QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 217-231, 1995

EL SISTEMA ADEST

J.M CARIDAD OCERIN* y R. ESPEJO MOHEDANO*

ADEST is an intelligent software enviroment, user friendly, oriented toward categorical data analysis. It's based on a generalization of Havranek procedure for log-linear model especification. It includes also a semi-automatic module to agregate categories whith low marginal expected frequencies, and several tools like a categorical data editor and a generator of BMDP programs.

ADEST System.

Key words: Categorical Data Analysis, Log-linear models, Multiple contingency tables.

1. INTRODUCCIÓN

La utilización de variables cualitativas es un hecho cada vez más frecuente en el ámbito de las Ciencias Sociales y Experimentales: análisis de encuestas, tratamiento de historias clínicas o epidemiológicas, y otras situaciones en análisis de datos.

Los investigadores y profesionales en estos campos no suelen ser expertos en Estadística, y en todo caso es más frecuente que no conozcan en profundidad las técnicas relativas a datos categorizados, por lo que se hace necesario la consulta a un estadístico profesional en las distintas fases del análisis de datos. Los desarrollos acaecidos en los últimos años en las aplicaciones software convencionales, sistemas potenciados y de inteligencia artificial en la construcción de sistemas expertos, permite, a veces, abordar la consultoria estadística de una forma sistemática hasta llegar a la elaboración de un logical de ayuda y apoyo al usuario final, disminuyendo la carga del consultor estadístico y facilitando la labor de éste.

*J.M. Caridad Ocerin y R. Espejo Mohedano. Departamento de Estadística. Escuela Superior de Ingenieros Agrónomos y Montes.
Avda. Menéndez Pidal s/n.
Apdo. 3048, 14080 Córdoba. Spain.

- Article rebut el febrer de 1994.
- Acceptat el setembre de 1994.

Tales herramientas son tremendamente poderosas para cualquier investigador ya que no sólo puede realizar un determinado análisis de forma más rápida, sino que pueden repetirlos tantas veces como se desee, quizás con una sola pequeña diferencia en el conjunto de parámetros. El ordenador ha cambiado la forma de pensar del estadístico en muchas áreas de la Estadística, en particular sobre ajuste de modelos. Según palabras de Hand (1992), estas ventajas también suponen un cierto peligro: el sobreajustar modelos es un peligro real, no en la forma convencional de sobreparametrizar un modelo, pero sí una forma más sutil e insidiosa de intentar ajustar muchos modelos. Existen métodos estadísticos para reconocer sobreparametrización, pero no para esta segunda clase de ajustes que quizás representa un problemas más serio. Por tanto, es necesario poner a disposición de usuarios de análisis de datos, un conjunto de herramientas inteligentes que eviten el uso inapropiado, y desgraciadamente muy frecuente, de los paquetes estadísticos, que dirijan su utilización e interpretación, lo que conlleva a la automatización de los procesos de decisión y selección de estrategias de análisis de datos. El disponer de tales herramientas supone un apoyo importante y cada vez más necesario en la utilización correcta de los métodos estadísticos y sus limitaciones.

Los modelos log-lineales constituyen un tipo de modelos utilizados en análisis de datos de tipo cualitativo. Además de representar la asociación entre los factores o variables en estudio, permiten cuantificar y determinar la influencia de los factores sobre las frecuencias de una tabla multidimencional, así como la frecuencia de efectos conjuntos de varios factores (interacciones) en cada celda de la tabla. En análisis exploratorio de datos, uno de los objetivos fundamentales es encontrar todos los modelos que describan de forma adecuada a los datos. Una primera aproximación para este problema, que puede utilizarse para tablas de contingencia pequeñas (menos de 3 o 4 factores), es ajustar todos los posibles modelos log-lineales jerárquicos y después utilizar alguna medida estadística, tipo AIC (Akaike 1973,1987) o BIC (Raftery 1986), para seleccionar aquellos modelos log-lineales que representen de forma óptima a los datos. Para tablas de más de 4 dimensiones, el problema empieza a dificultarse por la complejidad computacional que conlleva. Con 4, 5 y 6 variables, el número de modelos log-lineales jerárquicos posibles es 168, 7.581 y 7.828.354 respectivamente. De forma natural, mientras mayores términos de interacción contenga un modelo, mayor será la complejidad en la interpretación de éste; incluso cuando se consideran sólo modelos de orden 2 o menor, 5 y 6 variables producen 1.451 y 40.070 modelos log-lineales respectivamente. Esta complejidad computacional estriba no sólo en la generación y mantenimiento de estos modelos sino también a la hora de ajustarlos.

