Las nuevas tecnologías y la investigación educativa. El análisis de datos de variables categoriales

por Luis LIZASOAIN y Luis JOARISTI
Universidad del País Vasco-Euskal Herriko Unibertsitatea

Introducción

En el contexto general de un número especial monográfico dedicado a las aportaciones de las nuevas tecnologías a la investigación educativa, conviene en primer lugar centrar y delimitar el objeto de nuestro trabajo. Y éste se centra en el análisis estadístico de datos con variables categoriales, lo que en la terminología anglosajona se identifica habitualmente con el término *Categorical Data Analysis (CDA;* de ahora en adelante usaremos el acrónimo inglés por ser de uso más frecuente*).*

Una vez realizada esta afirmación, quedan aún cuestiones por aclarar. La primera y fundamental es qué se entiende por variables categoriales. Una posible respuesta es limitar este término a aquellos fenómenos cuyo nivel de medida sea estrictamente nominal. Pero junto a éste, la mayoría de los autores incluyen bajo este término las variables ordinales. Por ejemplo, Agresti (2007) afirma que las variables categoriales tienen dos tipos de niveles de medida: el nominal y el ordinal (cfr. página 2). Y en la misma línea apunta Friendly (2000) en su manual sobre visualización de datos categoriales. Dicho de otra manera, las variables categoriales serían las no cuantitativas, aquellos fenómenos que carecen de una métrica fuerte.

Pero las cuestiones de demarcación tampoco acaban aquí. Si asumimos un esquema clásico de dependencia, podemos distinguir entre variables explicativas y variables de respuesta, entre variables independientes y dependientes. En la mayoría de los manuales estadísticos de CDA se incluyen aquellas técnicas en las que la variable de respuesta sea categorial, pudiendo ser las explicativas de cualquier nivel de medida.

No creemos que el foco principal de este trabajo deba centrarse en discutir estas cuestiones en profundidad. Por simpli-

cidad y por razones de espacio, vamos a prestar una especial atención a aquellas técnicas de análisis de datos en que las variables cualitativas juegan un papel básico o preponderante empezando por aquellas en las que todas las variables son cualitativas.

No vamos a insistir tampoco en la importancia que este tipo de técnicas tienen en la investigación educativa y en las ciencias sociales y humanas en general, habida cuenta de que en nuestro campo de investigación son muchos –y relevantes– los fenómenos que carecen de una métrica fuerte y que sólo pueden ser incorporados a la investigación como variables nominales o en su caso ordinales.

La primera conclusión que queremos avanzar ya es que, dado el actual nivel de desarrollo de las técnicas estadísticas, no es preciso *forzar* el nivel de medida de un fenómeno considerándolo como cuantitativo por razones de tratamiento estadístico. El desarrollo de la Estadística, con el imprescindible soporte de los procedimientos e instrumentos de cómputo, pone a disposición del investigador un conjunto de técnicas y procedimientos estadísticos capaces de incorporar y operar con cualquier tipo de variables y con casi cualquier combinación de las mismas.

Y ya que acabamos de referirnos al soporte de cálculo y computación, bueno será que finalicemos esta introducción con una breve síntesis de dónde radican las principales aportaciones de las nuevas tecnologías al CDA, aunque quizá sería más preciso decir al análisis estadístico de datos en general, puesto que estas cuestiones son, en la mayoría de los casos, independientes del nivel de medida de las variables.

No se debe olvidar que la fundamentación matemática de muchas de las técnicas estadísticas que se van a abordar está desarrollada desde hace mucho tiempo, en muchos casos desde antes de la aparición de las tecnologías de cálculo. Lo que éstas han posibilitado ha sido su desarrollo y aplicación.

Pero si esto es cierto, no lo es menos que las tecnologías informáticas también han contribuido al desarrollo de nuevas técnicas y procedimientos y esto es así hasta el punto de que autores como Lindsay, Kettenring y Siegmund (2004) en su informe sobre el futuro de la Estadística afirman que "la actividad central de los estadísticos es la construcción de herramientas matemáticas, conceptuales y *computacionales* que puedan ser usadas para la extracción de información"; y más adelante afirman que la ciencia de la Estadística tiene sus raíces en la probabilidad y la Matemática junto con la más reciente influencia de la *ciencia informática*. (Cfr. pp. 392 y 405. El subrayado y la traducción son nuestros).

Y entre las aportaciones y desarrollos más sobresalientes que la tecnología informática conlleva señalemos los siguientes.

En primer lugar, el incremento exponencial de las capacidades de los sistemas informáticos permite procesar cada vez más información, lo que lleva aparejada la posibilidad de analizar mayor cantidad de datos y de realizar cálculos más complejos.

Ejemplos evidentes de todo esto lo constituyen ámbitos como el de la minería de datos orientado al análisis y extracción de información relevante en grandes bancos de datos.

Pero el desarrollo de las tecnologías informáticas no ha sido sólo una cuestión de potencia, de *fuerza bruta*. Los primeros ordenadores y los primeros programas ya llevaban a cabo operaciones y cálculos complejos. Recordemos, por ejemplo, que la primera versión del popular SPSS data de 1970 (Nie, Bent y Hull, 1970).

Pero junto a esto, la aparición y generalización en la década de los 80 de los ordenadores personales supuso también un cambio importante. Su capacidad de almacenamiento y proceso no ha parado de crecer, pero sobre todo, su uso trajo aparejada la interactividad. Los grandes sistemas informáticos eran de acceso complicado y restringido, y habitualmente los programas se ejecutaban por lotes (*batch*) de forma que el usuario enviaba su programa, éste se ejecutaba y posteriormente recibía los listados con los resultados. A la vista de los mismos podía elaborar un nuevo programa y reiniciar la secuencia. Era un procedimiento de trabajo habitualmente lento, farragoso y sin supervisión directa por parte del usuario, procedimiento que el modo interactivo de operar de los programas diseñados para ordenadores personales cambió drásticamente.

No vamos a extendernos sobre la operación interactiva (habitualmente mediante interfaces gráficos de usuario) de los recursos informáticos porque estamos ya tan acostumbrados a la misma que casi nos pasa inadvertida. Pero desde el punto de vista del análisis estadístico de datos es innegable la flexibilización que lleva aparejada y, sobre todo, las vías que ha posibilitado para el enfoque exploratorio de los datos en la línea que Tukey apuntó.

Y continuando con los planteamientos de Tukey (1977), las posibilidades gráficas de los actuales sistemas informáticos han contribuido al desarrollo de lo que es ahora una rama de la Estadística, a saber, la visualización de datos cuyo objetivo es presentar de forma gráfica la información más relevante contenida en datos y resultados. Más adelante veremos esta cuestión con algo más de detalle, pero no está de más que señalemos la importancia que este tipo de recursos tienen para la presentación de resultados a audiencias no expertas, cuestión ésta muy relevante en ámbitos como por ejemplo la evaluación de programas.

