Técnicas de analisis de datos nominales

Juan Javier Sánchez Carrión
Universidad Complutense de Madrid

EL ANALISIS DE DATOS NOMINALES

El objetivo de este trabajo es mostrar diferentes técnicas que permiten el análisis de datos nominales. Cada una de las técnicas se explica en otros libros y artículos, a los que haremos referencia cuando proceda. Aquí juntamos todas las técnicas para ilustrar su aplicación, y remitimos a libros y artículos para aprender su funcionamiento.

Las aquí llamadas variables nominales, otros autores las denominan cualitativas o categóricas. En todos los casos tenemos una variable, cuyas respuestas vienen expresadas en nombres y no en números. Ejemplos característicos de variables nominales serían el Sexo, con las categorías "hombre" y "mujer"; la Religión, etc. (Sánchez Carrión, 1989).

Dado que las técnicas de análisis de datos intervales ofrecen más información que las técnicas de análisis de datos nominales y, también, debido a un mayor desarrollo de las primeras, hay una tendencia generalizada a "elevar" de categoría a las variables nominales hasta convertirlas en intervales. Pensamos que no siempre está justificado el cambio y aquí vamos a mostrar técnicas de análisis que permiten trabajar con las variables nominales sin necesidad de transformarlas.

Digamos directamente que todas las técnicas de análisis de datos nominales pasan por la construcción de tablas de doble, triple, etc., entrada, a partir de las cuales se realizan ciertas manipulaciones. Estas son las manipulaciones que vamos a mostrar:

- Cálculo de proporciones (porcentajes) y construcción de un Sistema de Proporciones (Sistemas de la D).
- Construcción de un fichero de datos agregados a partir de las categorías de una de las variables nominales (Ficheros de Datos Agregados).
- Cálculo de razones y ajuste de un modelo Log-linear (Modelos Log-linear).
- Representación gráfica en forma de "Postes".
- Representación gráfica mediante el Análisis de Correspondencias.
- Análisis de Tablas con 3 o más dimensiones.

1.- El uso de los porcentajes y el análisis de variables nominales

La tabla 1 utiliza unos datos inventados para mostrar la relación entre las variables Estado Civil y Práctica Religiosa -datos tomados de Sánchez Carrión, 1989. Con el fin de poder establecer comparaciones entre los distintos estados civiles hemos puesto los datos en porcentajes.

	Estado civil			
Pra. religiosa	soltero	casado	viudo	separado/div.
Nunca	60	10	33	60
Alguna vez	30	30	33	30
Siempre	10	60	33	10
Total	100	100	100	100.-
	(200)	(500)	(65)	(765)

Tabla 1
Relación entre Estado Civil y Práctica Religiosa.(%)

Si queremos ver la relación entre ambas variables de nada servirían los estadísticos al uso en la investigación social (Gi-cuadrado, Lambda, 5 de Cramer, etc.). Supongamos que un investigador calcula un estadístico para resumir la relación y se presenta ante la persona que encargó la investigación diciendo que la relación entre ambas variables es de 0.6.

Si el estadístico va de 0.0 a 1.0, el valor de 0.6 significa una alta relación entre las 2 variables. Supongamos también que el patrocinador quiere aumentar la práctica religiosa de los entrevistados, ¿le informa este 0.6 sobre qué colectivo ha de actuar? No, tan sólo le dice que hay relación entre ambas variables. Si quiere conocer la naturaleza de esta relación necesariamente ha de acudir a mirar los porcentajes, comparando parejas de categorías.

Por ejemplo, mirando en la Tabla 1 las columnas de "soltero" y "separado/dvdo." vemos que ambos grupos se comportan de igual manera: el 60% "nunca" va a misa. Por el contrario, "solteros" y "casados" tienen comportamientos diferentes: agrupando las categorías "alguna vez" y "siempre" y utilizando las diferencias de porcentajes como medida de asociación, comprobamos que entre los "casados" hay un 50.0% más de individuos que van a misa ("alguna vez" o "siempre") que entre los "solteros".

Si quisiéramos ver la diferencia entre "casados" y "viudos" tendríamos que calcular un número diferente; y lo mismo ocurriría para comparar "solteros" con "viudos", etc., Con este estadístico (diferencia de porcentajes o de proporciones)

construimos un sistema de análisis, llamado Sistema de la D (Davis, 1976; Sánchez Carrión, 1989). En los Apartados 2 y 6 veremos algo más sobre la utilización de las diferencias de proporciones.

MORALEJA

Tratándose de variables nominales no hay ningún único estadístico que resuma su relación. Para estudiar este tipo de variables hay que olvidarse de las variables y mirar las categorías, comparándolas entre sí. El 0.6 es un estadístico que de tanto querer resumir la relación entre Estado y Práctica termina por ofrecer una información irrelevante.

