría y de esta forma la segmentación no se lleva a cabo.

El otro mecanismo de control de significación, en lugar de operar sobre la agrupación de categorías, afecta a la selección de predictores. Este procedimiento es una forma directa de finalizar la segmentación, porque, después de encontrar el predictor con menor significación, si no es inferior al límite establecido (generalmente 0.05), es obvio que no habrá otro predictor que cumpla también con esta propiedad, por lo que el proceso de segmentación se detiene. Visto desde sus posibilidades extremas, si se establece este parámetro (S.P., significación del predictor) en el valor 1, la segmentación se producirá por todas las variables existentes; pero si se determina que el parámetro sea 0, entonces la segmentación no se produce ni tan siquiera en el primer nivel, pues la significación empírica de un predictor, por muy pequeña que sea siempre es superior a 0.

Tabla 18



Si se aplica al ejemplo de la Tabla 13 un filtro de significación de predictor superior al establecido por omisión (por ejemplo, 0.70), es de esperar que la segmentación proporcione mayor número de niveles. En aquella tabla, no aparecía el predictor sexo porque sus diferencias eran muy pequeñas. Ahora bien, es preciso tener en cuenta que no basta cambiar el parámetro S.P., porque si sigue efectivo un valor inferior del S.C., al operar con anterioridad, éste elimina los efectos del primero. Es conveniente, por tanto, que S.C. > S.P. Por eso, en el ejemplo de la Tabla 18, aparecen los valores S.C. = 1 y S.P. = 0.70.

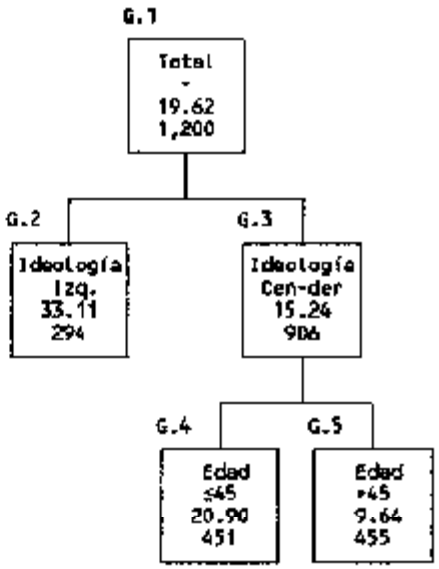
Como es de esperar con estos parámetros, la segmentación desciende al tercer nivel y aparece el sexo como un tercer predictor. De todas formas, la diferencia de porcentajes de hombres y mujeres que están a favor del aborto es pequeña en relación con el tamaño de estos segmentos, y, en el caso de los G.7, G.8 y G.9, ni tan siquiera con el nivel establecido en 0.7 se produce la segmentación.

Tabla 19

Variable dependiente
Opinión sobre el aborto en caso de que un matrimonio no desee tener más hijos.

% de posiciones favorables

Filtros de significación
S.C.= 0.05
S.P.= 0.0001
Na = 0
Nd .= 0
N.S.= 0



En cambio, si se aplica un filtro más severo, la segmentación sólo tendrá lugar cuando el predictor tenga una capacidad de diferenciación alta. Sobre el ejemplo matriz de la Tabla 13, aplicando en lugar del 0.05 por omisión, un S.P. de 0.0001, se obtiene una segmentación más reducida (Tabla 19) en la que los individuos de izquierda no aparecen segmentados, porque el predictor edad tiene una significación por encima del nivel establecido y por debajo del valor filtro por omisión.

2.3.2 *Filtros de asociación.*

Cumplen una función análoga a la de los filtros de significación de predictores. Se pueden aplicar a los siguientes coeficientes de asociación: Phi, V de Cramer, Coeficiente de Contingencia, T de Tschruprow y otros[8]. Se trata de detener la segmentación no porque un determinado cruce no obtenga un mínimo de significación, sino porque el coeficiente de asociación elegido no alcance un determinado nivel. Lo que principalmente diferencia a un procedimiento de otro es el hecho de que el que opera sobre la asociación no es sensible al número de casos sobre los que se trabaja. Por tanto, en valores equiparables de uno y otro,

[8]Véase nota 7.

los filtros de asociación son más permisivos en los niveles más bajos de segmentación, Como los de significación son muy sensibles al número de casos, es muy probable que en el tercer o cuarto nivel el análisis no cumpla las condiciones del filtro, porque los segmentos tengan un tamaño reducido. En cambio, los coeficientes de asociación, por el hecho de eliminar la influencia del número de casos, permiten segmentaciones aun en condiciones de escasos sujetos. En este caso hay mucho menos acuerdo sobre cuál debe ser el valor del filtro. En general, se consideran adecuados los valores 0.10 ó 0.20. Sin embargo, el programa CHATD del SPSS no contempla la posibilidad de utilizarlos para el control de la segmentación[9]. En todo caso, la opción recomendada para el uso de estos filtros es que se utilicen en conjunción con un filtro de significación, de forma que una segmentación que no sea significativa no se lleve a cabo por muy grande que sea su coeficiente de asociación. El caso contrario, que justifica especialmente el uso de estos filtros, también suele suceder. Se trata de relaciones entre variable dependiente y predictor muy significativas, pero con un coeficiente de asociación bajo, que se dan con frecuencia cuando se trabaja con muestras de elevado número de casos.

2.3.3. *Filtros de tamaño*

Su principal objetivo consiste en evitar que se formen grupos muy pequeños durante el proceso de segmentación, dado el problema que supone la generalización en estos casos. Si, por ejemplo se segmentara un grupo de 25 personas de las que un 30% es favorable al aborto, esto plantearía dos problemas: por un lado, este grupo no sería representativo en sí de la población; por otro, el valor del 30% tampoco sería un estimador muy preciso con un tamaño de muestra tan reducido.

Los filtros de tamaño pueden aplicarse en dos momentos: antes de la segmentación y después de la segmentación. En el primer caso, no se puede formar un grupo si no tiene un número establecido de componentes. En el segundo, la segmentación se detiene en el supuesto de que haya un grupo que haya descendido de un determinado número de individuos.

Supóngase que se arbitra que no haya ningún grupo con menos de 400 sujetos, en cuyo caso, si se aplica la segmentación a los datos de la Tabla 13, la ideología no sería un predictor adecuado porque genera un grupo, los individuos de izquierda, con menos de dicha cantidad establecida (294). Por tanto, en estas circunstancias, la segmentación (Tabla 20)

---

[9] No obstante, Sonquist y Morgan (1963), por utilizar la segmentación binaria con una variable dependiente de intervalo, confiaban más en los coeficientes de asociación que en los estadísticos de significación. Por ello, su AID seleccionaba las variables con mayor coeficiente de asociación y establecía como principal filtro la magnitud del coeficiente de determinación η^2, es decir, el cociente entre la suma cuadrática intergrupos y la suma cuadrática total. Otro filtro considerado por estos autores consiste en que la segunda cantidad mencionada alcance un mínimo nivel arbitrario. La razón estriba en evitar la segmentación de grupos muy homogéneos. Este último criterio sería inaplicable en el algoritmo CHAID.

presentaría un aspecto muy diferente de la original. Se formarían sólo dos grupos de edad, compuestos uno por 660 jóvenes y el otro por 540 mayores.

Tabla 20

Variable dependiente
Opinión sobre el aborto
en caso de que un matrimonio
no desee tener más hijos.

% de posiciones favorables

Filtros de significación
S.C.= 0.05
S.P.= 0.05
Na = 400
Nd .= 0
N.S.= 0

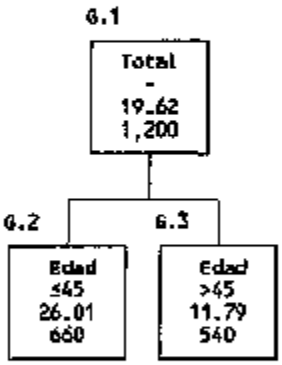
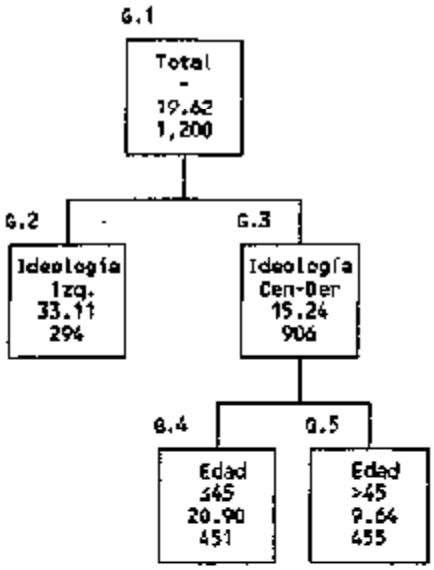


Tabla 21

Variable dependiente
Opinión sobre el aborto
en caso de que un matrimonio
no desee tener más hijos.

% de posiciones favorables

Filtros de significación
S.C.= 0.05
S.P.= 0.05
Na = 0
Nd .= 400
N.S.= 0



En cambio, si se opta por el filtro del tamaño después de la segmentación y se toma como cantidad el mismo número arbitrario, esto es, 400, el gráfico en forma de árbol toma una apariencia completamente distinta del anterior, porque con este nuevo criterio, la ideología sí funciona como predictor (Tabla 21). Lo que sucede es que el grupo de ideología de izquierdas no se segmenta porque su tamaño es inferior al establecido. Sin embargo, el grupo de centro-derecha, por tener 906 sujetos, se segmenta normalmente.