Otras cuestiones muy importantes como las comunicaciones o las redes informáticas ya son tratadas en otros trabajos de este número, por lo que no insistimos aquí en ellas; pero para finalizar, no queremos dejar de señalar una última cuestión que pensamos tiene su importancia, como es el auge y desarrollo de los programas informáticos de código abierto (*free and open software*).

Se trata de una corriente importante dentro de la comunidad de usuarios y desarrolladores de *software* consistente en desarrollar programas y aplicaciones que se pueden leer, modificar y distribuir sin apenas restricciones. Esto origina que el código fuente de un programa evoluciona, mejora, corrige errores y se adapta a las

nuevas necesidades de forma rápida y sin depender de las políticas comerciales de las empresas.

En nuestro caso, el ejemplo más notable de esto lo tenemos en el programa R de análisis estadístico de datos. En este momento no hay técnica o desarrollo estadístico que no lleve aparejado su *paquete* de R para poder ser utilizada. Como prueba de lo dicho, en el momento de escribir estas líneas (mayo de 2011) el repositorio de paquetes de R tiene disponibles 3049 *(http://www.r-project.org/).* Igualmente, si se entra en una revista como el *Journal of Statistical Software (http://www.jstatsoft.org/)* y se recorren los diferentes números, se puede ver la gran cantidad de trabajos centrados en R. Por ejemplo, el volumen 40 correspondiente a abril de 2011 consta de 14 artículos de los cuales 11 se refieren a paquetes de R.

Tomando en cuenta estas consideraciones, a continuación vamos a presentar de forma muy sucinta algunas de las técnicas estadísticas más empleadas con datos categoriales teniendo muy presente que, en nuestra opinión, la más relevante aportación del desarrollo conjunto de la tecnología informática y de la Estadística estriba en que posibilita que nuestro acercamiento a la realidad educativa que investigamos pueda llevarse a cabo tomando en consideración las principales notas distintivas de la misma, como son su complejidad y su carácter multivariado y multinivel.

Y para analizar los datos de una realidad compleja, multivariada y multinivel, los procedimientos más adecuados son los de modelización estadística a sabiendas de que, como afirman Little, Schnabel y Baumert (2000), los resultados obtenidos mediante estas técnicas adolecen de una cierta ambigüedad y relatividad. Esta *ambigüedad*, esta necesidad de formular y validar modelos surge del hecho de emplear diseños no experimentales; en palabras de los citados autores: las técnicas inferenciales clásicas tienen poca o ninguna ambigüedad en el modelo cuando se aplican a datos obtenidos en el contexto de diseños experimentales completamente aleatorizados. Así por ejemplo, la adecuación del modelo implícito subyacente al análisis de datos está determinada por el grado en que correctamente refleja el diseño experimental, siendo la única fuente de incertidumbre las suposiciones sobre la distribución de variables o variables en la población.

Pero, como sabemos, nuestras investigaciones raramente pueden llevarse a cabo en dichas condiciones, por lo que las técnicas de modelización estadística se nos ofrecen como procedimientos viables a sabiendas de que el precio que hemos de pagar sea esa cierta ambigüedad. Como afirman Cook y Campbell (1979) "el cambio del estricto contraste de hipótesis a la lógica de la validación de modelos no se realiza porque se empleen procedimientos de modelos de ecuaciones estructurales, sino porque se usan datos cuasi-experimentales".

En cualquier caso, y para una visión completa y detallada de los procedimientos y técnicas estadísticos para variables categoriales recomendamos el ya citado manual de Agresti (2007) que es un referente en la materia. Un excelente análisis

relativo a la evolución de estas técnicas se encuentra en el trabajo de Goodman (2007).

Análisis de frecuencia de configuraciones

Esta técnica estadística, más conocida como *Configural Frequency Analysis (CFA)*, fue inicialmente diseñada y propuesta por Lienert (1968) siendo Von Eye el autor de referencia que ha desarrollado la misma (von Eye, 2002; von Eye y Gutiérrez Peña, 2004). Tiene como objetivo el análisis de tablas de contingencia bi o multidimensionales generadas por la clasificación cruzada de variables categoriales.

A diferencia de otras técnicas como los modelos logarítmico lineales, que se centran en el examen y análisis de las relaciones entre las variables, aquí el enfoque se centra en los casos, en los individuos. Su objetivo es identificar grupos de individuos, configuraciones, que se puedan considerar –desde un punto de vista estadístico– *especiales*.

Y tal consideración se refiere al hecho de que sean casillas de la tabla de contingencia que no se ajusten a lo esperado, que sus frecuencias observadas sean significativamente distintas de las frecuencias estimadas esperadas. Por tanto, nos encontramos ante una técnica de clasificación de casos, pero, a diferencia del análisis de conglomerados o el de clase latente, que crean grupos desconocidos de casos a partir de los datos brutos, el CFA se centra en analizar si los grupos existentes en la clasificación cruzada contienen más o menos casos que los esperables.

Se trata de una herramienta fundamentalmente exploratoria cuyos resultados se consideran de fácil interpretación, pues se centra en analizar si un determinado patrón de categorías de las variables (configuraciones) se da más veces de lo esperado, menos o tanto como cabía esperar. Una configuración que contenga más casos que los esperados es denominada un *tipo* mientras que la que contiene menos casos de los esperados constituye un *antitipo*.

Un CFA se desarrolla en 4 etapas o fases (Von Eye, 2010): la primera concierne a la elección de un modelo base, la segunda se centra en la estimación de las frecuencias esperadas de cada casilla, la tercera es la selección y realización de la prueba estadística y la última es la identificación e interpretación de los tipos y antitipos. Veamos muy brevemente cada una de ellas.

El modelo base consta de todas las relaciones entre variables que no son objeto de hipótesis sustantivas, o dicho de otra manera, el modelo base está formado por todos los términos que *no* son de interés para el investigador. Tomando en consideración este modelo base, se estiman las frecuencias esperadas habitualmente basándose en modelos logarítmico lineales. Una vez estimadas las frecuencias esperadas se comparan con las empíricas y se realiza un contraste estadístico. Aquellas casillas, aquellas configuraciones cuyas frecuencias observadas sean significativamente mayores que las esperadas conforman un tipo y, simétricamente, aquellas donde la diferencia sea significativamente menor constituyen los antitipos.

El supuesto básico es que si emergen tipos y antitipos, estas configuraciones reflejan o representan los efectos que *sí* resultan de interés. El proceso finaliza interpretando estas configuraciones a la luz del papel que juegan las variables (por ejemplo explicativas vs. criterio), las características del modelo base y la teoría sustantiva.

Lógicamente en un enfoque como el presentado es crucial la elección del modelo base. Von Eye (2004) describe y analiza pormenorizadamente diferentes grupos de modelos base.