2.- Análisis de los datos agregados

En la Tabla 2 ofrecemos unos datos inventados con el resultado de entrevistar a 7 trabajadores, a los que se les pregunta -entre paréntesis incluimos los códigos con los que se habrían grabado los datos en el ordenador - :

- Tipo de Empresa en la que trabajan: Metal (1) o Textil (2).
- Su Voto en las últimas elecciones legislativas: partidos de Izquierda (1) o Derecha (2).
- Su Ocupación Manual (1) o No Manual (2)
- Su Voto en las elecciones sindicales: Sindicatos de Clase (1) o No- clase (2), y
- La Edad (en años).

Id	Empresa	Voto Legisl.	Ocupación	Voto Sindical	Edad
1	metal	izda.	manual	clase	40
2	metal	izda.	no-manual	clase	30
3	metal	dcha.	no-manual	clase	45
4	textil	dcha.	no-manual	no-clase	60
5	textil	dcha.	manual	no-clase	55
6	textil	dcha.	manual	no-clase	64
7	textil	izda.	manual	clase	60

Tabla 2
Matriz de Datos Individuales

A partir de esta matriz de datos podemos ver las características de los trabajadores, según sean del metal o del textil. Para ello podemos construir múltiples tablas de contingencia, que relacionen la variable Empresa con las restantes, o podemos crear un nuevo fichero de datos agregados, en el que los casos (las unidades) sean cada uno de los tipos de Empresa, y las variables sean derivaciones de las variables individuales.

Utilizando las siguientes instrucciones SPSS/PC+ (Norusis, 1985; Sánchez Carrión, 1988a) construimos la Tabla 3, en la que se incluyen 2 casos (metal y textil) y 4 variables.

- Edadmedi: edad media de los entrevistadores en cada empresa.

- Votoizda: tanto por ciento de trabajadores que votan a la izquierda en cada una de las empresas.

- Manuales: tanto por ciento de trabajadores con ocupación manual en cada una de las empresas.

- Votoclas: tanto por ciento de trabajadores que votan sindicatos de clase en cada una de las empresas.

```
AGGREGATE OUTFILE = ' Votoagre.sys'
/BREAK = Empresa
/Edadmedi = MEAN (Edad)
/Votoizda = PLT (Votolegi, 2)
/Manuales = PLT (Ocupacio,2)
/Votoclas = PLT (Votosind,2)
```

Empresa	Edadmedi	Votoizda	Manuales	Votoclas
metal	38.3	66.7	33.3	100.0
textil	59.7	25.0	75.0	25.0

Tabla 3
Fichero de datos agregados

Mirando la Tabla 3 se ve el "perfil" de los trabajadores de las empresas del metal y del textil -recordar que los datos son ficticios-: los del metal son jóvenes, votan mayoritariamente a la izquierda, tienen ocupaciones No-manuales y todos, sin excepción, votan sindicatos de Clase; todo lo contrario se podría decir de los trabajadores del textil.

Una vez que tenemos los datos agregados podemos juntarlos con los individuales para estudiar la influencia del contexto en el comportamiento de los individuos -para ver si la gente se atiene al refrán que dice "donde fueres haz lo que vieres". Realizamos esta operación utilizando de nuevo SPSS/PC+ (Sánchez Carrión, 1988a):

```
JOIN MATCH FILE = 'Voto. sys'/TABLE = 'Votoagre.sys'
/BY Empresa.
```

Nota: 'Voto.sys' contiene los datos individuales.

El resultado sería un fichero como el que incluimos en la Tabla 5.

Id.	Empresa	Voto Legis.	Ocupación	Voto Sindical	Edad	Med.	izda.	ales	clase
1	metal	izda	manual	clase	40	38.3	66.7	33.3	100.0
2	metal	izda	no-man	clase	30	"	"	"	"
3	metal	dcha	no-man	clase	45	"	"	"	"
4	textil	dcha	no-man	no-clase	60	59.7	25.0	75.0	25.0
5	textil	dcha	manual	no-clase	55	"	"	"	"
6	textil	dcha	manual	no-clase	64	"	"	"	"
7	textil	izda	manual	clase	60	"	"	"	"

Tabla 5
Unión de los ficheros de datos agregados e inventados

La forma de ver la validez del refrán pasa por hacer lo que en la investigación social se llama un Análisis Contextual. Para ello miramos la relación entre 2 variables, repitiendo el análisis con el añadido de una tercera (el "contexto"). Por ejemplo, podemos ver si el contexto político de la empresa (% de trabajadores que votan a la izda: Votoizda) afecta al voto de los trabajadores manuales.