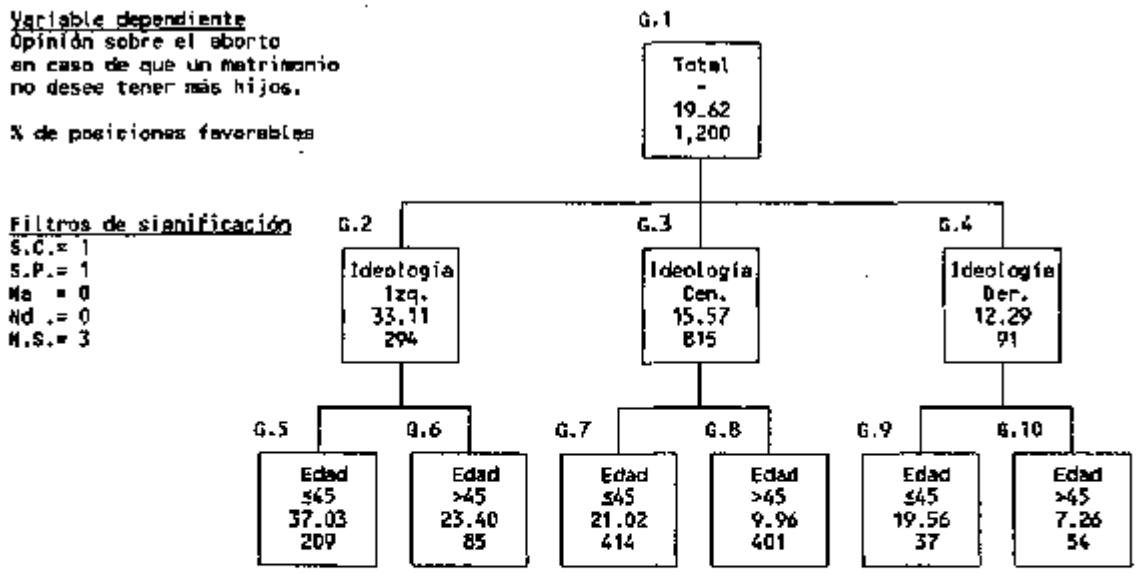
Es obvio que ambos filtros pueden utilizarse al mismo tiempo. Lo que no tiene sentido es que el filtro antes de la segmentación (Na) sea superior en número al de después

(Nd), puesto que de esta forma este último no se aplicaría. Más sentido tiene que Na sea inferior a Nd. Como regla general, se recomiendan unos parámetros de 100 para Na y 300 para Nd. Esto implica la no obtención de grupos inferiores a un centenar de personas y la no segmentación de conjuntos con menos de 300 componentes.

2.3.4 *Filtros de nivel.*

Por último, existe un cuarto tipo de mecanismo de detención de la segmentación. Consiste en arbitrar un nivel máximo de segmentación. Si se establece este criterio en 1, la segmentación no tendrá lugar; si en 2, sólo se realizará una segmentación; si en 3, dos. Por tanto, por nivel se entiende cada una de las franjas horizontales del dendograma en forma de árbol. La primera franja horizontal corresponde al total de la muestra, la segunda a la primera segmentación, la tercera a la segunda. Este filtro evita que se formen múltiples segmentaciones en segmentos desproporcionadamente grandes de la muestra. Asimismo, contribuye a simplificar los resultados en la medida en que reduce directamente el número de variables necesarias para predecir la variable dependiente.

Tabla 22



En el ejemplo de la Tabla 22, se han fijado los filtros de significación en 1, con objeto de que sólo operase el filtro de nivel. Por ello, a diferencia del de la Tabla 13, aparece la ideología escindida en tres segmentos. Pero, de forma distinta al de la Tabla 18, no prosigue la segmentación hasta el cuarto nivel, puesto que el valor del filtro Ns (nivel de segmentación) es 3.

3. Otros análisis de segmentación con variable dependiente nominal

El procedimiento explicado en el epígrafe 2 no es el único dentro del conjunto de técnicas de segmentación. Existen otros similares en los siguientes aspectos: a) se distingue entre variable(s) dependiente(s) y predictores y éstos son siempre de naturaleza nominal u ordinal, y b) se procede de forma jerárquica y paso a paso, esto es, la muestra total es segmentada inicialmente por el predictor que mejor cumple un determinado criterio y a partir de estos resultados se vuelve a aplicar el mismo procedimiento en cada uno de los segmentos formados.

Las diferencias entre estos procedimientos estriban fundamentalmente en las medidas que utilizan para agrupar categorías o segmentar grupos. La distinción primaria entre estos métodos consiste en el tipo de variable dependiente utilizada. El procedimiento CHAID es especialmente indicado para variables nominales. La segmentación también puede hacerse con una variable dependiente ordinal, con una de intervalo o con varias variables de intervalo. Pero en este trabajo sólo se contemplarán las medidas aplicables a variables nominales[10].

Como se acaba de decir, el χ^2, con su correspondiente ajuste de Bonferroni, no es el único medio para elegir los mejores predictores de una variable dependiente nominal. Una de las medidas clásicas, antecedente e inspiradora del análisis de segmentación, fue diseñada por Belson (1959) para el caso de variables dependientes dicotómicas. Se denomina *medida del desplazamiento máximo* y consiste en la semisuma de las desviaciones positivas entre los valores observados y los esperados en el supuesto de independencia entre la variable dependiente y el predictor. Indica el número de sujetos que tendrían que cambiar de valor en las variables para que la independencia entre ellas fuese completa y, evidentemente, a valores más altos, corresponde mayor capacidad predictiva de la variable independiente sobre la dependiente.

Sea una variable dependiente dicotómica, como puede ser la posesión de vídeo, y una variable independiente, como la clase social subjetiva, con tres valores: “alta”, “media” y “baja”. En la Tabla 23 se observan los tres pasos que han de llevarse a cabo para la segmentación, utilizando la distancia de Belson. En primer lugar, se dispone de la tabla de frecuencias observadas. La variable dependiente, la posesión de vídeo en este ejemplo, ha de ser dicotómica; el predictor puede tener inicialmente cualquier número de categorías. En segundo lugar, se calculan los valores de las frecuencias esperadas en el supuesto de independencia entre las variables. Por último, se calculan las diferencias entre los valores observados y los esperados. De esta forma, se obtienen

[10] Una exposición breve, pero completa, se encuentra en Fielding (1977).

Tabla 23

Cruce (supuesto) de posesión de video por clase social (frecuencias observadas)

Posesión de video	Clase social			
	Alta	Media	Baja	Total
Sí	24	126	6	156
No	6	84	54	144
Total	30	210	60	300

Valores esperados en el supuesto de independencia entre las variables

Posesión de vídeo	Clase social			
	Alta	Media	Baja	Total
Sí	15.6	109.2	31.2	156
No	14.4	100.8	28.8	144
Total	30	210	60	300

Diferencias entre los valores observados-esperados y desplazamiento máximo

Posesión de vídeo	Clase social			
	Alta	Media	Baja	Total
Sí	8.4	16.8	-25.2	25.2
No	-8.4	-16.8	25.2	25.2
Desplazamiento máximo				25.2

en cada columna cifras iguales en número absoluto, pero de signo contrario para cada uno de los valores de la variable dependiente (8.4 y -8.4 en la clase social alta). Prescindiendo de los resultados opuestos de la segunda categoría de la variable dependiente, la suma de los valores positivos da lugar al desplazamiento máximo. A partir de estos resultados, también pueden segmentarse dos categorías con el predictor. La primera con aquéllas cuyo valor residual sea positivo y la segunda con las que posean valor residual negativo. De esta forma se obtiene una agrupación óptima en dos grupos a partir de las k categorías de la variable independiente. En este caso, las clases sociales alta y media por un lado, frente a la clase social baja, por el otro. Si se calcula el nuevo desplazamiento máximo como fruto de la recodificación de valores realizada en el paso anterior, el resultado no varía, seguiría siendo 25.2.

Para el análisis de segmentación con variable dependiente nominal, existe otro algoritmo, llamado sucintamente THAID en virtud del nombre del programa informático que lo desarrolla (Messenger y Mandell, 1972), que utiliza una distancia basada en la métrica de *bloques urbanos* como criterio para maximizar la diferencia entre segmentos. Consiste esta distancia en el sumatorio de los valores absolutos de las diferencias entre los valores

observados de las frecuencias y sus correspondientes valores esperados en el supuesto de independencia entre el predictor y el criterio dependiente. Su fórmula general es:

$$\delta=\sum_{i}^{f}\sum_{j}^{o}|f_{ij}-f_{ij}^{*}| \qquad (1.1)$$

La principal diferencia entre δ y χ^2 es que la primera no está afectada por posibles desviaciones extremas de las frecuencias empíricas[11]. Existen otras diferencias. En primer lugar, en la distancia δ no se controla el tamaño inicial de las casillas. No es lo mismo una diferencia de cinco individuos en una casilla de diez personas que en otra de mil. Este inconveniente es débil porque todas la tablas a comparar poseen el mismo número de casos. Otra diferencia es el hecho de que los valores de δ dependan del número de casillas de la tabla. Esto se solventa en lo que concierne a la filas (variable dependiente) por el hecho de que siempre es constante y, en lo que respecta a las columnas, realizando sólo segmentaciones binarias de forma que el número de columnas (2) permanece siempre constante. Así con esta reducción, la fórmula (1.1) se convierte en:

$$\delta=\sum_{i}^{f}|f_{i1}-f_{i1}^{*}|+\sum_{i}^{f}|f_{i2}-f_{i2}^{*}| \qquad (1.2)$$

La última diferencia y principal inconveniente para δ es su carencia de un test de significación, en contraposición con el estadístico χ^2. Teniendo en cuenta que los primeros análisis de significación no se basaban directamente en la significación, sino en la asociación, su uso no rompe la tradición de los seguidores de esta técnica. Sin embargo, siguiendo estos criterios incontrastables, estas técnicas pierden adeptos entre los estadísticos más rigurosos.