Muchos procedimientos de selección de modelos son capaces de seleccionar un único modelo a través de una serie de tests de significación sobre parámetros o grupos de parámetros de modelos. Una revisión de la mayoría de estos procedimientos puede consultarse en Wrigley (1985). Los modelos seleccionados mediante este tipo de técnicas difieren a menudo dependiendo del procedimiento elegido, incluso uti-

lizando el mismo test estadístico. Según palabras de Tukey (1985) "La Ciencia es el resultado de trabajar con múltiples hipótesis", lo que incita a adoptar métodos que proporcionen respuestas múltiples. Las aproximaciones de optimización, con un criterio bien definido de optimabilidad de modelos, son más consistentes puesto que el criterio clasifica sin reparar en el método de búsqueda.

Este tipo de aproximaciones son computacionalmente más complejas que las basadas en test de significación sobre parámetros de modelos. Edwards y Havranek (1984, 1987) han desarrollado un procedimiento de búsqueda de modelos log-lineales jerárquicos basado en un criterio de optimización, y que en este trabajo se generaliza introduciendo modificaciones necesarias en procesos de modelización, selección de un punto inicial de partida en el procedimiento, además de dotarlo de elementos de decisión necesarios para su completa automatización.

Otro problema frecuente en el análisis estadístico en general, y en tablas de contingencia y modelos log-lineales en particular, radica en un tamaño muestral pequeño y que puede provocar obtener tablas dispersas, es decir, tablas en donde aparecen ciertas combinaciones de celdas con frecuencias bajas. Este tipo de tablas presentan graves inconvenientes tanto a nivel teórico, al no poder garantizar la convergencia de algunos estadísticos usuales, como computacionales, al tener que reproducir algunas frecuencias esperadas cero.

Poco puede hacerse en este tipo de problemas. La solución obvia sería buscar más datos, pero, desgraciadamente, en la mayoría de los casos esto es inviable. Tradicionalmente se han añadido constantes a valores muestrales, pero estas prácticas no son recomendables para todos los modelos y todas las circunstancias. También se han propuesto tests estadísticos alternativos (Cressie y Read 1984), pero la falta de potencia de éstos aún no ha sido resuelta. Colapsar (refundir) sobre categorías o variables, de tal forma que se garantice un tamaño mínimo en las frecuencias esperadas, es el camino por el que optan la mayoría de los investigadores, y aún cuando este tipo de estrategias generalmente afecta a la estimación de los parámetros del modelo log-lineal, mientras se indique que las relaciones encontradas pertenecen a la tabla actualmente utilizada, esta técnica puede ser la más prudente y en algunos casos la única viable (Agresti 1990).

En esta línea, el trabajo que nos ocupa presenta un mecanismo de refundición de categorías de variables en tablas de contingencia en las que ocurren los problemas aludidos. Dicho mecanismo es semiautomático en el sentido de que es el investigador el que tiene la última decisión sobre adoptar o no las recomendaciones ofertadas como las más óptimas, atendiendo a las características propias de las variables. Además se implementa un entorno inteligente de desarrollo adecuado y cómodo para el tratamiento de un conjunto de datos de tipo cualitativo, desde una primera fase de descripción de variables implicadas en el estudio y manipulación de datos, hasta la obtención de mecanismos y herramientas de control necesarias en procesos de modelización para

llegar a conclusiones y resultados que se ajusten a los requerimientos marcados. Esto se aborda mediante la elaboración de un lógical integrado que incluye cinco módulos principales: un módulo de creación de descripciones, otro de introducción controlada de datos, un tercer módulo de verificación, análisis y refundición semiautomática de categorías en tablas de contingencia dispersas, el módulo de selección automática de modelos log-lineales, y un último de generación de ficheros de instrucciones BMDP. El enfoque elegido es interactivo, de uso simple para no expertos con amplios sistemas de ayuda, y sobre sistemas estándares.

La fase de cálculo numérico está resuelta de forma satisfactoria con los paquetes estadísticos convencionales, por lo que se ha considerado más adecuado enlazar las técnicas previas de modelización con la generación automática de uno o varios ficheros de instrucciones BMDP, así como la importación de los ficheros de datos preparados con un logical de tratamiento de datos categóricos adecuado al usuario final, aunque en los procesos de decisión incorporados en el análisis de datos se necesitan procesos numéricos de estimación de modelos log-lineales, que se han desarrollado e incorporado al sistema.

El sistema ADEST está programado en Turbo-Pascal 6.0, bajo sistema operativo DOS en ordenadores compatibles IBM que dispongan de tarjeta gráfica VGA, coprocesador matemático y ratón compatible Microsoft.