La tabla 6.a muestra la relación entre Ocupación y Voto Sindical. A partir de esta tabla se observa que el 50% de los trabajadores manuales votan a sindicatos de clase. Mirando la Tabla 6.b vemos que cuando estos mismos trabajadores manuales se encuentran en un contexto de izquierdas (más del 50% de los trabajadores votan a la izquierda: Votoizda) el porcentaje de los que votan a sindicatos de clase se eleva al 100.0%. En un contexto de derechas, este mismo tanto por ciento queda reducido al 33.3%.

(a)

	Ocupación	
Voto sindical	manual	no-manual
clase	50.0	66.7
no-clase	50.0	33.3
Total	100.-	100.-
	(4)	(3)

(b)

Votoizda	
izquierda (+ 50%)	derecha (- 50%)

Voto Sind.	Ocupación manual	no-manual	Ocupación manual	no manual
clase	100.0	100.0	33.3	0.0
no-clase	0.0	0.0	66.7	100.0
Total	100.0	100.0	100.0	100.0
	(1)	(2)	(3)	(1)

Tabla 6
Relaciones entre Ocupación y Voto sindical (a) y entre Ocupación, Voto sindical y Voto legislativas (b)

3.- Modelos lineales logarítmicos

Los modelos lineales logarítmicos (en inglés, log-linear) también tienen por finalidad estudiar la relación entre variables. Mientras que los sistemas de las diferencias de proporciones están basadas en diferencias de proporciones, valga la redundancia, los modelos log-linear tienen su fundamento en las razones.

Para mostrar la aplicación de las razones como medida de asociación a partir de la Tabla 1 construimos una nueva tabla en la que agrupamos las categorías "alguna vez" y "siempre" y consideramos sólo a "solteros" y "casados". En la Tabla 7 se recogen los resultados.

Pra. religiosa	Estado civil soltero	casado	total
nunca	120	50	170
alguna + siempre	80	450	530
Total	200	500	700

Tabla 7
Cruce de Estado civil y Práctica religiosa

Podemos ver la razón de "nunca" a "alguna..." para los "solteros" y para los "casados". En el primer caso hay 1.5 individuos que nunca van a misa por cada uno que sí va (es decir, 120 entre 80). Entre los "casados" esta razón es de 0.11 a 1.00 (es decir, 50 dividido entre 450). Puesto que según se trate de solteros o de casados las razones son diferentes, podemos decir que estamos en presencia de una asociación entre ambas variables. ¿Cuál es la intensidad de esta asociación? Muy sencillo, basta dividir ambas razones para encontrar el estadístico medida de la relación:

$$(1.50/0.11) = 13.5$$

A este número le llamamos "razón de razones" (en inglés, odds ratio), y es igual al producto cruzado de las frecuencias de la Tabla 7:

$$(120x450)/(80x50) = 13.5$$

Una vez visto este estadístico digamos que en los modelos log-linear lo que vamos a hacer es tratar de ajustar modelos que expliquen la frecuencia de cada una de las casillas de las tablas. Por ejemplo, si comparamos la tabla 8 (versión de la Tabla 7, en la que los datos aparecen en tantos por uno) con otra donde las frecuencias en cada casilla fueran iguales (Tabla 9) podríamos preguntarnos a qué es debida la diferencia entre ambas, ¿qué factores (efectos) están actuando para que los números sean todos diferentes?

	Estado civil		
Pra. religiosa	soltero	casado	total
nunca	.60	.10	.24
alguna + siempre	.40	.90	.76
Total	.29	.71	1.00

Tabla 8
Cruce de Estado civil y Práctica religiosa, expresado en tantos por 1.0

	Estado civil		
Pra. religiosa	soltero	casado	Total
nunca	.25	.25	.50
alguna + siempre	.25	.25	.25
Total	.50	.50	.50

Tabla 9
Cruce de Estado civil y Práctica religiosa en el supuesto de equiprobabilidad de todas las casillas (%).

Hay 4 efectos que están determinando el tamaño desigual de las casillas de la Tabla 8:

- Por un lado está la influencia que tiene en las casillas el hecho de que no haya igual número de "solteros" que de "casados". Este es el efecto de la variable "columna".

- Igualmente influye en el tamaño desigual el hecho de que tampoco haya igual número de personas que van o que no van a misa (efecto de la variable columna).

- Una tercera influencia tiene que ver con la mayor probabilidad de no ir a misa cuando se es soltero que cuando se es casado (efecto atribuible a la relación entre ambas variables).

- Un último efecto es atribuible al tamaño medio de las casilla.