Finalmente hay que mencionar la aplicación estricta a las variables nominales del algoritmo clásico de segmentación de Morgan y Sonquist, diseñado en principio para variables cuantitativas. Para ello, se toma como dependiente una variable cualitativa con sólo dos valores (1 = presencia de una característica; 0 = ausencia de la misma característica). En estos casos se sabe: a) que la frecuencia (f_{ij}) de sujetos con una determinada característica en un grupo j a la que se atribuye discrecionalmente el valor 1 es igual al sumatorio de todos los valores de la variable (en un grupo muestral de cinco individuos, si hay tres casados -valor 1- y dos solteros -valor 0-, la suma de los valores es igual a los tres casados); b) que la proporción de individuos con una determinada característica (p_{ij}) en un grupo j es igual al sumatorio de los valores dividido por el número de sujetos del grupo, lo que equivale por definición a la media (en el anterior grupo, la media sería el sumatorio de los valores -3-,

[11] Entre χ^2 y δ existe una oposición similar a la que existe entre la varianza y la desviación media. Las primeras, al elevar al cuadrado las diferencias, tienden a un valor muy alto en el caso de valores muy dispares, mientras que las segundas respetan el tamaño inicial de las diferencias

dividido por el número de casos -5-; este cociente -0,6- es también la proporción de casados en el grupo muestral), y c) estas igualdades pueden aplicarse a la muestra total, de forma que la proporción de sujetos con una determinada característica es idéntica a la media calculada sobre la base de que se otorgue a los individuos con ese rasgo un 1 arbitrario y al resto un 0 de la misma naturaleza.

$$t_{.j}=\sum_{k}^{n_j} x_{jk}=n_j\bar{x}_j$$

$$p_{1j}=\frac{\sum_{k}^{n_j} x_{jk}}{n_j}=\bar{x}_j \qquad (1.3)$$

y

$$p_1=\frac{\sum_{j}^{c}\sum_{k}^{n_j} x_{jk}}{n}=\bar{\bar{x}}$$

Conocida es además la fórmula de la suma cuadrática interna del análisis de varianza:

$$SCI=\sum_{j}^{g} n_j(\bar{x}_j-\bar{\bar{x}})^2 \qquad (1.4)$$

siendo g el número de grupos, n_j el tamaño de cada grupo, $\bar{x}_j$ sus respectivas medias y $\bar{\bar{x}}$ la media del conjunto de la muestra. Si en lugar de trabajar con un número indeterminado de grupos, se opera sólo con dos, como es el caso de las segmentaciones binarias, la fórmula quedaría de la siguiente forma:

$$SCI=n_1(\bar{x}_1-\bar{\bar{x}})+n_2(\bar{x}_2-\bar{\bar{x}}) \qquad (1.5)$$

y dadas las igualdades presentadas en (1.3), se obtiene:

$$SCI=n_1(p_{11}-p_1)^2+n_2(p_{12}-p_1)^2$$

o

$$SCI=(f_{11}^o-f_{11}^e)^2+(f_{12}^o-f_{12}^e)^2 \qquad (1.6)$$

De esta forma, esta distancia es intermedia entre la δ y el χ^2, con la característica peculiar de que sólo trabaja con uno de los valores de la variable dependiente.

4 La aplicación del análisis de segmentación.

Los creadores de los primeros análisis de segmentación llamaron a esta técnica AID. Como su propio nombre indica, la principal intención era la búsqueda de interacciones entre un conjunto de variables independientes con una variable dependiente, para evitar los problemas del análisis de clasificación múltiple, que es un análisis de varianza con más de un criterio de clasificación (Sonquist y Morgan, 1964). Sin embargo, esta técnica ha sido muy utilizada con propósitos distintos de los originalmente pensados[12]; de ahí que sea conocida actualmente bajo el nombre de análisis de segmentación, que da una idea más precisa de su utilización en el campo de la investigación de mercados[13]. "El concepto «segmentación de mercado» es anterior al de «segmentación» entendido como herramienta de análisis estadístico al servicio de la investigación comercial. La segmentación de mercados surge como «filosofía» o estrategia de *marketing* aplicable a un mercado en el que, superada la fase evolutiva de un mercado de desarrollo de la demanda primaria de un producto, la oferta de marcas se traduce en una lucha abierta de presentación de ventajas diferenciales para la obtención de mayores participaciones de mercado" (Sánchez Cuenca, 1990, 512).

Para analizar las distintas posibilidades de aplicación de esta técnica en el análisis de cuestionarios, sus usos se clasificarán según los objetivos de la investigación. Se estudiarán, en primer lugar, las posibilidades descriptivas, se continuará con la utilidad exploratoria de esta técnica y, por último, se hará mención a la capacidad explicativa de este procedimiento.

4.1 *Utilidad descriptiva del análisis de segmentación.*

La función clasificadora del análisis de segmentación permite configurar una serie de grupos que se distinguen por su comportamiento distinto en una determinada variable dependiente. La especificación de las características de los grupos terminales formados por esta técnica es un excelente medio para describir grupos heterogéneos de la muestra. Segmentar significa dividir y este análisis permite con su algoritmo el hallazgo de grupos muy distintos en un determinado aspecto. Por tanto, uno de los usos que se le puede dar a la segmentación es la descripción de las muestras y, por extensión, de las poblaciones de las que son extraídas.

La mejor manera de efectuar la descripción con el análisis de segmentación es mediante la interpretación de los grupos terminales. Hay que recordar que para hacer

---

[12] Ejemplos de aplicación en el campo de las Ciencias Sociales en España son Horter (1978), García Ferrando (1982) y Alvira, García López y Horter (1982). Todos ellos utilizan el algoritmo AID basado en la suma cuadrática.

[13] El artículo donde se presenta por primera vez esta estrategia de mercados fue publicado en el *Journal of Marketing* (W. Smith, 1956).

una buena descripción es necesario introducir predictores adecuados en el procedimiento. A continuación se compararán dos ejemplos, ambos con la misma variable dependiente, opinión sobre el aborto, pero el segundo con más predictores. Así se mostrará la conveniencia de dos reglas: a) incluir predictores que sean relevantes para nuestra variable dependiente y b) introducir el máximo posible de predictores ya que el análisis en cuestión se encarga de filtrar los relevantes.

El primer ejemplo ya se ha presentado en el epígrafe anterior. Se trata de segmentar la opinión sobre el aborto en el caso de que un matrimonio no desee tener más hijos a partir de tres predictores: sexo, edad e ideología. De acuerdo con la Tabla 13 sólo las dos últimas variables discriminan la opinión de los sujetos y, de este modo, se formaron cuatro grupos terminales. Los más proabortistas (37.0%) son las personas de izquierda con menos de 45 años. Le siguen los de izquierda con más edad (23.4%). Los de ideología de centro-derecha, si son jóvenes presentan una actitud favorable en un 20.9%; pero si son mayores de 45 años, la posición a favor se reduce al 9.6%. Merece la pena destacar que los de izquierda siempre son más favorables al aborto que los de centro-derecha; aunque en el grupo de los mayores de cuarenta y cinco años de izquierda, la probabilidad de estar de acuerdo es poco más favorable que en el de los jóvenes de centro-derecha.

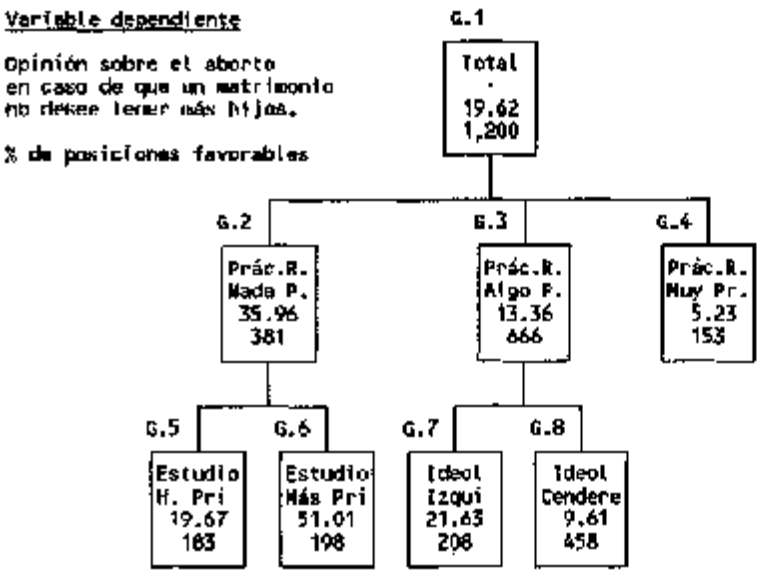
Además de los porcentajes, para la correcta descripción, hay que tener en cuenta la frecuencia de cada grupo. Se observa en la misma Tabla 13 que el grupo de mayores de izquierda es muy reducido: apenas cubre el 7% de la muestra (85 de los 1200). En cambio, los dos grupos de centro-derecha son los más numerosos, casi un 40% en cada uno de ellos[14].

Hay dos posibles maneras de resumir la información descriptiva de este análisis. La primera sería insistiendo en la oposición centro-derecha *versus* izquierda, destacando que las tres cuartas partes de la muestra (G.6 y G.7) presentan una baja aceptación del aborto practicado en las circunstancias antedichas, y los situados a la izquierda (24.5%, 294 de los 1200 entrevistados) son más favorables, sin llegarlo a ser en mayoría (en el supuesto más favorable, los jóvenes de izquierda sólo aprueban esta práctica en un 37.0%). La otra interpretación insistiría más en los grupos extremos: Aunque aproximadamente un 20% de la población apruebe que se lleve a cabo este tipo de aborto, hay dos grupos en los que las probabilidades de aprobación son considerablemente diferentes. Por un lado, están las personas mayores de centro-derecha (un 37.9% de la muestra), de los que menos de un 10%

[14] Existe en este caso gran mayoría de individuos de centro derecha porque los que no han proferido su ideología están también incluidos en este grupo. Se ha hecho así porque su posición ante el aborto es más similar a la de aquéllos que a la de los de izquierda. Aunque esto se haya realizado manualmente para simplificar el ejemplo, el análisis de segmentación los hubiese incorporado automáticamente de la misma forma.

darían su aprobación. Y, por otro, se encuentran los jóvenes de izquierda (un 17.4% de la muestra) entre los que la probabilidad de mantener esta opinión es superior al 35%.

Tabla 24



En el segundo ejemplo, además del sexo, la edad y la ideología, se introducen los siguientes predictores: estado civil, posición familiar, número de individuos en el hogar, nivel de estudios del entrevistado y del cabeza de familia, religión, práctica religiosa, clase social subjetiva e ingresos. En total, pues, se incluyen 12 variables para tratar de describir la misma variable dependiente: la opinión sobre el aborto en el supuesto de que un matrimonio no desee tener más descendencia. Los resultados del dendograma se presentan en la Tabla 24.

Lo que más resalta en este gráfico es el poder explicativo de la primera variable de segmentación: la práctica religiosa. Esta forma tres grupos: los nada practicantes, con un 36.0% de favorables a este tipo de aborto, los practicantes medios, con un 13.4% de favorables y los muy practicantes, con sólo un 5.2% de opiniones a favor. Sin embargo, en el dendograma aparecen cinco grupos terminales, pues cada uno de los dos primeros anteriores se subdivide en otros dos. De este modo, entre los nada practicantes se forman dos grupos muy distintos: por un lado, los individuos de bajos estudios con un 19.7% de favorables; por el otro, los de más estudios, de los que aproximadamente la mitad sostiene la posición a favor. Entre los practicantes medios, también se forman dos grupos distintos, pero no en función del nivel de estudios, sino de la ideología. Así, entre quienes practican algo o medianamente la religión, pero son de izquierdas, la aceptación del aborto en el caso

de que un matrimonio no desee tener más hijos es del 21.6%. Y, si son de centro-derecha, el porcentaje de favorables desciende al 9.6%.