Aunque estas técnicas son empleadas en la investigación en Psicología o en Medicina, no lo son tanto en nuestro campo. El trabajo de Wagner, Schober y Spiel (2008) ilustra el empleo de esta técnica en una investigación educativa reciente. En el mismo emplean CFA para analizar el efecto del género de los estudiantes en la relación entre el tiempo dedicado a la realización de tareas escolares y el rendimiento académico. Como suele ser habitual en este tipo de estudios, variables como el rendimiento o el tiempo son dicotomizadas en función de la mediana. En este caso de las 8 posibles configuraciones, sólo emerge un tipo que muestra que son más frecuentes las chicas que obtienen altas puntuaciones y que dedican mayor cantidad de tiempo al estudio. También en nuestro campo, y más específicamente en el de la orientación, Reitzle y Vondracek (2000) emplean CFA y análisis de correspondencias para estudiar las transiciones familiares y laborales de jóvenes alemanes resaltando las diferencias entre enfoques analíticos centrados en las variables y los centrados en las personas.

Para un análisis detallado de las posibilidades de esta técnica, de sus diversos campos y modos de aplicación (exploratorio, confirmatorio, predictivo o longitudinal), remitimos al libro recientemente publicado por Von Eye, Mair y Mum (2010) en el que abordan los últimos avances en CFA. Por último, el paquete de R que ejecuta este tipo de análisis se denomina *cfa* y se encuentra en *http://cran.r-project.org/web/packages/cfa/*.

Tablas de contingencia tridimensionales

El problema más sencillo con variables cualitativas y que presenta una estructura no simple se plantea como una tabla de contingencia bidimensional. Su complejidad es mínima desde el punto de vista inferencial, centrándose la complejidad en el estudio de la independencia entre dos factores bajo los diversos diseños de muestreo, obteniendo estimadores máximo verosímiles de la distribución empírica, la partición de tablas, las medidas de asociación, tanto simétricas como asimétricas, basadas en las razones de ventajas ordinarias (odds ratios, OR), locales y globales.

Sin embargo la complejidad crece considerablemente si se trata del análisis de una tabla tridimensional, complejidad por otra parte más acorde a la realidad educativa; surgen conceptos sobre los distintos tipos de independencia, así como las razones de ventajas, que desembocan en probar una numerosa serie de hipótesis sobre independencia global, conjunta entre dos factores, mutua entre tres factores, condicional, parcial y marginal, así como de ausencia de interacción y las relaciones entre ellas. El esquema más asequible se

refiere a tablas 2 x 2 x K que representan la relación entre dos factores dicotómicos controlando un tercer factor no dicotómico. Una vez especificada una medida de la asociación parcial entre los factores, se prueba la existencia de interacción o entre los factores; en su caso se plantea la hipótesis de homogeneidad de la medida de asociación entre niveles y aplicar por fin una prueba de no asociación entre factores. Para ello se utilizan como medidas de asociación las OR, que en este caso son las ordinarias.

Por su parte, en las tablas tridimensionales de factores no dicotómicos las OR a utilizar son las locales, lo que enmaraña considerablemente resultados e interpretaciones. Si además, para tablas de más de tres dimensiones, el número de estructuras jerárquicas (formas de independencia y asociación) se incrementa tanto que resulta muy trabajoso analizarlas individualmente mediante contrastes de Chi cuadrado de independencia, la generalización del análisis de tres a cuatro dimensiones, por poco práctica, dirige el enfoque metodológico a los modelos Log-lineales.

Modelos log-lineales

Siguiendo a Ruiz-Maya (1995), el análisis clásico de las tablas de contingencia no resuelve cuestiones como la estimación del peso sobre las frecuencias de cada factor a través de sus niveles, ni la influencia conjunta de varios factores. Además, por otras razones más arriba mencionadas, se recurre a los modelos Log-lineales, ya que tienen una interpretación natural en términos de probabilidad y de independencia condicional (Aguilera, 2006)

Se busca expresar en términos aditivos, por transformación logarítmica de los términos multiplicativos, la frecuencia de los casos en las combinaciones de niveles o modalidades de los distintos factores según las distribuciones marginales y las interacciones.

Correa (2002) plantea las siguientes etapas en la construcción de un modelo: proponerlo, derivar un conjunto de expectativas bajo él, someter a prueba el ajuste del modelo (Chi cuadrado o la razón de verosimilitudes de Chi cuadrado), decidir si mantener el modelo o no y por fin, estimar los parámetros del modelo definitivo.

Por lo tanto la primera etapa consiste en ajustar modelos Log-lineales jerárquicos a una tabla de contingencia multidimensional para encontrar las variables categoriales o factores que están asociados. Este proceso, que puede ser automático, se realiza habitualmente por pasos.

Se empieza por el modelo saturado. Hay que recordar que el modelo nulo es aquel que no tiene efectos y el saturado el que tiene tantos efectos como celdas tiene el hipercubo de contingencia (Lebart, 1985), lo que implica que no hay ninguna pérdida de información. Entre ambos extremos se encuentra el óptimo según el postulado de parsimonia. Inicialmente se elimina la interacción de orden más alto. Cada vez que se elimina un efecto, se evalúa el cambio del estadístico Chi cuadrado, de manera que si no es significativo, tal efecto se puede eliminar del modelo. En cada paso se elimina el efecto cuyo cambio en la razón de verosimilitudes tiene mayor nivel de significación. Se continúa así

hasta llegar a un modelo en que todos los efectos son significativos. Al tratarse de modelos jerárquicos, hay que tener presente que un modelo con algún efecto de interacción incluirá también a los efectos simples que lo componen y a las interacciones de orden inferior.

Siguiendo el proceso, la hipótesis sobre la nulidad de ciertos efectos ha generado un modelo; a partir de él se estiman las frecuencias teóricas (por el método de máxima verosimilitud y suponiendo que los datos siguen una distribución multinomial o una de Poisson, según que el tamaño de la muestra esté o no limitado previamente) utilizando el método iterativo de Newton-Raphson, para compararlas a las empíricas a través de pruebas de bondad del ajuste y así rechazar o no la hipótesis sobre las asociaciones incluidas en el modelo. Las pruebas más utilizadas son la de Chi cuadrado de Pearson y la G cuadrado de la razón de verosimilitudes, resultando que cuanto más se aproximan a 0 mejor es el ajuste. Así, una estrategia para obtener un modelo adecuado consiste en partir del saturado e ir eliminando sucesivamente las interacciones, haciendo que los estadísticos mencionados aumenten, pudiendo detener el proceso cuando se produzca un salto brusco. Es posible utilizar los criterios AIC y BIC para la selección del modelo.