Cada uno de estos efectos se puede calcular y todos juntos explicarán la frecuencia de cada una de las casillas de la Tabla 8. Por ejemplo, el efecto atribuible al tamaño medio de las casillas es igual a la media geométrica de la frecuencia de las 4 casillas, y se identifica con la letra griega Mu (μ):

$$\mu = \sqrt[4]{(120x50x80x450)} = 1/4\ (1g\ 120 + ... + 1g\ 450) = 4.797$$

El efecto debido a la asociación es igual a la raíz cuarta del producto cruzado y se identifica como λ_{AB} :

$$\lambda_{AB} = \sqrt[4]{(120x450\ /\ (80x50)} = 1/4\ 1g\ (120x450)\ /\ (80x50) = 0.650$$

El efecto de las filas y de las columnas es igual a la media geométrica de las frecuencias en una categoría respecto de las frecuencias en la otra categoría. Por ejemplo, el efecto de las columnas, λ_B , es igual a:

$$\lambda_B = \sqrt[4]{(120x50\ /\ (80x450)} = 1/4\ (1g\ 120 + 1g\ 50 - 1g\ 80 - 1g\ 450)' = -0.448$$

Calculando el efecto de la variable columna (-0.213) y poniendo todos los resultados juntos tenemos que:

$$lg_e\ 120 = 4.79 + 0.650 - 0.448 - 0.213 = 4.\ 786$$

Comprobamos de esta manera que el modelo que hemos ajustado, y que incluye todos los efectos que están influyendo en la tabla, explica la frecuencia de las casillas.

En el supuesto de que la tabla tuviera más de 2 dimensiones (variables) habría más efectos a considerar. En ambos casos, el problema no concluye con el cálculo de todos los efectos, sino que se hace necesario ver si no sería posible ajustar algún otro modelo con menor número de efectos, y que siguiera explicando nuestros datos. Entramos así en un problema que implica realizar sucesivos contrastes hasta encontrar el modelo con menor número de efectos y que ajusta los datos (ver Fienberg y Holland, 1975).

4.- Representación gráfica de las tablas: los *postes telegráficos*

Tomando datos de los Estados Unidos (Davis, 1987) la Tabla 10 muestra la relación entre las variables Estatus y Hábito de fumar. La primera variable tiene las categorías "alto", "medio" y "bajo". La variable Fumar admite las categorías "nunca" - nunca fumó -, "dejó" - fumaba, pero lo dejó - "empezó" - no fumaba, pero ahora fuma - y "siempre" - fumaba y sigue fumando.

			Fumar			
Estatus	nunca	dejó	empezó	siempre	Total	
bajo	34.6	18.0	33.1	14.3	100.-	(1111)
medio	40.3	20.4	28.2	11.2	100.-	(2710)
alto	45.1	22.9	24.4	7.6	100.-	(1154)
total	40.1	20.4	28.4	11.1	100.-	(4975)

Tabla 10
Relación entre Estatus y Hábito de fumar (%)

La tabla 11 se puede representar en un gráfico, tal como hacemos en la Figura 1. En este gráfico hacemos tantos postes como categorías tenga la variable dependiente, y colocamos en cada poste el tanto por ciento de individuos en cada una de las categorías de la variable independiente.

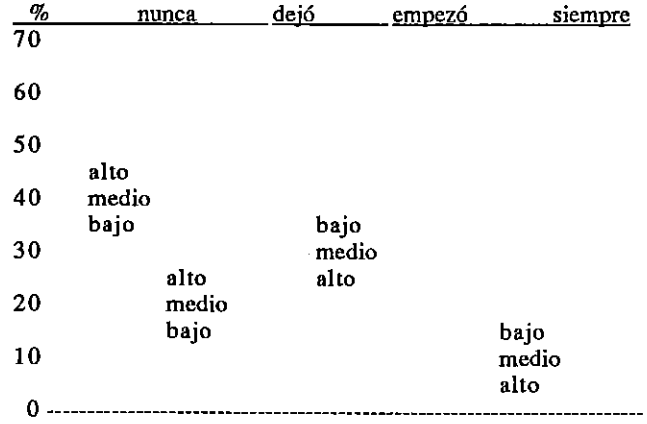

Figura 1
Representación gráfica de la relación entre Estatus y Hábito de fumar

Tanto a partir de la Tabla 10 como de la Figura 1 - pensamos que mejor a partir de la Figura 1 - se puede ver que lo que domina son las personas que "nunca" fumaron y lo que menos hay son fumadores de "siempre". Entre los que "nunca" fumaron son dominantes los individuos de estatus alto, justo lo contrario de lo que ocurre entre los fumadores -categoría "siempre"-, donde son mayoritarias las personas de estatus bajo.