Como consecuencia de este análisis, se forma una configuración de grupos muy distinta de la anterior, especialmente debido a que en el primer ejemplo no se introdujo una variable muy importante para describir la variable dependiente: la práctica religiosa. En esta ocasión son cinco los grupos terminales formados a partir de la muestra. Dos de ellos son muy poco favorables al aborto (menos del 10% a favor): los medio practicantes de centro-derecha y los muy practicantes. Otros dos grupos mantienen posiciones algo más favorables, pero aún bajas (en torno al 20%): los medio practicantes de izquierdas y los nada practicantes con bajo nivel de estudios. Por último, el quinto grupo está compuesto por no practicantes con estudios secundarios, medios o superiores. Entre estos últimos, más de la mitad da una respuesta positiva a este tipo de aborto.

El análisis de segmentación permite, pues, realizar una descripción de segmentos de la muestra con comportamiento u opinión distintos entre ellos. Por su propia lógica, tiende a encontrar grupos muy diferentes entre sí. Ahora bien, cuanto mejores sean los predictores introducidos, tarea que corresponde al analista, más nítida será la distribución de los distintos grupos. Por tanto, la mejor estrategia en la introducción de variables independientes es la inclusión en caso de duda: si se introduce un predictor poco relevante, el propio análisis se encarga de que no aparezca; en cambio si no se incluye un buen predictor, la calidad de la segmentación se reduce considerablemente.

4.2. *Utilidad exploratoria del análisis de segmentación*

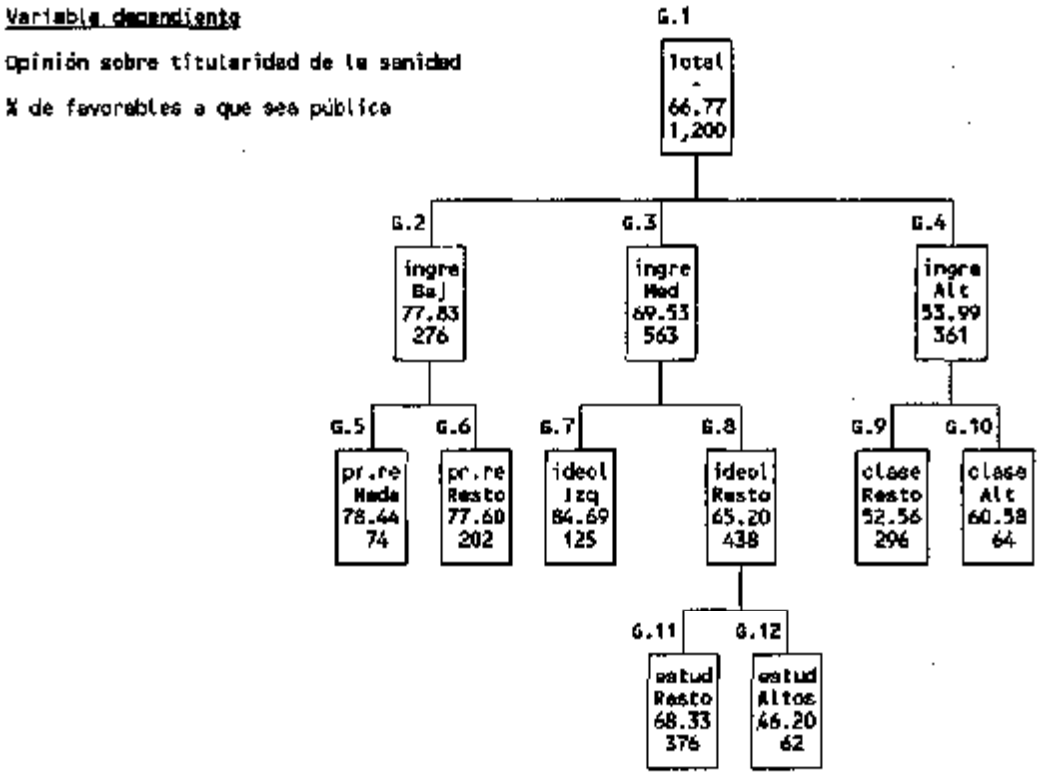
Otra utilidad adecuada del análisis de segmentación es la exploratoria. La razón radica en que su algoritmo consiste en la búsqueda de las mejores asociaciones de los predictores con la variable dependiente. En cierta medida, su potencia reside en la selección de aquellas variables que mejor expliquen una determinada distribución. Además, esta exploración permite la detección de interacciones, por lo que es un instrumento muy adecuado para buscar pautas de relaciones complejas entre variables.

Mediante el análisis del dendograma se puede: a) detectar qué variables son útiles para describir adecuadamente la variable dependiente; b) descubrir qué valores de una variable predictora son homogéneos en relación con la variable a explicar, y c) poner de manifiesto interacciones entre las variables independientes.

Si se introduce en el análisis un conjunto de variables superior al de segmentaciones que ha de efectuar el procedimiento, lógicamente algunas de ellas no se utilizarán para la

discriminación de los individuos. Esta técnica, por tanto, selecciona variables en función de su poder explicativo sobre la variable dependiente.

Tabla 25



Sea que se quiera ver qué variables predicen mejor la opinión sobre quién debiera ser el titular de los servicios de salud (tres posibles valores: titularidad pública, privada o mixta). Se dispone de trece variables sociodemográficas que podrían explicar la opinión en cuestión: sexo, edad, estado civil, posición familiar, número de individuos en el hogar, nivel de estudios del entrevistado y del cabeza de familia, religión, práctica religiosa, ideología, clase social subjetiva, identidad nacionalista e ingresos. Todas ellas se introducen en el análisis de segmentación, pero sólo una fracción de las mismas aparece en el dendograma (Tabla 25). En este caso, las variables seleccionadas son los ingresos, la práctica religiosa, la ideología, la clase social subjetiva y el nivel de estudios. Cabría decir que éstos son los mejores predictores de la opinión sobre la titularidad de los servicios de salud. Sin embargo, esto no es totalmente exacto en la medida en que las variables que aparecen en los niveles por debajo del primero (todas las mencionadas menos los ingresos) explican bien la opinión; pero no del conjunto de la muestra, sino de segmentos de ella.

Tabla 26

SIGNIFICACIONES DEL χ^2 PARA CADA GRUPO Y PREDICTOR
(Variable dependiente: Opinión sobre titularidad de la sanidad)

Predictor	G.1 (0)	G.2 (1)	G.3 (1)	G.4 (1)	G.5 (2)	G.6 (2)	G.7 (3)	G.8 (3)	G.9 (4)	G.10 (4)	G.11 (8)	G.12 (8)
Sexo	-	-	-	-	-	-	-	-	-	-	-	-
Edad	0.007	-	-	0.090	-	-	-	-	-	-	-	-
Estado civil	-	-	-	-	-	-	-	-	-	-	-	-
Posición familiar	-	-	-	-	-	-	-	-	-	-	-	-
Miembros familia	-	-	-	-	-	-	-	-	-	-	-	-
Estudios	8.6e-6	-	0.100	-	-	-	-	0.040	-	-	-	-
Estudios c.f.	5.7e-5	-	-	-	-	-	-	-	-	-	-	-
Religión	-	-	-	-	-	-	-	-	-	-	-	-
Práctica religiosa	0.300	0.020	0.400	-	-	-	-	-	-	-	-	-
Ideología	2.2e-6	-	0.092	-	-	-	-	-	-	-	-	-
Clase social subj.	2.2e-8	-	-	1.8e-6	-	-	-	-	-	-	-	-
Identidad nac.	-	0.300	-	0.060	-	-	-	-	-	-	0.040	-
Ingresos	5.9e-11	-	-	-	-	-	-	-	-	-	-	-
n:	1200	276	563	361	74	202	125	438	296	64	376	62

Por esta razón, se debe complementar la explicación con el examen de la tabla de las significaciones (Tabla 26), pudiéndose observar que en la segmentación de la muestra completa las variables mencionadas anteriormente poseen todas ellas capacidad predictiva sobre la opinión sobre la titularidad de los servicios de salud (no siempre tiene por qué ocurrir así). Sin embargo, se encuentra una variable que influye en la opinión y, no obstante, no aparece en el dendograma. Es el caso concreto de la variable edad, que presenta en el grupo inicial una significación por debajo de 0.01.

Además de seleccionar las variables que mejor predicen la variable dependiente escogida, el análisis de segmentación permite descubrir qué grupos de valores de una variable son homogéneos, y en qué sentido se produce la influencia. Contemplando el dendograma de la Tabla 25, se observa que, en función de los ingresos, existen tres grupos diferentes: los de bajos ingresos, con un porcentaje de preferencia por la sanidad pública del 77.8%, los de ingresos medios con un 69.5% que comparten esta posición y los de altos ingresos con un 54.0%. En otras segmentaciones se comprueba que a) los nada practicantes son más favorables a la sanidad pública; b) los de izquierda son más propensos a lo público; c)

paradójicamente, los que se consideran de clase alta son más partidarios de la medicina publica[13], y d) los de estudios superiores se decantan más por la privada. También en este caso, las tres últimas afirmaciones se desprenden de grupos ya segmentados.

En consecuencia, para ser más precisos, se recomienda analizar también las segmentaciones que produce cada variable en el primer nivel de la segmentación. En la Tabla 27, se observa claramente cómo son las personas mayores las más partidarias de la sanidad pública. De igual modo, se infiere que cuanto más bajos sean los estudios del entrevistado o del cabeza de familia, más favorable se será a la opción pública. Ser no practicante, de ideología de izquierda o considerarse de clase baja también son características que inclinan a defender el modelo público de medicina. Por último, la variable estadísticamente más significativa es la de los ingresos percibidos por la unidad familiar.