2. SELECCIÓN DE MODELOS

Para describir el procedimiento se necesita introducir la siguiente notación: la inclusión de modelos log-lineales se expresa en la forma habitual, así pues, $m_1 \leq m_2$ significa que m_1 es un submodelo de m_2, es decir en m_2 están los términos de m_1 y alguno más. Lógicamente, $m_1 < m_2$ se usa para decir que $m_1 \leq m_2$ y $m_1 \neq m_2$. A partir de este momento, y por razones de simplificación, las referencias al término "modelo" han de entenderse como "modelo log-lineal". Para cualquier conjunto de modelos $\mathcal{S}$ en $\mathcal{F}$ se definen $\max(\mathcal{S})$ y $\min(\mathcal{S})$ como:

$$\max(\mathcal{S}) = \{s \in \mathcal{S} \ / \text{ si } \ s < t \Rightarrow t \notin \mathcal{S}\}$$
$$\min(\mathcal{S}) = \{s \in \mathcal{S} \ / \text{ si } \ t < s \Rightarrow t \notin \mathcal{S}\}$$

Los principios básicos en los que se basa el método de selección de modelos hacen referencia al principio de coherencia (Gabriel, 1969) y la utilización de dos reglas de uso frecuente en el ámbito del análisis multidimensional. El principio de coherencia implica que un procedimiento que involucre contrastar un conjunto de modelos, no debe de aceptar un modelo mientras se rechace un modelo más general, entendiéndose éste como un modelo que incluye al primero. Claramente, si se considera un modelo

compatible con los datos, es absurdo estimar un modelo más general incompatible con ellos. Las dos reglas a las que se ha hecho referencia son las que siguen:

1. Si un modelo se acepta, entonces todos los modelos que lo incluyan se consideran aceptados.

2. Si un modelo es rechazado, entonces todos sus submodelos se consideran rechazados.

En lo que sigue, se dice que un modelo es $d-$aceptado (débilmente aceptado) si incluye un modelo más simple aceptado, es decir deducimos que puede ser aceptado, y $d-$rechazado si está incluido en un modelo rechazado.

Estas reglas verifican el principio de coherencia y son prácticas puesto que reducen de forma considerable el número de modelos a ser ajustados. Así pues, los modelos $d-$aceptados o $d-$rechazados no necesitan ser considerados como especificaciones alternativas.

El procedimiento propuesto busca un conjunto $\mathcal{A}$ de modelos aceptados y un conjunto $\mathcal{R}$ de modelos rechazados, de tal forma que cualquier otro modelo de $\mathcal{F}$ (familia de todos los modelos log-lineales jerárquicos posibles que se pueden formar a partir de un conjunto de variables dadas) o contiene un modelo de $\mathcal{A}$, y por tanto es $d-$aceptado, o está contenido en un modelo de $\mathcal{R}$, y por tanto es $d-$rechazado.

Para cualquier modelo $m \in \mathcal{F}$, dado un nivel de significación α, un test de significación adecuado permitirá aceptar o rechazar su ajuste a los datos. Por tanto, en cualquier estado se pueden clasificar los modelos en $\mathcal{F}$ en tres subconjuntos:

$$\begin{aligned} \mathrm{A} &= \{m \in \mathcal{F}/m \text{ se acepta}\} \\ \mathrm{R} &= \{m \in \mathcal{F}/m \text{ se rechaza}\} \\ \mathrm{I} &= \{m \in \mathcal{F}/m \text{ no ha sido contrastado}\} \end{aligned}$$

Si $m \in \mathcal{F}$ se ajusta a los datos, puede ser incluido en el conjunto A; si $m \in \mathcal{F}$ se rechaza pertenecerá a R, y si m no ha sido contrastado estará incluido en I. El propósito del procedimiento de selección de modelos es reducir el número de modelos de I hasta conseguir que $\mathrm{I}=\emptyset$ con el menor coste computacional y de tiempo posibles.

Las construcciones básicas usadas para determinar qué modelo hay que contrastar son los conceptos de $a-$dual y $r-$dual de un conjunto dado. El $a-$dual de un conjunto de modelos $\mathcal{S}$, escrito $D_a(\mathcal{S})$ se define como el conjunto de modelos minimales en $\mathcal{F}$ que no están contenidos en ningún modelo de $\mathcal{S}$. Este concepto es de interés cuando los modelos en $\mathcal{S}$ han sido rechazados.

$$D_a(\mathcal{S}) = \min\{m \in \mathcal{F}/m \not\leq m^*, \forall m^* \in \mathcal{S}\}$$

De forma similar, el r-dual de $\mathcal{S}$, escrito $D_r(\mathcal{S})$, se define como el conjunto de modelos en $\mathcal{F}$ que no contienen ningún modelo en $\mathcal{S}$, siendo de utilidad cuando los modelos

en $\mathcal{S}$ han sido aceptados.

$$D_r(\mathcal{S}) = \max\{m \in \mathcal{F} / m^* \not\leq m, \quad \forall m^* \in \mathcal{S}\}$$

Si $\mathcal{S}$ está vacío, se define: $D_a(\mathcal{S}) = \min(\mathcal{F})$ y $D_r(\mathcal{S}) = \max(\mathcal{F})$

El primer paso, como ya se ha dicho, consiste en contrastar un conjunto inicial de modelos $\mathcal{M}I$ y construir los conjuntos $\mathcal{A}$ y $\mathcal{R}$.

$$\begin{aligned} \mathcal{A} &= \min\{m \in \mathcal{M}I / m \text{ se acepta}\} \\ \mathcal{R} &= \max\{m \in \mathcal{M}I / m \text{ se rechaza}\} \end{aligned}$$