Conviene dejar claro que no conviene aplicar esta técnica con más de tres factores, pues téngase en cuenta que con 4 factores hay 167 modelos jerárquicos posibles (Lebart, 1985)

En SPSS, el Análisis *Log-lineal* se encuentra en el submenú *Loglineal* con sus tres posibilidades: *General, Logit y Selección de modelo.*

En R se pueden usar varias funciones para ajustar modelos *Log-lineales.* En la biblioteca "*stats*" (que se carga por defecto) está la función *loglin* y *glm*. En el paquete "*MASS*" las funciones *loglm* y *loglin* ajustan modelos jerárquicos *Log-lineales.* Como tutorial, proponemos el de William B. King que está accesible en la web *http://ww2.coastal.edu/kingw/statistics/R-tutorials/loglin.html.*

Modelos *logit, probit* y análisis de regresión logística

A diferencia de lo que se ha expuesto hasta aquí, introduciremos una variable dependiente y una o más independientes. Por otro lado hay que tener presente la inviabilidad del modelo lineal de probabilidad, pues puede dar resultados con probabilidades estimadas fuera del intervalo [0, 1]. Sin embargo, ésta y otras condiciones imprescindibles son cumplidas por dos tipos de funciones, la logística y la normal. A partir de estas funciones de enlace, surgen los modelos denominados *Logit* y *Probit,* pertenecientes ambos a los denominados Modelos lineales generalizados (GLMM).

En general se aborda este tema distinguiendo claramente entre la respuesta binaria y la multinomial, pues al basarse en las OR de la variable dependiente hay que recordar la complejidad de las OR locales, asociadas sólo a las variables politómicas. Y como el modelo *Logit* es una reformulación del *Log-lineal,* lo relativo a la bondad de ajuste y estimación de los parámetros se realiza empleando las herra-

mientas estadísticas del modelo *Log-lineal.*

En su estructura, la Regresión logística es semejante a los modelos *Logit* (tanto que es frecuente no distinguir entre ellos, dada la asociación del modelo *Logit* a la Regresión logística) y en sus objetivos, al Análisis discriminante, siendo menos restrictivo que éste. Su principal aplicación se centra en variables dependientes dicotómicas y en estimar la probabilidad de que se produzca el suceso definido por ellas (Silva, 2004), aunque recientemente se desarrollan cada vez más modelos más generales y complejos, como los de variable de respuesta politómica y ordinal.

La regresión logística en R es realizable mediante la función *glm* que está incorporada en el módulo básico. El tutorial de Manning (2007) aporta información detallada al respecto.

Análisis factorial de correspondencias

Las exposiciones y desarrollos sobre análisis factorial en general, muy numerosas, difieren entre sí en aspectos meramente matemáticos y didácticos, siendo frecuentes textos en los que se prima la comprensibilidad sobre el rigor. Autores como Benzécri (1984), Volle (1997), Bertier (1981), etc. apuestan por la profundización matemática para conseguir una considerable competencia en la aplicación práctica (Joaristi, 1999)

Son técnicas del grupo de las de interdependencia y su objetivo son las tablas de contingencia. A diferencia del análisis *Loglineal,* es preponderante el enfoque exploratorio. En el caso de sólo dos dimensiones, hablamos del AFC simples. Si fuesen más de dos dimensiones, se trataría del AFC múltiples.

Centrándonos en las correspondencias simples, definir una correspondencia entre dos conjuntos consiste en asociar a cada elemento del producto cartesiano de ambos conjuntos un número no negativo. Lo más clásico es tratar una tabla de contingencia, aunque puede ser aplicado a cualquier tabla de números positivos. Si todos los valores son enteros se trata de una correspondencia estadística, pues los números indican cuántas veces se presenta cada par; estas son las tablas de contingencia. Sin embargo, si son probabilidades, se trata de una correspondencia probabilística. También se pueden analizar tablas de medidas, tablas de intensidades con notas de mérito, tablas de preferencias y tablas de Burt (Lizasoain, 1999).

Entre los problemas que más habitualmente se plantean está el de descomponer una tabla de grandes dimensiones, es decir, que sea inasequible a la inspección visual, en otras "fáciles de entender". Se puede representar la información fundamental de una tabla grande en unas pocas dimensiones, dando lugar a unas pocas representaciones gráficas planas (del tipo de un diagrama de dispersión), que son más asequibles e ilustrativas. Otra aplicación consiste en transformar una información sin métrica en otra bajo factores ortogonales y con métrica euclídea con el fin, entre otros, de realizar una clasificación de los sujetos.

En este tipo de técnicas multivariantes de análisis de datos hay que caracterizar a

los elementos (filas y columnas) por sus coordenadas en el espacio correspondiente, además es preciso tener presente su masa y por fin hay que definir una distancia entre ellos; el problema matemático consiste habitualmente en la optimización de una función objetivo. En el caso que nos ocupa, la ponderación de un elemento (fila o columna) se realiza por las marginales correspondientes. En cuanto a la distancia, se trata de Chi-cuadrado, la cual posee la importante propiedad de la equivalencia distribucional, consistente en que la distancia entre filas no se altera si se fusionan dos columnas de perfil semejante. Esta propiedad es deseable frente a arbitrariedades de la codificación, garantizando así la robustez, pues ni se gana información descomponiendo una clase en subclases homogéneas, ni se pierde fusionando clases homogéneas en otra. El objetivo matemático es encontrar los ejes principales, o de máxima inercia o varianza de las nubes de puntos de las filas y de las columnas.

Los elementos utilizados para calcular los planos factoriales se denominan elementos activos y deben formar un conjunto homogéneo y exhaustivo (describir completamente el tema) para que las distancias entre elementos puedan ser fácilmente interpretables. Así, se analizan como elementos suplementarios o ilustrativos las observaciones recogidas bajo condiciones poco claras o distintas de las del resto, o bien, elementos aberrantes, o casos nuevos, o elementos de distinta naturaleza del resto; asimismo se tratarían como suplementarios los elementos recogidos con posterioridad a la realización del análisis. Además la dicotomía entre elementos activos e ilustrativos es tan fundamental como la distinción que se establece entre variables endógenas (a explicar) y exógenas (explicativas) en la regresión múltiple.

En cuanto a la interpretación de los resultados, una importante ventaja al tratar una tabla de números por AFC es la asociada a las relaciones de transición: el poder representar las coordenadas de una fila en función de las coordenadas de las columnas, lo que implica que los dos conjuntos se pueden representar gráficamente de forma simultánea.

Para representar adecuadamente la importancia de un elemento en la creación de un eje factorial se recurre a las contribuciones relativas. Además, con el fin de evaluar cómo están representadas una fila o una columna por los distintos factores, se recurre a las contribuciones relativas del factor sobre el elemento en cuestión; se entiende por calidad de la reconstitución de un elemento por medio de los primeros ejes como la suma de las contribuciones relativas de esos ejes sobre él. Para interpretar un factor es conveniente elegir un reducido número de modalidades cuya contribución a la inercia del factor sea fuerte. Para interpretar un factor conviene buscar los puntos para los que la contribución relativa de modalidad a factor es elevada.