5.-Representación gráfica de las tablas: el Análisis de Correspondencia

Tomando datos de García Santesmases (A.C., 1985) vamos a ilustrar la técnica del Análisis de Correspondencias -sólo haremos referencia a las correspondencias simples. El A. de C. es un método descriptivo que pretende representar en un espacio de la menor dimensión posible la relación entre las categorías de dos o más variables nominales. Supongamos que tenemos la distribución del Producto Nacional Bruto entre los sectores Agrícola, Industrial y de Servicios para un conjunto de países, tal como se muestra en la Tabla 11.

País	Agrícola	Industria	Servicio	Total
Argentina	13	46	41	100
Bolivia	17	29	54	100
Brasil	11	38	51	100
Chile	8	37	55	100
Colombia	29	28	43	100
C. Rica	19	26	55	100
Ecuador	15	37	48	100
Salvador	28	28	43	100
Guatemala	26	20	54	100
Honduras	32	26	42	100
Méjico	10	38	52	100
Nicaragua	29	28	43	100
Panamá	23	21	56	100
Paraguay	31	24	45	100
Perú	10	43	47	100
R. Domin.	19	26	55	100
Uruguay	13	37	50	100
Venezuela	6	47	47	100
U.S.A.	3	34	63	100
Canadá	4	33	63	100
Alemania	2	49	49	100
Bélgica	2	37	61	100
Dinamarca	5	59	36	100
España	9	31	60	100
Francia	5	34	61	100

Italia	7	43	50	100
P. Bajos	4	37	59	100
Portugal	13	47	40	100
G. Bretaña	2	36	62	100
Japón	5	42	53	100
Total	400	1055	1545	3000

Tabla 11
Distribución Producto Nacional Bruto por Países

A partir de la Tabla 11 se puede ver qué países tienen una mayor o menor participación en cada uno de los sectores productivos y, en función de esta información, cuáles son las semejanzas o diferencias entre los países. Por ejemplo, Italia y Japón son bastante parecidos entre sí, y diferentes a Colombia o Nicaragua.

Si quisiéramos representar gráficamente la Tabla 11 podríamos construir 2 nubes de puntos: una en la que los puntos fueran los países (30), siendo sus coordenadas los valores de sus PNB respectivos en cada uno de los sectores (3); otra, con los sectores como puntos (3) y sus valores para cada país como coordenadas (30). La segunda nube de puntos, en un espacio de 30 dimensiones, no puede ser representada gráficamente. Sí podemos representar los 30 puntos (países) en un espacio de 3 dimensiones (sectores), tal como hacemos en la Figura 2.

Tomando Argentina como ejemplo vemos que sus coordenadas son igual a:

Argentina = (13/100 + 46/100 + 41/100)
(ver figura 2 en el apéndice final)

El Análisis de Correspondencias trata de representar en un único subespacio las 2 nubes de puntos. Este espacio ha de formarse siguiendo un par de criterios:

- Que sea del menor número de dimensiones posibles, y

- Que respete las distancias originales entre los puntos: parejas de puntos distantes en los datos (las nubes de puntos), también han de estar distantes en el subespacio difinido por el A. de C.

Con el fin de calcular la distancia entre los puntos se va a utilizar una métrica espacial, la Distancia de Benzecri. Esta distancia se caracteriza por el hecho de que pondera las distancias entre los puntos de manera inversamente proporcional a sus frecuencias. A continuación ofrecemos un ejemplo de utilización de esta distancia, calculando la distancia entre Argentina y Bolivia (nube de puntos de los países).

$$d^2(i,i')=\sum_{j=1}^{q}\frac{1}{f.j}\left(\frac{fij}{fi.}-\frac{fi'j}{fi'.}\right)^2$$

$$d^2(AArg,Bol)=\frac{1}{400}\left(\frac{13}{100}-\frac{17}{100}\right)+\ldots+\frac{1}{1545}\left(\frac{41}{100}-\frac{54}{100}\right)^2=.00008$$

Estas distancias, calculadas para todas las parejas de puntos en ambas nubes, son las que hay que respetar en la solución que proporcione el A. de C. El procedimiento que se sigue para encontrar la solución final no lo vamos a explicar aquí; digamos simplemente que como resultado del análisis se obtiene la representación gráfica de los puntos en varias dimensiones, junto con una serie de informaciones que nos permiten ver la bondad de la representación y hacer su interpretación.