El siguiente paso consiste en contrastar o $D_a(\mathcal{R}) \setminus \mathcal{A}$ (modelos en $D_a(\mathcal{R})$ que no pertenecen a $\mathcal{A}$) o bien $D_r(\mathcal{A}) \setminus \mathcal{R}$, y así sucesivamente actualizando los conjuntos $\mathcal{A}$ y $\mathcal{R}$. Si suponemos que en cualquier paso los modelos en $D_r(\mathcal{A}) \setminus \mathcal{R}$ son contrastados y rechazados, entonces después de haber actualizado $\mathcal{R}$, se tiene que $D_r(\mathcal{A}) \setminus \mathcal{R}$ está vacío y por tanto en este punto puede parar el procedimiento. A las mismas conclusiones se llega si los modelos en $D_a(\mathcal{R}) \setminus \mathcal{A}$ se contrastan y todos han sido aceptados.

Cuando se aplica el procedimiento a un problema particular hay que considerar cuidadosamente (a) qué test de ajuste se va a usar, (b) qué conjunto de modelos inicial se va a contrastar y (c) si contrastar $D_r(\mathcal{A}) \setminus \mathcal{R}$ o $D_a(\mathcal{R}) \setminus \mathcal{A}$ en cada paso.

Una modificación al procedimiento anterior es la que sigue. En muchas aplicaciones podrían considerarse modelos que sólo contengan un modelo dado m_0, es decir un submodelo, o bien que una determinada interacción aparezca en el modelo final. Esto puede ser posible en el contexto del procedimiento, al asignar inicialmente los modelos en $D_r(m_0)$ a $\mathcal{R}$, sin contrastar. Por tanto, para cualquier modelo $m \in \mathcal{F}$, m_0 es un submodelo de m si y sólo si m no está contenido en algún modelo de $D_r(m_0)$, con lo que se obtiene el resultado deseado, es decir, si se asigna $D_r(m_0) = \mathcal{R}$, se rechazan a priori todos aquellos modelos que no contengan a m_0.

El test estadístico de ajuste que se ha utilizado es el clásico test chi-cuadrado que garantiza el principio de coherencia, uno de los pilares básicos del procedimiento.

El punto de partida, es decir el conjunto de modelos inicial, es un punto crucial en el desarrollo del procedimiento no en cuanto a la solución final que, obviamente, ha de ser la misma a un mismo nivel de significación independientemente también del conjunto $D_a(\mathcal{R})$ o $D_r(\mathcal{A})$ seleccionado en cada paso (Edwards, 1987), sino al tiempo de cómputo necesario para llegar a dicha solución. No es lo mismo partir de un conjunto de modelos próximo a la solución que otro mucho más alejado; el número de pasos necesarios para converger es mucho mayor en el segundo caso. Por tanto, si el investigador conoce o sospecha algún tipo de relación o interacción entre variables, puede comenzar el procedimiento con esta información, pero si no es éste

el caso, el proceso automático de selección ha de construir un conjunto de modelos inicial que de una forma no muy compleja intente reducir el tiempo de proceso y a la vez tenga en consideración la naturaleza de las variables en estudio. Los contrastes STP (Simultaneous Test Procedure) sobre la existencia de efectos de un orden k o superior, pueden también ayudar en la selección del conjunto inicial.

Partiendo de la hipótesis no restrictiva y lógica de que al menos existe una variable con mayor peso respecto a las demás, entre el conjunto de variables implicadas, el primer objetivo es construir un conjunto de modelos anidados (un conjunto m_2 se dice anidado respecto de m_1 cuando los parámetros de m_2 constituyen un subconjunto de los parámetros de m_1) que tenga en consideración tales variables y sus interrelaciones. Es norma usual en análisis exploratorios de datos incluir el término de mayor interacción entre variables con mayor peso y concentrar el proceso de modelización sobre términos que relacionen variables con menor peso.

Así pues, si $X = \{x_1, x_2, \ldots, x_n\}$ representa el conjunto de variables sobre las que se va a operar, e $Y = \{y_1, y_2, \ldots, y_m\}$ $(m < n)$ el subconjunto de X formado por las variables de mayor peso, se puede considerar el conjunto de modelos iniciales $\mathcal{MI}$ como:

$$\mathcal{MI} = \{\{G \cup C_i\} \cup \{\acute{C}_j\}\}_{i=1,\ldots,k}^{j=k+1,\ldots,n-1}$$

donde:

- k es el orden de la mayor interacción G posible entre las variables en Y
- $\{\{G \cup C_i\}\}$ es el conjunto de modelos de la forma $\{G \cup C_i\}$ $i=1,\ldots,k$ y $C_i = \bigcup_x c_i$ con c_i interacción de orden i en X
- $\{\{\acute{C}_j\}\}$ es el conjunto de modelos de la forma $\{\acute{C}_j\}$ $j=k+1,\ldots,n-1$ y $\acute{C}_j = \bigcup_x \acute{c}_j$ con $\acute{c}_j$ interacción de orden j en X.