Una vez resumidas las ideas fundamentales del AFC simples, poco queda por decir sobre el AFC múltiples, pues básicamente es análogo. Habitualmente los datos vienen bajo el formato que Diday (1982) denomina tabla de modalidades, en que las filas son casos y las columnas son las variables cualitativas; por lo tanto, una

diferencia sustancial con el AFC simples es que las filas no son modaliades de variables. Además no es ésta la tabla que se va a analizar, debido a que modificaciones –lícitas– en la codificación numérica de las modalidades, afectarían a los resultados. Es por lo que se recurre a la codificación disyuntiva completa de las modalidades en un procesamiento automático que se realiza de forma temporal.

Otra forma de analizar una tabla de contingencia multidimensional es a través de la matriz de Burt, que es una matriz simétrica que contiene todas las tablas de contingencia simple entre las variables y su principal ventaja es que es más "económica" desde el punto de vista del procesamiento informático. Su análisis no es de correspondencias múltiples sino el de la tabla binaria asociada a la correspondencia múltiple, siendo analizados los casos como filas suplementarias de la matriz de Burt.

En SPSS el AFC simples se realiza en el submenú Reducción de datos, en Análisis de correspondencias y el AFC múltiples en el mismo submenú en Escalamiento óptimo.

El análisis de correspondencias, tanto simple como múltiple, se puede ejecutar en R mediante los paquetes *ca* y *anacorr*. En la página web específica de R dedicada a los modelos y métodos psicométricos (*CRAN Task View: Psychometric Modles and Methods*) hay un apartado específico dedicado al análisis de correspondencias con información adicional muy detallada. Como trabajo de referencia recomendamos el artículo de Nenadic y Greenacre (2007).

Clasificación automática con variables categoriales

Todo ello, sin embargo, produce un volumen de resultados cuya interpretación, por no ser sencilla ni estandarizada, está sometida a un aspecto más subjetivo que lo habitual en las técnicas estadísticas de análisis de datos. Ante ello se recurre a realizar una clasificación; por medio del AFC se pueden obtener las puntuaciones factoriales, tanto de filas como de columnas, pudiendo así someter a ambos tipos de elementos a alguna de las técnicas de clasificación, aclarando así la disposición y agrupamientos de los elementos perceptibles en los distintos planos factoriales en clases homogéneas.

Pero al tratarse de técnicas de clasificación con variables cuantitativas, quedan al margen de este artículo. Abordaremos sin embargo las técnicas de clasificación que operan directamente sobre variables cualitativas.

Conocida bajo varios nombres (análisis de clusters, análisis de conglomerados, taxonomía numérica), constituye un conjunto de técnicas que se puede situar en el grupo de las de interdependencia. Es además eminentemente exploratoria. Se trata de superar una visualización plana y continua de las asociaciones estadísticas, para poner en evidencia clases de casos o clases de variables (Lebart, 1985). Se diferencia del análisis discriminante en que éste asigna los casos a grupos preexistentes.

Las técnicas requieren de cálculos por medio de algoritmos, son por tanto recursivos y repetitivos. Como consecuencia, es prácticamente imprescindible la utiliza-

ción de herramientas informáticas: "El cálculo electrónico es uno de los dominios de la ciencia aplicada en que el progreso es en la actualidad el más rápido y sorprendente...Tales problemas no se resolverán si no es mediante algoritmos de clasificación o de reducción del número de dimensiones", afirma Benzécri (1984). Evidentemente cuanto mayor sea la rapidez y la capacidad de proceso, problemas con mayor volumen de datos podrán ser tratados. Por otra parte, los desarrollos matemáticos son en general elementales, basándose en el sentido común más que en teorías formalizadas (Lebart, 1985).

Como ya se ha mencionado, en las técnicas multivariantes los elementos se caracterizan por sus coordenadas y su masa, definiendo también una distancia entre ellos. La situación aquí no es muy distinta. Partiendo de la posición de los elementos en el espacio correspondiente (los casos en el de las variables y viceversa), el proceso consiste en obtener una medida de la semejanza, similitud o disimilitud (que no hay que confundir con la distancia, pues ésta es más restrictiva) entre ellos y desarrollar algún procedimiento que construya las clases basándose en ellas.

Como ya ha sido apuntado, en lo que sigue vamos a limitarnos a las técnicas de clasificación con variables cualitativas. Ello va a condicionar tanto la elección de la métrica o distancia como el algoritmo de clasificación. En general, los algoritmos de clasificación pueden ser jerárquicos o no.

La clasificación jerárquica se basa en una familia de algoritmos calificables de deterministas" (Lebart, 2000), pero están mal adaptados a conjuntos de datos amplios. Un algoritmo de clasificación jerárquica consiste en transformar la disimilitud inicial para convertirla en una distancia ultramétrica y a continuación construir la jerarquía indexada (Hernández, 2001); esto es, a través de la ultramétrica se pueden ir agrupando las clases más próximas hasta llegar a una jerarquía indexada. El proceso para realizar una clasificación jerárquica ascendente se puede resumir en, partiendo de elementos sin agrupación alguna, buscar los 2 elementos más próximos y agregarlos en uno nuevo; a continuación se calcula la distancia entre el nuevo elemento (grupo) y el resto. Los métodos de agregación habituales son los del Vecino más próximo, Vecino más lejano, Vinculación inter-grupos, Vinculación intra-grupos, Agrupación de centroides, Agrupación de medianas y Método de Ward. El proceso iterativo consiste en repetir este proceso hasta que al final todos los elementos estén en un sólo grupo. Esto da origen a una jerarquía indexada que se representa en forma de árbol o dendrograma, representando en un eje los elementos ordenados según se van agregando y en el otro las distancias correspondientes a los distintos niveles de agregación.

Por su parte la clasificación en dos pasos o fases permite clasificar los casos, que pueden ser muy numerosos, basándose en variables tanto cuantitativas como cualitativas, utilizando para ello como medida para la agregación la verosimilitud. Se supone que las variables categoriales siguen leyes multinomiales y se parte de que todas las variables son independientes; sin embargo, el procedimiento es lo suficientemente robusto ante violaciones de estos

dos supuestos. Esto hace que este algoritmo sea preferible a las técnicas tradicionales. Además no hay que especificar un número concreto de clases. Siguiendo a Yu (2010), en el primer paso o de preclasificación se crea el árbol de características de los conglomerados, explorando cada elemento y aplicando la distancia de verosimilitud para determinar si el elemento debe agregarse con otros o formar una clase nueva; explorados todos los casos, las clases creadas son tratadas como nuevos datos. En el segundo paso se aplica la clasificación jerárquica a estas clases, obteniéndose una jerarquía indexada. Para seleccionar el número óptimo de clases se dispone de los criterios de información AIC de Akaike, que es un índice del ajuste de un modelo, que compromete la bondad de su ajuste con su complejidad. Ya que el índice AIC es optimista, es más conveniente aplicar el criterio BIC de Bayes, que penaliza la complejidad con más intensidad. La gráfica de la importancia de las variables en cada clase resulta una valiosa herramienta; en ella se muestra el valor de Chi-cuadrado o su significación como índice de la importancia de cada variable categórica en cada clase.