Para deducir el número de dimensiones miramos el valor de los "autovalores". Cada uno de ellos indica el tanto por ciento de variabilidad (en la terminología del A. de C. se suele hablar de "inercia") explicada por el autovalor o eje factorial. El primer autovalor explica el 81.0% de la inercia y el segundo el 19'0%. En conjunto, ambos autovalores explican el 100% de la inercia. Por lo tanto, en nuestro ejemplo podemos representar los puntos en un espacio de 2 dimensiones, tal como se muestra en la Figura 3.

En la Figura 3 la proximidad entre 2 puntos significa que entre ambos existe una relación positiva. Mirando la Figura 3 vemos que Honduras, Paraguay, etc. son países muy próximos a Agricultura, como corresponde a la importancia que tiene este sector en los países en cuestión - ver Tabla 11. En el caso contrario se encuentran Alemania o España, en los que predominan la industria (Alemania) o los Servicios (España).

Cuando una de las variables tiene 3 categorías siempre es posible encontrar un espacio bi-dimensional que represente adecuadamente los puntos. En el supuesto de tener más categorías y elegir sólo los primeros autovalores que expliquen, por ejemplo, un 40% de la inercia, el problema que se plantea es que los puntos aparecerán deformados: las distancias en la representación no coincidirán con las distancias en los datos originales.

Con el fin de "matizar" la representación gráfica, además de esta representación y de los autovalores, el A. de C. ofrece información sobre la importancia que tiene cada punto en la definición de los ejes y sobre la calidad de la representación de los puntos situados en un eje. La primera información recibe el nombre

de "contribución absoluta" y la segunda "contribución relativa". Las Tablas 12 y 13 muestran las contribuciones absolutas de los puntos fila y columna a los 2 ejes, así como las contribuciones relativas de los ejes a los puntos fila y columna.

	(a) Factores 1	2
Argentina	2	86
Bolivia	7	9
Brasil	2	1
Chile	9	4
.....		
.....		
.....		
P. Bajos	26	21
Portugal	2	103
G. Bretaña	37	46
Japón	26	0

	(b) Factores 1	2
Agricultura	812	54
Industria	175	473
Servicios	13	472

Tabla 12
Contribuciones absolutas de los puntos fila (a)
y columna (b) a los 2 ejes

	(a) Factores 1	2
Argentina	85	915
Bolivia	765	235
Brasil	919	81
Chile	913	87
.....		
.....		
.....	.....	
P. Bajos	843	157
Portugal	83	917
G. Bretaña	774	81
Japón	998	2

	(b) Factores	
	1	2
Agricultura	985	15
Industria	613	387
Servicios	103	897

Tabla 13
Contribuciones relativas de los dos ejes a los puntos fila (a) y columna (b)

El que un punto tenga un contribución absoluta muy alta en un eje puede sugerir una posible interpretación de ese eje. Así, mirando la Tabla 12.b vemos que el primer eje está muy bien definido por Agricultura, con una contribución absoluta de 812 sobre 1000, mientras que el segundo queda definido por Industria y Servicio, con una oposición entre ambos sectores dadas las coordenadas opuestas de ambos puntos.

En el caso que nos ocupa las contribuciones relativas no vienen sino a confirmar algo que ya sabíamos a partir de los autovalores: que los puntos no aparecen deformados. En otros casos, en los que el tanto por ciento de inercia explicada por los ejes sea pequeño, las contribuciones relativas permitirán ver la bondad de la representación de los puntos en cada uno de los ejes que consideremos.

6.- Análisis de tablas con 3 o más variables

Una vez que hemos visto en el Apartado 1 el uso de los porcentajes en situaciones en las que tenemos 2 variables veamos ahora su extensión a problemas con 3 o más variables. En otro lugar (Sánchez Carrión, 1989 y 1988b) explicamos detenidamente este problema: aquí sólo vamos a hacer una introducción que permita comprender el posible interés del tema.

Lo primero que tenemos que explicar es la pertinencia de añadir nuevas variables a la situación bivariada. Los beneficios son múltiples (ver Sánchez Carrión, 1989); aquí sólo vamos a elegir aquél que tiene que ver con el hecho de que al introducir una nueva variable podemos conocer mejor la relación existente entre otras dos.

Supongamos que tenemos información sobre los Estudios, los Ingresos y el Voto de una muestra de cabezas de familia. La Tabla 14 muestra la relación entre estas 3 variables.

Estudios	Ingresos	Voto PSOE	Otros	Total
inferiores	altos	2	3	5
	medios	43	25	68
	bajos	184	80	264
medios	altos	7	3	10
	medios	42	30	72
	bajos	37	27	64
superiores	altos	12	14	26
	medios	30	31	61
	bajos	12	10	22
Total		369	223	592

Tabla 14
Cruce de las variedades Estudios, Ingresos y Voto

A partir de estos datos, en la Tabla 15 podemos ver que hay un mayor porcentaje de votantes del PSOE entre los individuos con estudios inferiores que entre aquéllos que tienen estudios superiores: un 13.0% más (es decir, 67.9-54.9).