Además de estas tendencias, en la misma tabla se observa que hay valores con opinión muy parecida: aquéllos con cabeza de familia de estudios bajos se han agrupado con los entrevistados con persona principal de estudios medios, y los poco y medio practicantes se han fusionado con los muy practicantes. También es relevante que la categoría "No sabe, no contesta", nunca forma grupo específico, sino que se une a los valores bajos estudios del entrevistado, bajos y medios estudios del padre, práctica religiosa no nula, ideología de derechas, clase social baja e ingresos altos. Esto no quiere decir que estas categorías son las que den mayor proporción de no contestaciones; sino que los que no contestan a estas preguntas tienen opiniones sobre la titularidad de la sanidad muy parecidas a las categorías a las que se han agregado. Así, los que no contestan a la pregunta sobre los ingresos tienen actitudes semejantes a los que han declarado tener ingresos altos.

Por último, es necesario mencionar la posibilidad de detectar interacciones entre las variables. Existe interacción cuando la relación entre dos variables depende de los valores que asume una tercera. La interacción es un modelo que atañe al menos a tres variables. Por tanto, es condición necesaria para su detección que la segmentación tenga como mínimo 3 niveles, puesto que con 2 sólo, únicamente se podrá estudiar la relación bivariada entre la variable dependiente y el predictor utilizado en el primer nivel de segmentación.

El dendograma del análisis de segmentación es un medio muy útil para encontrar las principales interacciones en un conjunto de variables que tratan de explicar una dependiente. Sin embargo, tiene, como se explicará más adelante, una serie de limitaciones. De momento, véase cómo se analiza un dendograma para diagnosticar la existencia de interacciones.

---

[15]Esta paradoja procede del hecho de que la segmentación se produce entre los de ingresos altos. Si se observa el cruce original de opinión con clase social subjetiva (Tabla 27), los autodenominados de clase alta son los menos partidarios de la sanidad pública. Esto es un ejemplo claro de interacción de las variables ingresos y clase social con esta opinión.

Tabla 27

Opinión sobre la titularidad de la sanidad según los distintos valores de los predictores.

VARIABLES/Valores	Pública	Mixta	Privada	Ns/Nc	n	χ^2	p	V
EDAD						14.5	0.007	0.110
Joven	64.73	31.73	2.80	0.74	552			
Mayor	68.50	27.30	1.19	3.01	648			
ESTUDIOS						37.0	8.6E-6	0.124
Bajos, Ns/Nc	71.56	24.50	1.45	2.49	743			
Medios	64.47	31.79	2.66	1.09	274			
Altos	50.77	45.30	2.80	1.14	183			
ESTUDIOS DEL C.F.						25.6	5.7E-5	0.146
Bajos, medios, Ns/Nc	68.78	27.31	1.84	2.06	1094			
Altos	46.03	50.18	2.84	0.95	106			
PRACTICA RELIGIOSA						8.4	0.5	0.083
Ninguna	69.77	28.23	1.72	0.28	358			
Poca...Mucho, Ns/Nc	65.49	29.81	2.02	2.68	842			
IDEOLOGIA						40.0	2.2E-6	0.129
Izquierda	81.81	15.02	1.37	1.80	222			
Centro	60.56	37.11	1.60	0.73	391			
Derecha, Ns/Nc	65.22	29.57	2.36	2.85	587			
CLASE SOCIAL SUBJETIVA						50.1	2.2E-8	0.144
Alta	46.23	51.41	1.18	1.18	88			
Media	64.47	31.48	2.73	1.32	736			
Baja, Ns/Nc	76.10	19.95	0.54	3.41	376			
INGRESOS						62.9	5.9E-11	0.162
Bajos	77.83	15.77	2.23	4.17	276			
Medios	69.53	27.93	1.46	1.08	563			
Altos, Ns/Nc	53.99	41.91	2.44	1.67	361			

La primera regla es que un dendograma simétrico es indicador de ausencia de interacciones. Por simétrico se entiende que para cada segmento de un nivel dado las variables predictoras son las mismas. De los ejemplos que se han incluido en este capítulo, el primero (Tabla 13) muestra la estructura peculiar de los modelos sin interacción. En el primer nivel la variable que segmenta es la ideología; ésta divide la muestra en dos grupos: izquierda *versus* centro-derecha. Y, lo que es más importante para este caso, en cada uno de ellos el predictor más adecuado es la edad, aunque exista otro alternativo que es el sexo. No

obstante, hay que adoptar una precaución importante porque esta regla es incompleta. Para que realmente no exista interacción, las diferencias de porcentajes[16] entre los grupos segmentados han de ser semejantes en cada una de las divisiones. Siguiendo con el mismo ejemplo, la diferencia de porcentajes entre edades en el grupo de izquierdas es de 13.6 puntos, mientras que la existente entre esas mismas edades en el grupo de centro-derecha es de 11.3. Ambas cantidades son similares y si se aplicara un test de hipótesis, se evidenciaría una diferencia no significativa, esto es, posiblemente debida a errores de muestreo.

El caso contrario lo constituye el segundo ejemplo, expuesto en la Tabla 25. En esta ocasión, la primera segmentación crea tres grupos: ingresos bajos, medios y altos. En el segundo nivel de segmentación, para el primer grupo, la variable mejor predictora es la práctica religiosa; para el segundo, la ideología, y para el tercero, la clase social subjetiva. Por tanto, puede hablarse de interacción porque supuestamente -más adelante se aclarará el matiz- la práctica religiosa es la variable más influyente en la opinión sobre el aborto (la variable dependiente) sólo si los ingresos de los individuos son bajos. Si los ingresos son medios o altos, su poder de explicación desciende, al menos comparativamente con otras variables. Lo mismo ocurre con la ideología entre los de ingresos medios, y con la clase social subjetiva entre los entrevistados con altas retribuciones. Sin embargo, aquí también hay que tomar precauciones y observar la tabla de las significaciones en cada segmentación, porque en ocasiones, la introducción de una variable u otra para la segmentación efectiva es cuestión de décimas, apenas significativas. En la Tabla 26, en las columnas correspondientes a los G.2, G.3 y G.4, se confirma la existencia de interacción, en el caso de las variables práctica religiosa, ideología y clase social subjetiva (con los ingresos para la explicación de la opinión sobre el aborto), porque sólo tienen coeficiente significativo en el grupo en el que se seleccionan para la segmentación. Así, la práctica religiosa sólo influye en la opinión sobre el aborto cuando los ingresos son bajos, la ideología, cuando los ingresos son medios, y la clase social subjetiva, cuando son altos.

4.3. *Utilidad explicativa del análisis de segmentación*

El análisis fue concebido y es utilizado principalmente con una finalidad exploratoria. Su capacidad de seleccionar las asociaciones más relevantes entre un conjunto indefinido de variables puede conducir fácilmente al establecimiento de relaciones empíricas que no sean causales. La principal objeción que se le puede hacer a este análisis es que busca las asociaciones

[16] No es este el lugar para explicar las propiedades y posibilidades que ofrecen las diferencias de porcentajes, técnica empleada especialmente en las tablas de contingencia. El lector interesado puede consultar Davis (1975) y Sánchez Carrión (1984 y 1989).

empíricamente más fuertes y éstas no necesariamente tienen que ser indicadoras de una relación causa efecto entre los fenómenos.[17]

Esta objeción pierde peso siempre y cuando se introduzca entre los predictores el conjunto de fenómenos que causan realmente la variable dependiente. Aunque este requerimiento debe aplicarse a toda técnica estadística de causalidad, es particularmente importante que se aplique a esta técnica, a menos que se quiera incurrir en un alto riesgo de error.

Aprovechando el conocimiento de distintos modelos de relación entre tres o más variables[18], se detalla a continuación el comportamiento que esta técnica adopta frente a cada uno de ellos. Ello se hace así con una doble intención: mostrar la bondad de esta técnica para detectar relaciones complejas entre variables y ayudar a la interpretación de las distintas formas que presentan los dendogramas que se generan con este análisis.

4.3.1. *Modelos espurios y de intervención*

Como ambos modelos presentan las mismas características estadísticas y es el investigador quien decide si se trata de uno u otro estableciendo con la ayuda de la teoría la posición de las variables en la cadena causal, el comportamiento del análisis de segmentación es similar ante uno y otro. La cuestión reside en mostrar cómo se comporta la técnica de segmentación ante la presencia de variables antecedentes (relaciones ficticias o espurias) o intervinientes (relaciones indirectas).

Hay tres posibilidades: a) que el predictor sea totalmente ficticio b) que el predictor también contribuya a explicar por sí solo la variable dependiente y c) que haya un predictor muy efectivo no incluido en el análisis, pero que esté relacionado con un predictor falso utilizado en la segmentación.

Sean tres variables con una relación bajo el modelo espurio; por ejemplo, edad, estado civil y autodenominación liberal. Esta última será considerada variable dependiente y las dos

[17] Un análisis similar al de la segmentación es el análisis de regresión múltiple paso a paso. Se tiene una variable dependiente y del conjunto de variables independientes o predictores se introduce en la ecuación aquella variable con un coeficiente de correlación más alto con la primera. Posteriormente, se van introduciendo nuevas variables a la ecuación ya formada, asumiendo que las ya incluidas controlan la relación entre las que no están en la ecuación y la variable dependiente. Esta técnica goza de poca simpatía entre los estadísticos por su naturaleza inductivista. La posición más aceptada sería la hipotética-deductiva por la que la estadística juzga la bondad de los modelos previamente establecidos por el investigador en función de teorías plausibles.

[18] Se utiliza en este artículo la metodología y terminología que siguen distintos autores en al análisis de las tablas de contingencia. Véase Lazarsfeld (1955), Zeisel (1962), Rosenberg (1968) y Lieberson (1987). Una presentación breve puede encontrarse en el capítulo 10 del manual de Mayntz, Holm y Hübner (1975).

Tabla 28

CRUCE DE A. LIBERAL SEGUN ESTADO CIVIL Y EDAD

	EDAD			
	Joven		Mayor	
	E.CIVIL		E.CIVIL	
	Soltero	Otros	Soltero	Otros
A. LIBERAL				
No.....	40.0%	40.0%	75.0%	75.0%
Si.....	60.0%	60.0%	25.0%	25.0%
TOTAL.....	200	100	40	160

primeras predictores. Los datos tridimensionales se incluyen en la Tabla 28. Se sabe que la relación del estado civil con ser liberal no es directa, sino que viene otorgada por la relación que existe entre, por un lado, ser joven y estar soltero, y, por el otro, ser mayor y estar casado.