Se presenta de forma resumida la clasificación alrededor de centros móviles (K medias) ya que puede ser la primera etapa auxiliar de otros métodos de clasificación, pues este algoritmo es mucho más rápido y requiere menor capacidad de procesamiento informático y almacenamiento. Las distancias a utilizar son la euclídea y la Chi-cuadrado. Este algoritmo consiste en especificar una serie de centros provisionales de las clases, constituyéndose una primera partición de los elementos, asignando cada elemento a la clase de centro más próximo. Después se determinan los nuevos centros de las clases formadas en la etapa anterior, dando origen a una nueva partición. Repitiendo el proceso, se llega a la estabilidad de la clasificación, lo que da fin al proceso.

Respecto de la clasificación mixta, es adecuada para grandes volúmenes de datos. La primera fase consiste en realizar una clasificación no jerárquica, agrupando los elementos en torno a centros móviles en muchas clases (pero considerablemente menos que el número de elementos). En la fase siguiente se realiza una clasificación jerárquica ascendente a las clases obtenidas en el paso anterior. El proceso finaliza obteniendo la partición final por corte del árbol de la clasificación jerárquica ascendente (habitualmente cortando por las ramas más largas).

Programas informáticos con técnicas de clasificación para datos categoriales:

En SPSS desde la versión 11.5 hasta las actuales, se encuentran en el submenú Clasificar tres subprogramas: Conglomerados en dos fases y Conglomerados jerárquicos y Conglomerados de K medias.

Las técnicas de análisis de conglomerados son realizables en R mediante el paquete *cluster* que viene incorporado con el programa, de forma que cualquier instalación de R dispone del mismo sin que sea necesaria ninguna carga adicional. La función *dendrogram* permite la realización de este tipo de representación gráfica. Información adicional muy detallada se puede encontrar en la página específica que

R tiene sobre clasificación: http://cran.es.r-project.org/web/views/Cluster.html

Por último, como textos de ayuda recomendamos los tutoriales de Holland (2006) y de Oksanen (2010).

Visualización de datos categoriales

Aunque el uso de recursos gráficos es casi tan antiguo como la Estadística [1], el desarrollo de las tecnologías informáticas ha supuesto un enorme impulso en este campo permitiendo en primer lugar el diseño y realización de gráficos cada vez más complejos y completos; segundo, la generación interactiva de los mismos mediante la selección por parte del usuario de variables, parámetros y opciones; y tercero, la vinculación dinámica de los gráficos a las tablas y a los archivos de datos.

La visualización de datos se puede considerar una parte de la Estadística existiendo revistas científicas centradas en este campo como *Journal of Computational and Graphical Statistics* o *Information Visualization* siendo además muy numerosas las páginas web centradas en esta cuestión.

Este conjunto de técnicas tiene como objetivos complementar las habituales tablas y resúmenes estadísticos mediante recursos gráficos y visuales orientados a facilitar la exploración de los datos, así como el diseño y ajuste de modelos mostrando patrones en los datos.

Y todo ello con la doble finalidad de ser una herramienta que facilite tanto la presentación de datos y resultados, como el propio análisis de los mismos. Friendly (2000) en la introducción de su obra sobre visualización de datos categoriales afirma que el foco de la visualización de datos ha de estar en los gráficos intuitivos de forma que estos sean capaces de revelar aspectos de los datos que no pueden ser percibidos mediante otros medios (cfr. pág. 1).

En definitiva, la visualización trata de ser una herramienta para facilitar la comprensión de los datos. Este enfoque es asumido por la obra de Young, Valero-Mora y Friendly (2006) que se centra en el uso de gráficos dinámicos interactivos como herramienta del análisis estadístico de datos.

Aunque en sus inicios, la visualización estaba sobre todo centrada en variables cuantitativas, en la actualidad se dispone también de un amplio repertorio de gráficos (y de programas informáticos) para las variables categoriales y la citada obra de Friendly (2000) en un referente en la materia a la que remitimos al lector para un examen detallado y pormenorizado.

Ahora vamos a mostrar un ejemplo y algunas direcciones útiles tanto para el estudio de estas cuestiones como para localizar los recursos informáticos necesarios.

Normalmente los gráficos para variables categoriales se agrupan en aquellos centrados en la representación de distribuciones discretas, en los que su interés está en las tablas de contingencia bi o multidimensionales y en los asociados a técnicas o modelos específicos como la regresión logística, el análisis de correspondencias o los modelos logarítmicos-lineales.

Dado que los primeros se refieren a un enfoque univariado y que los terceros han de contemplarse junto con los procedimientos asociados, vamos a centrarnos en las tablas de contingencia. Y aquí los gráficos más habituales son los mosaicos, y para el caso de tablas de 2x2 los gráficos cuádruples (*fourfold displays*) y los diagramas que muestran los niveles de acuerdo o desacuerdo (diagrama de Bangdiwala). Dado que los mosaicos son de aplicación a tablas de cualquier dimensión y que se usan también con los modelos logarítmico lineales, vamos a ver sucintamente un ejemplo, remitiendo para el resto a la bibliografía y páginas web recomendadas.

Los gráficos de mosaico fueron planteados por Hartigan y Kleiner (1984) y por Friendly (1994). La idea central es que, dada una tabla de contingencia, se representan tantos rectángulos como casillas tenga tabla, siendo el área de los rectángulos proporcional a la frecuencia observada y la intensidad del sombreado a las frecuencias esperadas o a los residuos (de este planteamiento general hay diversas variaciones, pero éste es el modelo más habitual).

Por ejemplo, en la Tabla 1 se muestra la tabla de contingencia que cruza el género de 300 estudiantes con el tipo de carrera universitaria que cursan (Científico-Matemática, Ciencias Sociales, Humanidades).