Otro posible conjunto inicial de modelos podría ser considerar el modelo no saturado de mayor orden, es decir aquel que contiene todas las interacciones de orden $n-1$, garantizando, de igual forma, que las interrelaciones máximas entre variables de mayor peso están incluidas. Análisis sobre tiempos de cómputo necesarios para obtener la solución final mediante procesos de simulación de tablas, confirman que iniciando el proceso de selección partiendo de modelos anidados, el tiempo empleado se reduce de forma considerable con una ganancia aproximada de 4 a 1.

Otro estado a considerar dentro del procedimiento de selección supone la elección entre la construcción del $D_a(\mathcal{R})$ o el $D_r(\mathcal{A})$. Dada la naturaleza lógica y algebraica del método para la construcción de estos conjuntos (Havranek,1987), el número de combinaciones entre generadores de modelos de $\mathcal{A}$ o $\mathcal{R}$, y por tanto el número de modelos intermedios para tales construcciones, puede llegar a ser excesivo y supone otro punto crítico en el proceso de control, tanto en la velocidad de proceso como

en las limitaciones de la memoria disponible. Lo ideal en cuanto a tiempo de proceso sería construir mediante el algoritmo oportuno todos los modelos intermedios y minimizar éstos para obtener el $D_a(\mathcal{R})$ o el $D_r(\mathcal{A})$, pero tal número de modelos intermedios, aunque se utilizan estructuras de datos dinámicas para almacenarlos, supone un desborde de la memoria disponible relativamente rápido en equipos con pocos recursos.

La solución dada al problema planteado consiste en pasar un proceso de construcción-minimización de $D_a(\mathcal{R})$ o $D_r(\mathcal{A})$ después de la generación de cada modelo y no esperar a que se generen todos. De esta forma el proceso consume más tiempo pero las necesidades de memoria disminuyen de forma considerable ya que se evita manejar muchos modelos repetidos o incluidos en otros.

Por último, se han implementado dos mecanismos de control y decisión para automatizar de forma completa el método de selección de modelos. El primero para decidir si construir el $D_a(\mathcal{R})$, el $D_r(\mathcal{A})$ o ambos, analizando el número de combinaciones entre generadores de modelos necesarios para tales construcciones. Si el número de estas combinaciones es relativamente pequeño se considera indiferente uno u otro, en caso contrario se opta por aquel que requiere menor número de modelos intermedios para su construcción. El segundo mecanismo de control está implementado para decidir si contrastar los modelos del $D_a(\mathcal{R})$ o el $D_r(\mathcal{A})$, ateniéndose a la siguientes normas y orden: se selecciona el conjunto con un menor número de modelos y dentro de éstos, el que tenga un menor número total de generadores, y por último, si coinciden ambos criterios, el de menor varianza de generadores.

El hecho de adoptar estas normas o reglas viene dado, principalmente por el tiempo de proceso consumido en el procedimiento de cálculo de frecuencias esperadas estimadas para cada modelo (Bishop, Fienberg y Holland, 1975). Aunque no es posible conocer cuanto tiempo puede consumir este proceso (al ser un proceso iterativo, el número de iteraciones dependerá de la naturaleza de los datos), si se conoce el número de pasos necesarios en cada iteración: tantos como generadores tenga el modelo. Por otra parte, también se ha seguido otro proceso lógico y tradicional en la elección de modelos: el criterio de parsimonia, lo que supone elegir los modelos o conjunto de modelos más simples.

3. REFUNDICIÓN DE CATEGORÍAS

Un problema frecuente en el análisis estadístico, en particular en análisis de tablas de contingencia y modelos log-lineales, radica en un tamaño muestral pequeño. Si muchas frecuencias esperadas de celdas de una tabla de contingencia multidimensional

tiene valores bajos ($< 5o < 1$), no puede esperarse que los tests estadísticos G^2 de razón de verosimilitudes y X^2 de Pearson se aproximen a la distribución teórica chi-cuadrado (Haberman, 1988). Es más, los efectos estandarizados (Goodman, 1971) y los errores ajustados no se aproximan a la distribución normal estándar. Cochran estudió en una serie de artículos la aproximación chi-cuadrado para X^2: en 1954 sugirió que para contrastar la independencia con más de un grado de libertad, un valor mínimo de 1 era permisible siempre y cuando no existan más de un 20% de valores esperados de celdas por debajo de 5. Larntz (1978), Koehler y Larntz (1980) y Koehler (1986) muestran que el estadístico X^2 se comporta mejor que G^2 para tamaños muestrales pequeños y tablas dispersas. Muestran que la distribución muestral de G^2, generalmente, se aproxima en menor grado a una chi-cuadrado que X^2 cuando n/N es menor que 5, siendo n el tamaño muestral y N el número de celdas.