Tabla 29

Análisis de la muestra completa

PREDICTOR	A. LIBERAL		
Estado civil	No	Si	(100%)
Soltero	45.83	54.17	240
Otros	61.54	38.46	260
Total	54.00	46.00	500

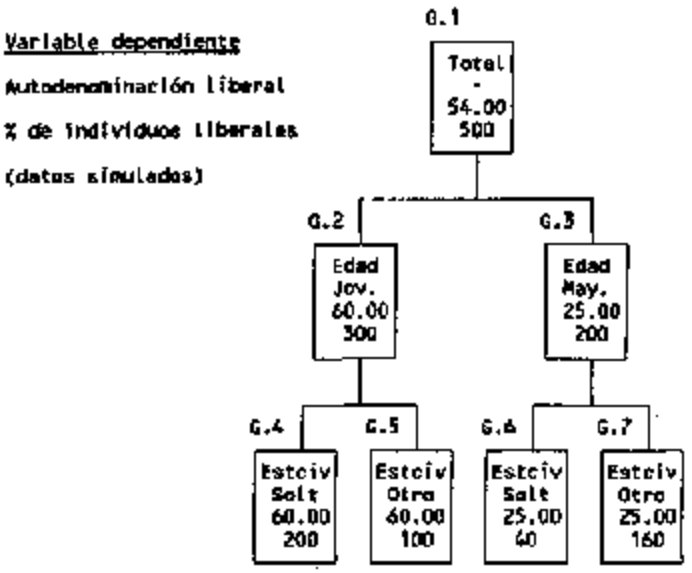
χ^2: 12.4 (g.l.=1 ; $p \leq 0.0004$)

PREDICTOR	A. LIBERAL		
Edad	No	Si	(100%)
Jóvenes	40.00	60.00	300
Mayores	75.00	25.00	200
Total	54.00	46.00	500

χ^2: 59.2 (g.l.=1 ; $p \leq 1.4e\text{-}14$)

Si la relación entre la variable espuria (predictor ficticio) y la dependiente es efectivamente nula, la segmentación sabrá discriminar adecuadamente entre la influencia real y la ficticia, porque siempre y cuando la asociación entre los predictores no sea perfecta, las asociaciones binarias con la variable dependiente serán mayores con el predictor real que con el ficticio. La razón reside en que la asociación ficticia es el producto de dos asociaciones: la que existe entre el predictor real y la variable dependiente, y la de la relación entre ambos predictores. En la Tabla 29 se verifica que la asociación de la autodenominación liberal con la edad es mayor que con el estado civil. Por tanto, la primera segmentación siempre se realiza con la variable que produce la asociación directa.

Tabla 30



Una vez realizada esta segmentación, la asociación de la variable está controlada por los valores del predictor anterior, puesto que estos últimos se mantienen constantes en cada grupo que ha de ser segmentado. En consecuencia, la variable con relación ficticia, o la variable antecedente en un modelo de intervención, no aparece como predictor relevante en la segmentación. En la Tabla 30 se presenta el dendograma del análisis incluyendo la segunda segmentación (G.4 a G.7), que en realidad no debiera aparecer, debido a la nula diferencia entre los porcentajes de las categorías de estado civil provenientes del mismo grupo.

En el segundo caso, cuando el predictor también contribuye a explicar la variable dependiente, es decir, cuando entre un predictor y la variable a explicar existe un doble efecto, directo e indirecto o ficticio, el comportamiento del análisis de segmentación es correcto, puesto que la asociación bivariada total será mayor con el predictor que posea mayor asociación directa, siempre y cuando no haya predictores importantes no incluidos en el análisis. Sin embargo, a medida que la correlación entre los predictores sea más alta, la significación del primer predictor será comparativamente mayor que la del segundo, puesto que la de aquél incorpora los efectos indirectos o espurios, mientras que la de éste sólo tiene en cuenta los directos, puesto que después de la primera segmentación, en el segundo nivel, sólo se da cuenta de los efectos directos entre el segundo predictor y la variable dependiente.

Un ejemplo de esta segunda situación sería la predicción de la misma variable dependiente por dos variables que ejercen influencia sobre ella y además están relacionadas entre sí. Sean por ejemplo, edad y práctica religiosa. Observando el cuadro de las χ^2 (Tabla 31), se ve cómo la significación de la práctica religiosa crece considerablemente en

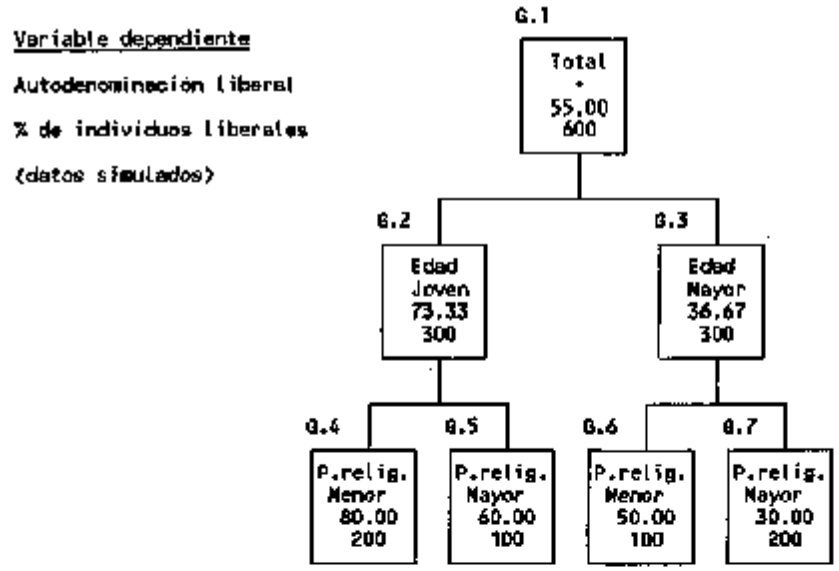
Tabla 31

SIGNIFICACIONES DEL χ^2 PARA CADA GRUPO Y PREDICTOR
(V.dependiente: autodenominación liberal)

Predictor	G.1	G.2(1)	G.3(1)
Edad	1.8e-19		
Práctica religiosa	1.5e-13	2.0e-4	7.0e-4
n:	600	300	300

el segundo nivel de segmentación (de 1.5E-13 a 2.0E-4 ó 7.0E-4). Esto, además del efecto del tamaño de la muestra, es debido a la correlación entre los dos predictores. De igual manera, cabe pensar que el coeficiente del G.1 correspondiente a la edad, está sobreestimado por el efecto indirecto de la edad hacia la actitud liberal a través de la práctica religiosa.

Tabla 32



A pesar de estos problemas, el análisis de segmentación no yerra en la discriminación de la variable más influyente. Como puede comprobarse en el dendograma, la edad precede en nivel a la práctica religiosa por su mayor capacidad de segmentación (Tabla 32).

Cosa muy distinta sucede cuando en el análisis no se incluye como predictor una variable relevante, especialmente si ésta está asociada con otra que no posee efectos directos sobre la dependiente (Tabla 33). Si se realiza una segmentación con estado civil y práctica religiosa para explicar la autodenominación liberal, se observa que el primero es mejor predictor que la segunda.

Tabla 33

Variable dependiente

Autodenominación liberal

% de individuos liberales

(datos simulados)

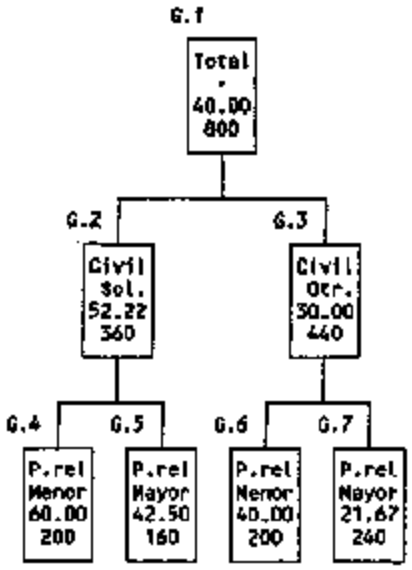


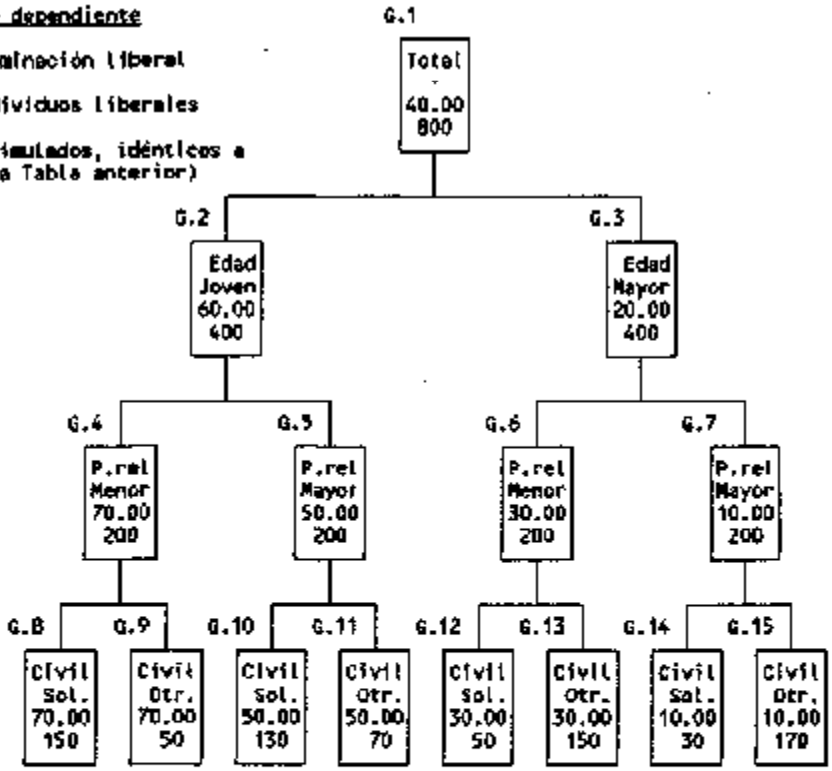
Tabla 34

Variable dependiente

Autodenominación liberal

% de individuos liberales

(datos simulados, idénticos a los de la Tabla anterior)