TABLA 1: *Frecuencias de las variables género y tipo de carrera*

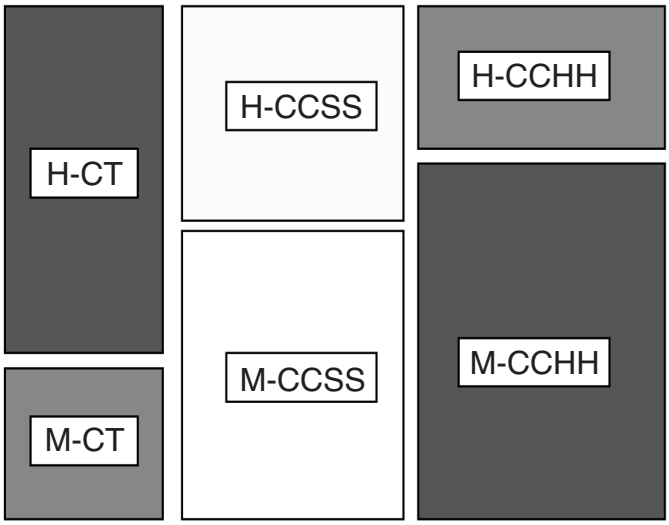
	Científico-Matemática (CT)	Ciencias Sociales (CCSS)	Humanidades (CCHH)	Suma
Hombres (H)	51	52	37	140
Mujeres (M)	21	61	78	160
Suma	72	113	115	300

La Figura 1 muestra el gráfico de mosaico correspondiente a dicha tabla. El lado horizontal de cada uno de los 6 rectángulos es proporcional al marginal por columnas y el vertical al marginal por filas. De esta forma, el ancho –por ejemplo– de los dos rectángulos de la primera columna es proporcional a la suma de estudiantes matriculados en carreras científico matemáticas con respecto al total (72/300). Y para cada columna el alto relativo de los dos rectángulos –por ejemplo de la tercera– lo es a la proporción de hombres y mujeres que cursan grados de ese tipo (humanida-

des: hombres 37/140; mujeres 78/160). Y para cada rectángulo, su área es proporcional a la frecuencia de cada casilla.

Pero además, la intensidad del color (aquí sombreado en blanco y negro que hace perder algunos matices) está en función de las frecuencias esperadas o, si se quiere, de los residuos. De esta forma, cuanto más intenso sea el color o el sombreado, más importante será el efecto de interacción. Y así vemos como en la columna central el sombreado es muy tenue, casi inexistente, denotando que los estudiantes matriculados en grados de ciencias sociales no difieren en su número con respecto a su sexo. Por el contrario, el rectángulo superior de la primera columna o el inferior de la tercera muestran un sombreado intenso pues la proporción de estudiantes masculinos en carreras científico-matemáticas es muy superior a la de estudiantes femeninas, dándose la relación contraria en el caso de las humanidades. En los gráficos en color, el signo de los residuos se representa empleando colores distintos y la magnitud mediante la intensidad de dicho color.

FIGURA 1: *Mosaico correspondiente a la Tabla 1*



Es muy difícil plasmar en un trabajo de estas características y en soporte papel las enormes potencialidades que tienen los actuales programas de visualización de datos, por lo que remitimos al lector a las páginas web que a continuación reseñamos y donde podrá encontrar una detallada galería de gráficos con multitud de efectos y posibilidades. Por ejemplo, el paquete *vcdExtra* de R permite desarrollar mosaicos en tres dimensiones para tablas de contingencia que crucen tres variables.

De entre la amplia variedad de sitios y páginas web centrados en visualización de datos, nos permitimos reseñar los siguientes:

La galería gráfica de R *(http://addictedtor.free.fr/graphiques/)* es un referente obligado donde aparecen todos los recursos gráficos de R. En el caso concreto de la visualización de datos categoriales los paquetes más habituales son *vcd y vcdExtra*:

• *(http://cran.r-project.org/web/packges/vcd/)*

• *(http://cran.r-project.org/web/packages/vcdExtra/)*

La página de Friendly sobre visualización de datos (http://www.datavis.ca/) a la que aludimos al inicio de este apartado, es una referencia en la materia. Contiene una excelente galería con útiles (y en ocasiones divertidos) ejemplos de buenas y malas gráficas, enlaces, material de sus cursos, etc. Por ejemplo, en el apartado de cursos se puede descargar un excelente tutorial escrito por Friendly (2010) sobre los paquetes de R *vcd* y *vcdExtra (http://www.datavis.ca/courses/VCD/vcd-tutorial.pdf).*

Por último, en la página web del profesor Valero Mora de la facultad de Psicología de la Universidad de Valencia *(http://www.uv.es/valerop/)* hay un apartado dedicado a la visualización de datos y otro al programa informático ViSta (Young y Bann, 1997; Valero-Mora, Young y Friendly, 2003) de visualización estadística que es de descarga libre y gratuita.

Dirección para la correspondencia: Luis Lizasoain. Departamento de Métodos de Investigación y Diagnóstico en Educación. Universidad del País Vasco, Avda. Tolosa 70, 20018 San Sebastián.
E-mail: luis.lizasoain@ehu.es.

Fecha de recepción de la versión definitiva de este artículo: 15.VI.2011

Notas

[1] Véase al respecto la página web de visualización mantenida por el profesor Friendly http://www.datavis.ca/milestones/ centrada en los hitos en la historia de los gráficos y la visualización, o el trabajo del propio Friendly (2009) titulado *Milestones in the history of thematic cartography, statistical graphics, and data visualization.*

Bibliografía

AGRESTI, A. (2007) *An Introduction to categorical data analysis* (New Jersey, John Wiley and Sons, 2nd Ed.).

AGUILERA, A. M. (2002) *Tablas de contingencia bidimensionales* (Madrid, La Muralla).

AGUILERA, A. M. (2006) *Modelización de tablas de contingencia multidimensionales* (Madrid, La Muralla).

BENZECRI, J. P. (1984) *L'Analyse des Données, Vol.1. La Taxinomie* (París, Dunod).

BERTIER, P. y BOUROCHE, J. M. (1981) *Analyse des données multidimensionnelles* (París, PUF).

COOK, T. D. y CAMPBELL, D. T. (1979) *Quasi-Experimentation: Design & Analysis Issues for field settings* (Boston, Houghton Mifflin Company).

CORREA, A. D. (2002) *Análisis logarítmico lineal* (Madrid, La Muralla).

DIDAY, E.; LEMAIRE, J.; POUGET, J. y TESTU, F. (1982) *Élements d'analyse de données* (París, Dunod).

FRIENDLY, M. (1994) Mosaic displays for multi-way contingency tables, *Journal of the American Statistical Association*, 89, pp. 190-200.

FRIENDLY, M. (2000) *Visualizing Categorical Data* (Cary, NC, SAS Institute Inc.).

FRIENDLY, M. (2009) *Milestones in the history of thematic cartography, statistical graphics, and data visualization.* Ver http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf (Consultado el 16.V.2011).

FRIENDLY, M. (2010) *Working with categorical data with R and the vcd and vcdExtra packages.* Ver http://www.datavis.ca/courses/VCD/vcd-tutorial.pdf (Consultado el 16.V.2011).

GOODMAN, L. A. (2007) Statistical Magic and/or Statistical Serendipity: An Age of Progress in the Analysis of Categorical Data, *Annual Review of Sociology*. 33, pp.1-19.