Estudios	Voto PSOE	Otros	Total		
inferiores	.679	.321	1.000	(337)	*
med-superio	.549	.451	1.000	(255)	
Total	.623	.377	1.000	(592)	

*El valor 337 se obtiene como resultado de sumar 5,68 y 264 en la Tabla 14.

Tabla 15
Cruce de estudios y voto

Parece que una pregunta lógica es preguntarse cuál es la razón de esta relación entre los Estudios y el Voto. Una explicación plausible consiste en atribuir el menor voto al PSOE de los individuos con estudios superiores a sus mayores ingresos, y no a los estudios en sí mismos. Es decir, si los entrevistados con estudios superiores votan al PSOE, ello no es debido a razones de tipo cultural-académico, sino a motivos económicos.

Con el fin de comprobar nuestra suposición podemos recurrir a los diferentes métodos que se utilizan en la investigación social. El mejor de todos, siempre y cuando sea factible, es realizar un experimento. Bastaría dar los mismos ingresos a todos los trabajadores, independientemente de sus estudios, para observar después qué ocurre con su voto.

Si este método no es viable podemos recurrir a hacer un pseudo-experiento en el que estadísticamente se ajusten los datos con el fin de hacer que todos los trabajadores tengan los mismos ingresos, para comprobar posteriormente qué le ocurre a la relación entre Estudios y Voto. Un par de instrucciones en el programa CHIP (Bogart y Conner, 1986) nos permite realizar este ajuste y obtener la Tabla 16, en la que se muestra la relación entre Estudios y Voto en el supuesto de que no hubiera relación entre Estudios e Ingresos - dicho de otra manera, en el supuesto de que tanto los trabajadores con estudios inferiores como los que tienen estudios superiores ganasen lo mismo.

	Voto			
Estudios	PSOE	Otros	Total	
inferiores	.664	.336	1.000	(337)
med-superio	.557	.443	1.000	(255)
Total	.623	.377	1.000	(592)

Tabla 16
Cruce de las variables Estudios y Voto (datos estandarizados)

Vamos a comprobar los datos de las Tablas 15 y 16: en los datos estandarizados el porcentaje de votantes al PSOE entre los individuos con estudios inferiores es inferior (66.4% frente a 67.9%). Es decir, si las personas con estudios inferiores tuvieran los mismos ingresos que el resto, su voto al PSOE disminuiría. En el caso de aquéllos que tienen estudios superiores, el efecto de igualar sus ingresos con el de las restantes personas aumentaría su voto al PSOE (55.7% frente a 54.9%).

Puesto que mirar los números de ambas tablas puede ser confuso vamos a presentar los mismos resultados en forma gráfica. En la Figura 4, que se expone en la página siguiente, ofrecemos los resultados conjuntos de las Tablas 15 y 16. En vertical construimos unos postes en los que se refleja el tanto por ciento de votantes al PSOE para cada categoría de Estudios, y ello con los datos originales y con los estandarizados.

Tal como muestra esta figura, si todos los individuos tuvieran los mismos ingresos la diferencia de voto al PSOE entre los que tienen estudios inferiores y los que tienen estudios superiores se reduciría algo (pasaría de 13.1% a 10.7%). Pero aún con los mismos ingresos, el comportamiento político de ambos colectivos seguiría siendo diferente.

La conclusión sociológica que se saca de estos datos es que los menores Ingresos de los individuos con estudios inferiores explican un poquito (la diferencia entre 13.1 y 10.7) su mayor preferencia por el PSOE. Sin embargo hay algo en los Estudios, independientemente de que faciliten ganar más dinero, que es lo que en mayor medida explica esta preferencia política.

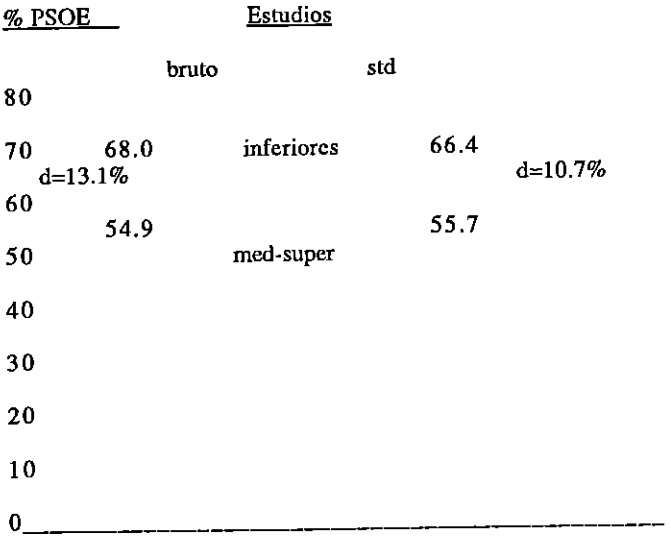

Figura 4
Influencia de los Estudios sobre el Voto.
(Datos brutos y estandarizados)