Agresti (1990) aborda este problema y realiza una recopilación de técnicas alternativas, entre las que cita a Cressie y Read (1984) con los tests de potencia divergente basados en la familia de estadísticos:

$$\frac{2}{\lambda(\lambda+1)}\sum n_i\left[\left(\frac{n_i}{\hat{m}_i}\right)^{\lambda}-1\right] \quad ; \quad -\infty<\lambda<\infty$$

Cita también que otras adaptaciones de test chi-cuadrado han sido propuestas por Berry y Mielke (1988) y Zelterman (1987) y comenta que las ventajas de estas adaptaciones sobre los tests estadísticos convencionales no son claras en muchas circunstancias y además, la falta de fuerza de estos tests aún no ha sido resuelta.

En el sentido referenciado en la introducción sobre refundición de categorías o variables, de tal forma que se garantice un tamaño mínimo en las frecuencias esperadas, y considerando tablas multidimensionales, se ha trabajado en criterios que realicen la refundición semiautomática de categorías. Para ello se han definido los cuatro criterios de decisión y que se exponen a continuación, refiriéndose todos ellos a la tabla de frecuencias esperadas bajo un determinado modelo log-lineal. Los criterios de decisión adoptados son los siguientes:

- Se considera que una tabla no necesita una refundición cuando al menos el 80% de sus valores son mayores que 5 y todos mayores que 1, aunque este criterio se puede alterar en cada caso concreto.

- Se construye el vector ordenado *IRV* (índice de refundición de variables) cuyos elementos son de la forma (V^{X_i}); $1 \leq i \leq n_v$ (n_v=número de variables) y cada elemento V^{X_i} se corresponde con el número de líneas (filas o columnas en el caso bidimensional) "defectuosas" en relación al número total de categorías la variable X_i, siempre y cuando el número de categorías de la variable sea mayor o igual a 3. Si en algunas variables estos valores coinciden, la ordenación

se realiza conforme a aquella que tenga más categorías (esta última decisión parte de la lógica de que al refundir se pierde una categoría). La variable correspondiente al primer elemento de *IRV* será aquella en la que se ha de producir una refundición entre dos de sus categorías. Para saber si una línea es defectuosa se procederá de la siguiente forma:

- Se busca la línea con más valores menores que 5 calculándose dicho número de valores.
- Se le calcula el "tope" (número entero entre 0 y 100 dado como información a priori) por ciento a ese número de valores encontrado.
- Si una línea tiene más valores menores que 5 que el módulo del porcentaje calculado anteriormente, se considera defectuosa.

De cualquier forma es el investigador el que decide si acepta refundir alguna categoría de la variable propuesta por el programa según el criterio anterior, o por el contrario prefiere otra.

- La categoría "peor" dentro de una variable (y por tanto la que necesita ser refundida con otra) será aquélla correspondiente al primer elemento del vector ordenado IRC^{X_k}. IRC^{X_k} es un vector cuyos elementos son de la forma $(C_j^{X_k})\ 1 \leq j \leq n_{ck}$ (n_{ck}=número de categorías de X_k) para la variable X_k seleccionada en el paso anterior. Cada elemento $C_j^{X_k}$ es el número de valores menores que 5 que contiene la categoría j de X_k. En el caso de existir más de una categoría con el mismo valor, la ordenación se realiza conforme a aquella que tenga menor valor medio.

- Para decidir con qué categoría ha de refundirse la encontrada con anterioridad, hay que distinguir entre que la escala de medida de la variable en cuestión sea nominal u ordinal, es decir, si sus categorías están ordenadas o no:

  - *Ordinal:* Sólo puede refundirse con una categoría adyacente. Si la peor categoría es la primera o la última, sólo hay una posibilidad: La segunda y la penúltima, respectivamente. Si no es ése el caso, existen dos y se tomará la peor. TABCONT encuentra ésa peor buscando en IRC^{X_k} los valores correspondientes a las categorías adyacentes y seleccionando aquella que figure primero dentro de IRC^{X_k}, pero permite al usuario seleccionar la otra posibilidad.
  - *Nominal:* La peor categoría puede refundirse con cualquier otra. En este caso, TABCONT presenta al usuario las categorías correspondientes a los restantes elementos de IRC^{X_k}. En este punto es el usuario el que ha de decidir la primera refundición de la lista que le parezca lógica (de esta forma nos aseguramos haber realizado la mejor refundición posible).

Tanto si las opciones presentadas son aceptadas o se introducen otras distintas, el proceso continua adicionando las frecuencias de las categorías obtenidas o seleccionadas, reajustando tablas y actualizando la información sobre variables y categorías, hasta que se cumplan las condiciones del test.

4. GENERADOR DE FICHEROS DE INSTRUCCIONES BMDP

Como un módulo más integrado en el sistema ADEST, GEN-BMDP tiene como objetivo la construcción de un fichero de instrucciones BMDP. Parte de la idea de algunos autores de que los sistemas estadísticos pudieran imitar de manera automática los pasos seguidos por un estadístico experimentado y generar igualmente de forma automática los correspondientes interfaces con los paquetes estadísticos, en este caso BMDP.