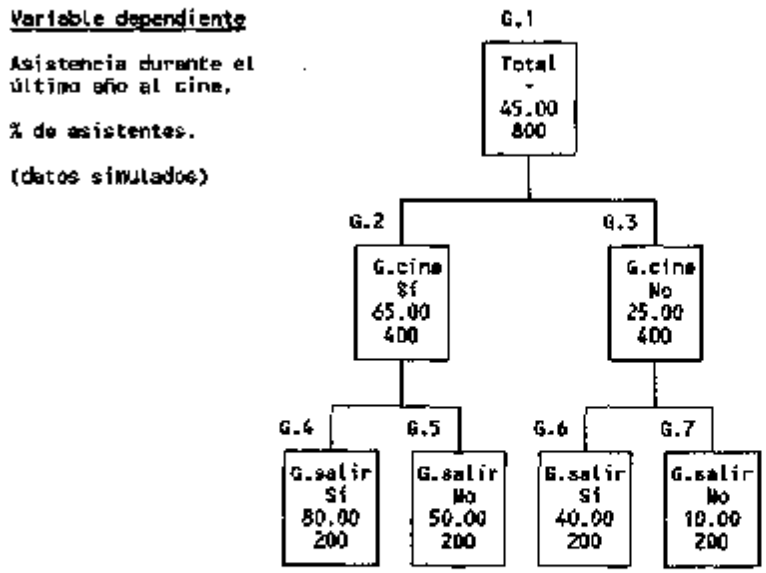
Sin embargo, esta conclusión es ficticia en la medida en que el estado civil no es el verdadero predictor de la autodenominación liberal, sino la edad. Si en el análisis se incluye esta última variable, el esquema resultante es muy distinto, pues el estado civil se queda sin

explicar nada y la práctica religiosa le supera en poder discriminante, al contrario de lo que ocurría en el ejemplo anterior donde no se incluía la edad. Como se puede ver reflejado en el último nivel de la Tabla 34, los grupos formados por la variable estado civil no son distintos en la variable dependiente.

4.3.2 *Modelos multicausales.*

Siempre y cuando la segmentación se produzca en más de un nivel, puede hablarse de multicausalidad. La razón es muy simple pues en estos casos más de una variable contribuye a explicar la variable dependiente. Dentro de los modelos multicausales se pueden distinguir los *aditivos*, si las variables influyen de forma independiente entre sí, y los *interactivos*, en el caso de que una determinada configuración de los valores de dos o más predictores produzcan valores singulares en la variable dependiente. En la práctica, un modelo es aditivo cuando en un determinado nivel k son semejantes todas las diferencias de porcentajes calculadas entre los grupos que proceden del mismo segmento. Y es interactivo cuando esas diferencias no son de la misma magnitud. Por ejemplo en la Tabla 35, la diferencia de porcentajes entre los G.4 y G.5 (80% y 50% respectivamente) es similar, en este caso idéntica, a la de los G.6 y G.7 (40% y 10%). En cambio, en el ejemplo B) de la Tabla 37, la diferencia de porcentajes entre los grupos G.4 y G.5 (90% y 70%) es sensiblemente menor que la de los G.6 y G.7 (70% y 10%).

Tabla 35

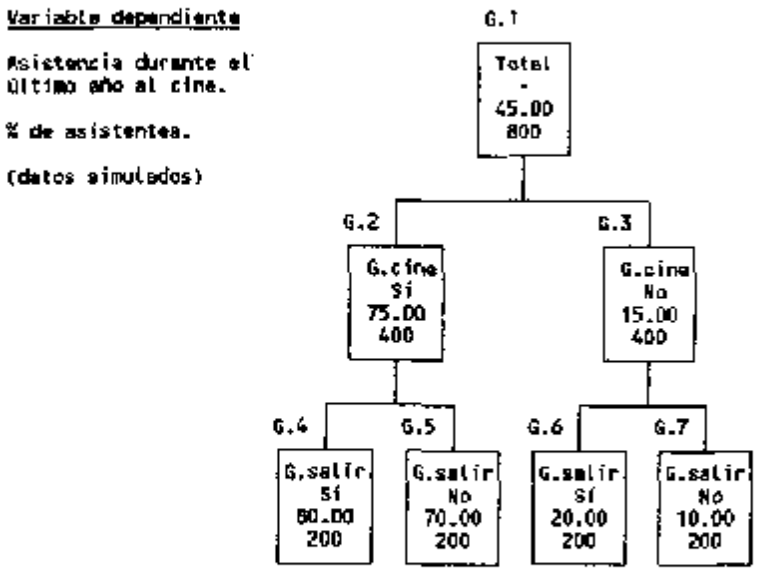


A su vez, los modelos aditivos pueden clasificarse en *equilibrados* y *no equilibrados*, según la importancia relativa de los predictores que influyan sobre la variable dependiente. La segmentación posee la virtualidad de poner en primer término el predictor con mayor

asociación con la variable dependiente. Así pues, en la situación extrema de equilibrio, los dos predictores tendrán la misma significación y aparecerán indistintamente en el primer o el segundo nivel de segmentación. Dado que esto ocurre muy rara vez, discrecionalmente se define un modelo aditivo equilibrado cuando la mínima diferencia de porcentajes entre dos grupos del nivel k+1 pertenecientes a *distintos* grupos del nivel k es menor que la mínima diferencia de porcentajes entre dos grupos procedentes del *mismo* grupo del nivel k.

Un par de ejemplos con datos simulados aclararán estas comparaciones de diferencias de porcentajes. Tómese como variable dependiente la asistencia a un cine durante el último año y como predictores si a los individuos les gusta esta actividad y si les gusta salir de casa para disfrutar del ocio. En la Tabla 35 aparece un modelo aditivo equilibrado. Ya se ha visto que es aditivo, porque la diferencia de porcentajes entre los G.4 y G.5 (formados en la segunda segmentación a partir del G.2) es idéntica a la de los G.6 y G.7 (también formados en la segunda segmentación a partir esta vez del G.3). Además, es equilibrado, porque la diferencia entre el G.5 y el G.6 (segmentos, que perteneciendo a distintos grupos del primer nivel, uno al G.2 y el otro al G.3 ofrecen la mínima diferencia de porcentajes[19]) es menor que la que se produce entre los G.4 y G.5 ó G.6 y G.7 (ambas similares por ser un modelo aditivo). En este caso, se podría decir que la asistencia al cine depende tanto de que a uno le guste este arte como de que le guste salir de casa.

Tabla 36



[19] Otras posibilidades de comparación hubiesen sido el G.4 con el G.6 (diferencia= 40%), el G.4 con el G.7 (diferencia= 70%) y el G.5 con el G.7 (diferencia= 40%). Por tanto, la mínima es la que se produce entre el G.5 y el G.6 (diferencia=10%).

En cambio, con los datos de la Tabla 36, la influencia relativa de las dos variables varía. Con estos datos, para determinar la asistencia al cine, que al individuo le guste esta actividad es bastante más importante que le agrade salir o no. Aunque le encante salir, si no le gusta el cine, tendrá poca probabilidad (20%) de entrar en él. Y, si le agrada el cine, aunque no le guste salir, la posibilidad de que acuda a ver una película es alta (70%). En este ejemplo hay desequilibrio déla multicausalidad. Una variable influye mucho más que la otra, porque la mínima diferencia entre grupos derivados de diferentes segmentos del nivel anterior (50%) es mayor que las diferencias producidas por la segunda segmentación (10%).

La utilización de datos reales no confirmaría probablemente ninguno de los dos modelos anteriores y, en su lugar, proporcionaría indicios de interacción entre los dos predictores para la explicación de la asistencia al cine. Si a un individuo no le gusta el cine, el que le guste salir o no va a tener un efecto pequeño sobre su asistencia. Sin embargo, gustándole ver películas, la influencia del gusto por salir será considerable.

De la misma forma que se podían distinguir dos submodelos aditivos, entre los modelos interactivos hay también distintos tipos. Un modelo recibe la denominación de interactivo de efectos *acumulados,* cuando la variable dependiente tiene un valor anómalo en una especial combinación de valores de dos variables. Por ejemplo, un determinado objeto de consumo suntuario sólo será comprado por las personas que cumplan dos condiciones: que tenga el dinero y que le guste o encuentre útil la posesión de dicho objeto. El porcentaje de compradores será muy bajo para los grupos 1) con escasez de recursos e interés por el bien, 2) con escasez de recursos pero interesado por el producto y 3) sin escasez de medios económicos y no interesados. En cambio, en el grupo de dotados de medios y adictos al bien habrá un alto porcentaje de propietarios de éste.

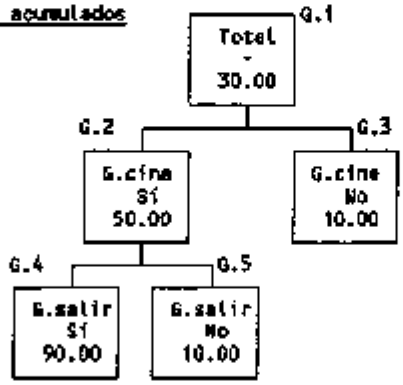
La interacción será de efectos *diferenciados,* cuando haya distintas diferencias de porcentajes entre los grupos procedentes de segmentos distintos y dichas diferencias presenten el mismo signo (todas positivas o negativas). Este modelo implica que la segunda variable segmentadora tiene mayor influencia para un grupo de sujetos que para otros. Se diferencia del anterior en que el segundo predictor tiene influencia, aunque con distinto grado, en todos los grupos y no en un subconjunto de ellos como es el caso de la interacción de efectos acumulados. Por ejemplo, en la compra de un bien no suntuario, la posesión de recursos tiene mayor influencia entre aquellos a los que les gusta el bien, que entre aquellos a los que no les gusta.

En tercer lugar, se denomina de efectos *opuestos* a aquel submodelo interactivo en el que las diferencias de porcentajes entre los grupos son de signo distinto. Ello implica que el efecto de una segunda variable es radicalmente distinto según sean los valores de la

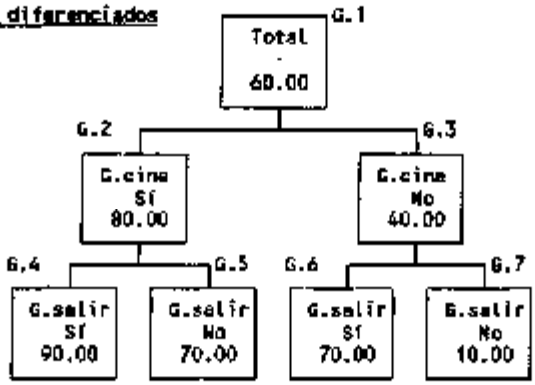
primera. Un determinado medicamento puede tener efectos muy positivos para un enfermo y consecuencias desastrosas para una persona con salud.