7.- Conclusiones

En las páginas precedentes hemos mostrado la aplicación de una serie de técnicas al análisis de datos nominales. Todas ellas parten de la tabla de contingencias, a partir de la cual realizan diferentes manipulaciones. En unos casos resumen las frecuencias de las tablas utilizando algún estadístico (diferencias de porcentajes o razones) y en otros representan gráficamente la información contenida en la tabla ("postes" y análisis de correspondencias). También hemos mostrado un procedimiento de análisis que sustituye las tablas por la creación de un fichero de datos agregados.

Una característica común de todas las técnicas que hemos introducido es su interés por mostrar las relaciones entre las categorías antes que las relaciones entre las variables, a las que pertenecen esas mismas categorías. Tal como hemos intentado explicar, a la hora de analizar variables nominales lo importante son las categorías y no las variables.

BIBLIOGRAFIA

BENZECRI, J.P. (1979): **L'analyse des donnés.** Dunod. París

BISHOP, Y.M.; FIENBERG, S.E. y HOLLAND, P.W. (1975): **Discrete Multivariate Analysis.** Mass: MIT Press. Cambridge.

BOGART, R. y C. CONNER, C. (1986): **CHIP.** N.H.: TrueBASIC Inc. Hannover.

DAVIS, J.A. (1976): *Analyzyng contingency tables with linear flow graphs: D. Systems.* En D.R. Heise (ed), **Sociological Methodology.** Joseey Bass. San Francisco.

DAVIS, J.A. (1987): **Social differences in contemporary America.** Harcourt Brace Jovanovich, Inc. Nueva York.

GARCIA SANTESMASES, J. (1984): *Análisis factorial de correspondencias.* En J.J. Sánchez Carrión (ed), **Introducción a las técnicas de análisis multivariable.** Centro de Investigaciones Sociológicas (CIS). Madrid.

NORUSIS, M.J. (1986): **SPSS/PC+.** Chicago III.: SPSS Inc.

SANCHEZ CARRION J.J. (ed) (1984): **Introducción a las técnicas de Análisis Multivariable aplicadas a las Ciencias Sociales.** Centro de Investigaciones Sociológicas (CIS). Madrid.

SANCHEZ CARRION, J.J. (1988a): **Análisis de datos con SPSS/PC+.** Alianza Universidad Textos, Madrid.

SANCHEZ CARRION, J.J. (1988): *Extending Rosenberg's idea about conjoint effects* Quantity 22: 49-64.

SANCHEZ CARRION, J.J. (1989): **Análisis de tablas de contingencia: el uso de los porcaentajes en las Ciencias Sociales.** Centro de Investigaciones Sociológicas (en prensa), Madrid.

PROGRAMAS INFORMATICOS

-CHIP (Análisis de Tablas de Contingencia: sistemas de las Diferencias de Proporciones).

Ruth Bogart y Chip Conner
True BASIC Inc.
39 S. Main Steet.
EE.UU.

-ECTA (Análisis de Tablas de Contingencia: modelos Log-linear).

Leo A. Goodman
Dpt. of Sociology
University of Chicago
1126 East 59th Street
Chicaggo III., 60637
EE.UU.

-TRI-DEUX (Análisis de Correspondencias)

Ph. Cibois
LISH
54 Bd. Raspail
755006 Paris
Francia

ESCUELA UNIVERSITARIA POLITECNICA Biblioteca CACERES

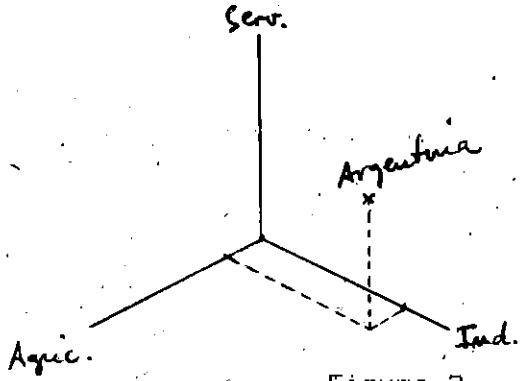

Figura 2
Nube de puntos de los paises