Está enfocado hacia el investigador que ya familiarizado con alguna técnica estadística y conozca bien sus objetivos (en este apartado el sistema puede dejarlos claros al seleccionar un conjunto de modelos log-lineales óptimo, para posteriormente ser tratados y analizados con el paquete estadístico), pueda disponer de una herramienta de fácil manejo que le descargue de la tarea de programar en el lenguaje propio del paquete estadístico BMDP.

Naturalmente, se han seguido unos pasos que se corresponden con la lógica seguida en la codificación de instrucciones BMDP:

- generación automática de párrafos generales de descripción de datos,
- generación automática de los párrafos particulares correspondientes a los objetivos específicos del estudio.

A partir de un fichero de descripción sobre las características de las variables, se obtiene la información necesaria para concluir el primer paso, con lo que se generarían los párrafos:

/INPUT	CASES	FILES
	FORMAT	TITLE
	VARIABLES	
/VARIABLES	NAMES	MAX
	USE	MIN
	MISSING	GROUP
/GROUP	CODES	NAMES
	CUTPOINT	

Para el segundo, una vez leído el fichero de descripción, se presenta al investigador un protocolo de preguntas distinto según el programa BMDP seleccionado y las características descritas para las variables en uso. Dicho protocolo de preguntas está basado en las posibilidades que permite el programa seleccionado. Si éste es muy amplio, como podría ser 4F, habrá que realizar más preguntas al usuario para concretar qué tratamiento se desea exactamente. Las preguntas en cada caso son distintas según las respuestas precedentes y la naturaleza de las variables utilizadas. Puede decirse, en este sentido, que GEN-BMDP es seudointeligente y además, ofrece la posibilidad de ayuda en cada momento al usuario, con referencia a las preguntas formuladas y posibilidades presentadas a éste. Como ya se ha comentado con anterioridad, el sistema ADEST está orientado hacia el tratamiento de datos de tipo categórico, con lo que el generador de ficheros de instrucciones cubre los siguientes programas BMDP:

- 1D, 2D, 5D — descripción y tabulación de datos,
- 4F — tablas de contingencia y modelos log-lineales,
- CA, CAM — análisis de correspondencias simple y múltiple,
- LR, PR — regresión logística dicotómica y policotómica paso a paso.

Una vez seleccionado el programa BMDP y finalizado el protocolo de preguntas correspondiente, GEN-BMDP dispone de la información necesaria para poder generar el fichero de instrucciones.

REFERENCIAS

[1] **Agresti, Alan** (1984). *Analysis of ordinal categorical data.* Wiley Interscience.

[2] **Agresti, Alan** (1990). *Categorical data analysis.* Wiley interscience.

[3] **Akaike, H.** (1973). "Information theory and an extension of the maximun likelihood principle". In B.N. Petrov and B.F. Csaki (Eds), *Second International Symposium on Information Theory*, 267–281. Budapest: Academiai Kiado.

[4] **Akaike, H.** (1987). "Factor Analysis and AIC". *Psychometrika,* **52**, 317–332.

[5] **Baker, R.J.** and **Nelder, J.A.** (1978). *The GLIM system. Release 3. Generalized linear interactive modelling manual.* Oxford: N.A.G.

[6] **Berry, K.J.** and **Mielke, P.W.** (1988). "Montecarlo comparisons of the asymptotic chi-square and likelihood ratio test with the nonasymptotic chi-square test for sparse rxc tables". *Psychol. Bull.,* **103**, 256–264.

[7] **Bishop, Y.S., Fienberg, P.** and **Holland, P.** (1975). *Discrete multivariate analysis: Theory and practice.* MIT Press, Cambridge.

[8] **Caridad, J.M.** and **Espejo, R.** (1991). *Sistema experto en análisis de datos categorizados.* I Seminario Internacional de sistemas expertos en la agricultura mediterránea. Córdoba.

[9] **Caridad, J.M.** and **Espejo, R.** (1991). *Inteligencia artificial y estadística aplicada.* I Seminario Internacional de sistemas expertos en la agricultura mediterránea. Córdoba.

[10] **Clogg, C.C.** and **Becker, M.** (1986). "Log-linear model with SPSS". *Computer science and statistics.*, 263–269. North Holland. Amsterdam.

[11] **Clogg, C.C.** and **Eliason, S.R.** (1987). "Some common problems in log-linear analysis". *Sociological Methods and Research*, **16**, 8–14.

[12] **Cochran, W.G.** (1954). "Some method of strengthening the common chi-square test". *Biometrics*, **10**, 417–451.

[13] **Cressie, N.** and **Read, T.R.** (1984). "Multinomial goodness of fit". *J. Roy. Statist. Soc.*, **46**, 440–464.

[14] **Cressie, N.** and **Read, T.R.** (1989). "Pearson X^2 and the loglikelihood ratio statistic G^2: A comparative review". *Internat. Statist. Rev.*, **57**, 19–43.