HABERMAN, S. J. (1973) The Analysis of Residuals in Cross-Classified Tables, *Biometrics*, 29, pp. 205-220.

HARTIGAN, J. A. y KLEINER, B. (1984) A mosaic of television ratings, *The American Statistician*, 38, pp. 32-35.

HERNÁNDEZ, L. (2001) *Técnicas de taxonomía numérica* (Madrid, La Muralla).

HOLLAND, S. M. (2006) *Cluster Analysis*. Ver http://www.uga.edu/strata/software/pdf/clusterTutorial.pdf (Consultado el 16.V.2011).

JOARISTI, L. y LIZASOAIN, L. (1999) *Análisis de correspondencias* (Madrid, La Muralla).

LEBART, L.; MORINEAU, A. y FENELON, J. P. (1985) *Statistique exploratoire multidimensionelle* (París, Dunod).

LEBART, L.; MORINEAU, A. y PIRON, M. (2000) *Tratamiento estadístico de datos* (Barcelona, Marcombo).

LIZASOAIN, L. y JOARISTI, L. (1999) Análisis de Correspondencia, en TEJEDOR, F. J. y NIETO, S. (coords.) *Técnicas de análisis multivariante* (Salamanca, Tesitex), pp. 15-62.

LIZASOAIN, L. y JOARISTI, L. (1999) El programa SPAD para Windows, en TEJEDOR, F. J. y NIETO, S. (coords.) *Técnicas de análisis multivariante* (Salamanca, Tesitex), pp. 63-92.

LIENERT, G. A. (1968) *Die "Konfigurationsfrequenzanalyse" als Klassifikationsmethode in der klinischen Psychologie [Configural Frequency Analysis – A classification method of clinical psychology]*. Paper presented at the 26. Kongress der Deutschen Gesellschaft für Psychologie in Tübingen 1968.

LINDSAY, B. G.; KETTENRING, J. y SIEGMUND D. O. (2004) A Report on the Future of Statistics, *Statistical Science*, 19:3, pp. 387-413.

LITTLE, T. D.; SCHNABEL, K. U. y BAUMERT, J. (2000) *Modeling longitudinal and multilevel data* (New Jersey, Erlbaum).

MANNING, C. (2007) *Logistic regression (with R)*. Ver http://nlp.stanford.edu/~manning/courses/ling289/logistic (Consultado el 16.V.2011).

NENADIC, O. y GREENACRE, M. (2007) Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package, *Journal of Statistical Software*, 20:3.

NIE, N.; BENT, D. H. y HULL, C. H. (1970) *SPSS: Statistical Package for the Social Sciences* (New York, McGraw-Hill).

OKSANEN, J. (2010) *Cluster Analysis. Tutorial with R*. Ver http://cc.oulu.fi/~jarioksa/opetus/metodi/sessio3.pdf (Consultado el 16.V.2011).

REITZLE. M. y VONDRACEK, F. W. (2000) Methodological Avenues for the Study of Career Pathways, *Journal of Vocational Behavior*, 57, pp. 445-467.

RUIZ-MAYA, L.; MARTÍN, F. J.; MONTERO, J. M. y URIZ, P. (1995) *Análisis estadístico de encuestas: datos cualitativos* (Madrid, AC).

SILVA, L. C. y BARROSO, I. M. (2004) *Regresión logística* (Madrid, La Muralla).

TUKEY, J. W. (1977) *Exploratory Data Analysis* (Boston, Addison-Wesley).

VALERO-MORA, P. M.; YOUNG, F. W. y FRIENDLY, M. (2003) Visualizing categorical data in vista, *Computational Statistics & Data Analysis*, 43:4, pp. 495-508.

VOLLE, M. (1997) *Analyse des données* (París, Economica).

VON EYE, A. (2002) *Configural frequency analysis – Methods, models, and applications* (Mahwah, NJ, Erlbaum).

VON EYE, A. (2004) Base models for Configural Frequency Analysis, *Psychology Science*, 46, pp. 150-170.

VON EYE, A. y GUTIÉRREZ PEÑA, E. (2004) Configural frequency analysis – The search for extreme cells, *Journal of Applied Statistics*, 31, pp. 981-997.

VON EYE, A.; MAIR, P. y MUN, E. Y. (2010) *Advances in Configural Frequency Analysis* (New York, Guilford).

WAGNER, P.; SCHOBER, B. y SPIEL, C. (2008) Time students spend working at home for school, *Learning and Instruction,* 18, pp. 309-320.

YOUNG, F. W. y BANN, C. (1997) ViSta: A Visual Statistics System, en STINE, R. y FOX, J. (eds.) *Statistical Computing Environments for Social Research* (Thousand Oaks, Sage), pp. 207-235.

YOUNG, F. W.; VALERO-MORA, P. M. y FRIENDLY, M. (2006) *Visual Statistics Seeing Data with Dynamic Interactive Graphics* (New J., Wiley).

YU, C. H. (2010) Exploratory data analysis in the context of data mining and resampling, *International Journal of Psychological Research*, 3:1, pp. 9-22.

Resumen:
Las nuevas tecnologías y la investigación educativa. El análisis de datos de variables categoriales

El objetivo de este artículo es analizar el impacto de las nuevas tecnologías de la información y la comunicación en la investigación educativa, y más espcíficamente en el análisis de datos categoriales. Desde este punto de vista se presentan y describen las principales características de técnicas estadísticas como las siguientes : Análisis de Frecuencia de Configuraciones, Análisis de Tablas de Contingencia Tridimensionales, Modelos *Log-lineales,* Modelos *Logit, Probit* y Regresión Logística, Análisis Factorial de Correspondencias, Clasificación Automática con Variables Categoriales y Visualización de Datos Categoriales. En todos los casos se describen los programas informáticos adecuados con especial atención a los programas de código abierto, libres y gratuitos como R.

Descriptores: Análisis de datos categoriales/Análisis de Frecuencia de Configuraciones/ Modelos Log-lineales/ Regresión Logística/Análisis de Correspondencias,/-Clasificación Automática/Visualización de Datos Categoriales/R.

Summary:
The new information and communication technologies and the educational research. Categorical Data Analysis

The aim of this paper is to analyze the impact of the information and communication technologies on the educational research, and more specifically on the categorical data analysis. From this point of view, the main features of several statistical techniques for categorical data are described: Configural Frequency Analysis (CFA), 3-way Contingency Tables, Log-linear Models, Logit and Probit Models and Logistic Regression, Correspondence Analysis, Categorical Cluster Analysis and Categorical Data Visualization. In all the cases, the software is described and special attention is paid to the open source and free software such as R.

Key Words: Categorical Data Analysis/Configural Frequency Analysis (CFA)/Log-linear Models/Logistic Regression/Correspondence Analysis/Cluster Anaysis/Categorical Data Visualization/R.