Tabla 37

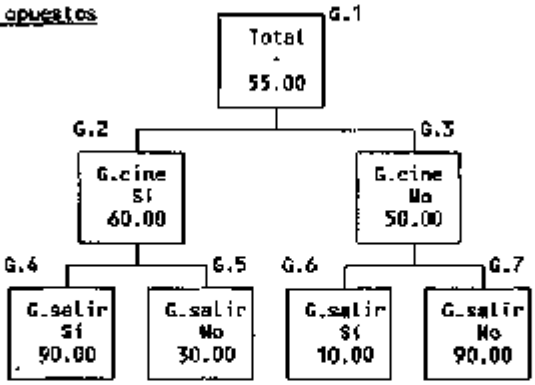
(A) Efectos interactivos acumulados



(B) Efectos interactivos diferenciados



(C) Efectos interactivos opuestos

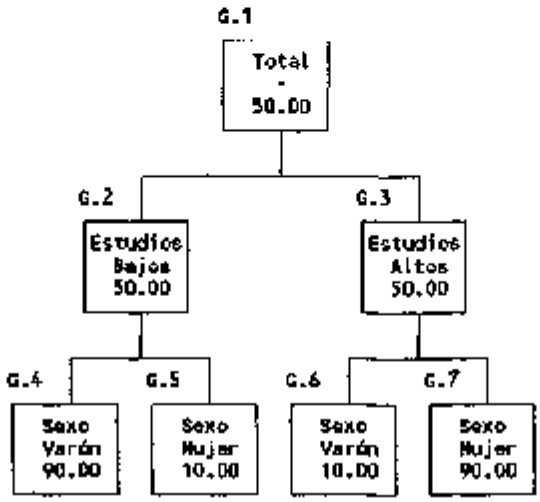


Según fueran los tipos de interacción, los gráficos de segmentación presentarían las siguientes formas y valores de los porcentajes (Tabla 37): Un modelo de efectos interactivos acumulados saldría asimétrico y el porcentaje de uno de los grupos segmentados debería ser semejante al del grupo no segmentado del nivel anterior (modelo A); en un modelo de interacción diferenciadora, aun pudiendo aparecer simétrico, las diferencias de porcentajes

serían distintas para cada uno de los grupos de los que procede (modelo B); por último, un modelo de efectos interactivos opuestos, presentaría diferencias de porcentajes de diferente signo (modelo C).

Tabla 38

Modelo de efectos interactivos opuestos. Caso de no detección



Un caso extraordinario que, paradójicamente, no es capaz de resolver este análisis automático de interacción se produce si existen efectos interactivos opuestos que se anulen entre sí. Debido a este hecho, esta técnica se revela totalmente incapaz de detectar este tipo de interacciones, incluso llegando a ocultar verdaderas relaciones entre variables. En la Tabla 38, los datos dan muestra clara de una interacción entre sexo y estudios. Sin embargo, el análisis de detección automática de interacciones sería incapaz de descubrirla, porque en el momento de hacer la primera segmentación se encontraría que ni el sexo, ni los estudios poseen capacidad explicativa sobre la variable dependiente. Y nada más lejos de la realidad, por lo que se refiere al sexo, por citar el más evidente, que diferencia en gran medida a la variable dependiente en cada uno de los subgrupos formados por la variable estudios.

Sin embargo, en el análisis de segmentación lo más frecuente es encontrar las interacciones bajo la forma de dendogramas asimétricos. Como se explicó en el apartado de la utilidad exploratoria de esta técnica, la segmentación en un mismo nivel por distintas variables es indicio de interacción. Ahora bien, en estos casos, resulta más difícil calificar el tipo de interacción. El dendograma no ofrece, en principio, ningún dato para la clasificación. En cambio, la tabla de significaciones del χ^2, permite, al menos, distinguir las interacciones de efectos acumulados de las otras dos. Para que una interacción sea del primer tipo, es condición necesaria, aunque no suficiente, que en uno de los grupos de un nivel de

segmentación, la relación con la variable dependiente sea no significativa. En la Tabla 26, por ejemplo, la interacción de la práctica religiosa con los ingresos para explicar la opinión sobre la titularidad de los servicios públicos podría ser de tipo acumulado, porque la relación de la práctica religiosa con la opinión sólo es significativa cuando los ingresos son altos. La otra condición se obtiene a partir del dendograma y es que haya al menos un grupo del segundo predictor con un valor de la variable dependiente semejante al resto de grupos del primer predictor. Así, en la Tabla 25, el G.6 tiene un porcentaje de favorables a la sanidad pública del 77.6%, bastante superior al resto de grupos de la variable ingresos (G.3 y G.4, con valores medios y altos y porcentajes 69.5% y 54.0% respectivamente), por lo que este ejemplo no se puede clasificar como interactivo de efectos acumulados.

Para distinguir los otros dos tipos de modelos interactivos es preciso realizar los cruces necesarios y estudiar el comportamiento de las distintas diferencias de porcentajes. Otra posibilidad sería hacer que las variables segmentadoras sean las mismas para cada nivel de segmentación y, de esta manera, estudiar el comportamiento de los porcentajes que presentan los distintos grupos formados por los valores de los predictores. Pero esto también implicaría volver al punto de partida: la tabla de contingencia.

5 Sumario a modo de conclusiones.

El análisis de segmentación es una técnica de análisis de datos basada en la dependencia entre variables, cuya finalidad es la de formar grupos, configurados con valores de las variables independientes, que sean muy distintos entre sí en la variable dependiente. La lógica de su procedimiento se sustenta en los siguientes pasos: a) agrupación de categorías de los predictores, b) selección de los mejores predictores y c) sucesivas segmentaciones sobre los grupos formados a partir de los pasos anteriores.

Hay distintos algoritmos que reciben la denominación de análisis de segmentación. Este artículo se centra en el algoritmo CHAID, basado en la métrica del x^2; pero también da cuenta de las similitudes y diferencias de éste con otros procedimientos como el clásico AID de Morgan y Sonquist y el algoritmo THAID para variables dependientes nominales.

La utilidad del análisis de segmentación es múltiple. Está especialmente diseñado para propósitos descriptivos o exploratorios. Sin embargo, también puede ser útil para un previo análisis causal de las variables. A través de los resultados del análisis se pueden realizar hipótesis de modelos de causalidad, como el espurio, el de intervención y los distintos modelos de causalidad, entre los que destaca el tipo de interacción entre las variables.

Referencias bibliográficas

Alvira, F., García López, J. y Horter, K. (1982), "La situación de la vivienda en España", *Papeles de Economía Española*, 10, pp. 208-247.

Belson, W.A. (1959), "Matching and Prediction on the Principle of Clasification", *Applied Statistics*, 8, pp. 195-202.

Bouroche, J.M. y Tennenhaus, M. (1972), "Some Segmentation Methods", *Metra*, 7, pp. 407-418.

Cellard, J.C., Labbe, B. y Cox, G. (1967), "Le programme Elisée. Presentation et application", *Metra*, 3, pp. 511-519.

Davis, J.A. (1975), *Análisis elemental de encuestas*, México, Trillas.

Fielding, A. (1977), "Binary Segmentation: The Automatic Interaction Detector and Related Techniques for Exploring Data", en C.A. O'Muircheartaigh y C. Payne (eds.), *The Analysis of Survey Data*, New York, Wiley.

García Ferrando, M. (1982), *Regionalismo y autonomías en España, 1976-1979*, Madrid, Centro de Investigaciones Sociológicas.

Hawkins, D.M. y Kass, G.V. (1982), "Automatic Interaction Detection" en D.M. Hawkins (ed.), *Topics in Applied Multivariate Analysis*, Cambridge, Cambridge University Press.

Horter, K. (1978), "Análisis multivariable de los votos político y sindical", *Revista Española de Investigaciones Sociológicas*, 1, pp. 145-158.

Kass, G.V. (1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data", *Applied Statistics*, 29, pp. 119-127.

Lazarsfeld, P.F. (1955), "Interpretation of Statistical Relations as a Research Operation", en P.F. Lazarsfeld y M. Rosenberg (eds.), *The Language of Social Research*, Glencoe, The Free Press.

Lieberson, S. (1987), *Making it Count*, Berkeley, University of California Press.

Madgison, J. (1989), *SPSS/PC+ CHAID*, Chicago, SPSS Inc.

Mayntz, R., Holm, K. y Hübner, P. (1975), *Introducción a los métodos de la sociología empírica*, Madrid, Alianza.

Messenger, R.C. y Mandell, L.M. (1972), "A Modal Search Technique for Predictive Nominal Scale Multivariate Analysis", *Journal of the American Statistical Association*, 67, pp. 768-772.

Morgan, J.N. y Sonquist, J.A. (1963), "Problems in the Analysis of Survey Data", *Journal of the American Statistical Association*, 58, pp. 415-434.

Rosenberg, M. (1968), *The Logic of Survey Analysis*, New York, Basic Books.

Ruiz-Maya, L. *et al.* (1990), *Metodología estadística para el análisis de datos cualitativos*, Madrid, Centro de Investigaciones Sociológicas.

Sánchez Carrión, J.J. (1984), "Análisis de las tablas de contingencia. Sistema de las diferencias de proporciones" en J.J. Sánchez Carrión (ed.), *Introducción a las técnicas de análisis multivariable aplicadas a las ciencias sociales*, Madrid, Centro de Investigaciones Sociológicas.

Sánchez Carrión, J.J. (1989), *Análisis de tablas de contingencia*, Madrid, Centro de Investigaciones Sociológicas.

Sánchez Cuenca, I. (1990), "La segmentación", en E. Ortega, *Manual de investigación comercial*, Madrid, Pirámide.

Smith, W. (1956), "Product Differentiation and Market Segmentation as Alternative Marketing Strategies", *Journal of Marketing*.

Sonquist, J. A. y Morgan J.N. (1964), *The Detection of Interaction Effects*, Survey Research Center Monograph, nº 35, Ann Arbor, Institute for Social Research, University of Michigan.

Sonquist, J.A. (1971), *Multivariate Model Building. The Validation of a Search Strategy*, Ann Arbor, Institute for Social Research, University of Michigan.

Zeisel, H. (1962), *Dígaselo con números*, México, Fondo de Cultura Económica.