[15] **Dixon, W.J.** (1990). *Ed's Statistical software manual. Vol 1 y 2.* University of California Press.

[16] **Edwards, D.** and **Havránek, T.** (1985). "A fast procedure for model search in multidimensional contingency tables". *Biometrika*, **72**, 339–351.

[17] **Edwards, D.** and **Havránek, T.** (1987). "A fast model selection procedure for large families of model". *J.A.S.A.*, **82, 397**, 205–213.

[18] **Gabriel, K.R.** (1969). "Simultaneous test procedures — some theory of multiple comparisons". *Annals of Mathematical Statistics*, **40**, 224–250.

[19] **Goodman, L.A.** (1971). "The analysis of multiple contingency tables: stepwise procedures and direct estimation methods for buildings models for multiple clasifications". *Technometrics*, **13**, 33–61.

[20] **Goodman, L.A.** (1984). *The analysis of cross-classified data having ordered categories.* Harvard University Press.

[21] **Haberman, S.J.** (1973). "The analisys of residuals in cross-classification tables". *Biometrics*, **29**, 205–220.

[22] **Haberman, S.J.** (1988). "A warning on the use of chi-square statistics with frequency tables with small expected cell counts". *J.A.S.A.*, **83**, 555–560.

[23] **Havránek, T.** (1984). "A procedure for model search in multidimensional contingency tables". *Biometrics*, **40**, 95–100.

[24] **Koehler, K.** and **Larntz, K.** (1980). "An empirical investigation of goodness of fit statistics for sparse multidimensiona tables". *J.A.S.A.*, **75**, 336–344.

[25] **Koehler, K.** (1986). "Goodness of fit test for log-linear models in sparse contingency tables". *J.A.S.A.*, **81**, 483–493.

[26] **Larntz, K.** (1987). "Small-sample comparison of exact levels for chi-squared goodness of fit statistics". *J.A.S.A.*, **73**, 253–263.

[27] **Nelder, J.A.** and **Baker, R.J.** (1985). "Statitstical software: progress and prospects". *Computer Science and Statistics. Proc. of de 16th symposium on the interface*, 33–37. Amsterdam.

[28] **Raftery, A. E.** (1986). "Choosing models for cross-classifications". *Amer. Socciol. Rev.*, **51**, 145–146.

[29] **Tukey, J.W.** (1985). "Comment on more inteligence statistical software and statistical expert systems: Future directions". *The American Statistician*, **39**, 12–14.

[30] **Wrigley, D.** (1985). *Categorical data analysis for geographers and environmental scientists.* Longman. London.

[31] **Zelterman, D.** (1987). "Goodness of fit test for large sparse multidimensional distributions". *J.A.S.A.*, **82**, 624–629.

ENGLISH SUMMARY:

ADEST SYSTEM

J.M. Caridad Ocerin and R. Espejo Mohedano

ADEST is an intelligent software environment for categorical data analysis. It is oriented to end users of log-linear models and contingency tables, who belong to medical or social sciences backgrounds, that is, to professionals who are familiar with the main concepts of multivariate categorical data, but who are not experts in this field. It is an enhanced package, which helps and guides in log-linear model building, and solves some of the practical problems associated with these statistical techniques, like dealing with low expected frecuencies.

In log-linear model specifications, the computational problems become really serious when the number of variables exceeds four; for example there are 7.581 five variable hierarchical models and 7.828.354 with six variables.

Some stepwise procedures for model building lead to one final model after a sequence of goodness of fit test. This can be misleading, as in many situationes it is not possible to specify an optimun model, but a set of alternative models. The stepwise method could also produce errors in the specification depending on the first variables included or excluded.

One alternative is the Edwards and Havranek (1984, 1987) selection procedure, and a generalization of it is implemented in ADEST. From an initial set of possible models $\mathcal{F}$ we use a sequence of acceptable models $\mathcal{A}$, non-acceptable models $\mathcal{R}$, and non-tested models I. This last set is reduced, in a finite number of steps, to the empty set.

A model m is weakly accepted if there is a submodel $m \leq m$ that belongs to the set $\mathcal{A}$, and weakly rejected if $m \in \mathcal{R}$. It is not necessary, then, to test the weakly accepted or rejected models. The set:

$$D_a(\mathcal{R}) = \min\{m \in \mathcal{F} / m \not\leq m \ , \quad \forall m \ \in \mathcal{R}\}$$

includes the minimal log-linear models not included in rejected models, and $D_a(\mathcal{R}) \setminus \mathcal{A} = D_a(\mathcal{R}) \cap \mathcal{A}^c$ (where $\mathcal{A}^c$ is the set of models in $\mathcal{F}$-$\mathcal{A}$) is the set of models to be tested; if one of these models is rejected, it must be included in $\mathcal{R}$ and the set $D_a(\mathcal{R}) \setminus \mathcal{A}$ is then reduced. The procedure stop when $D_a(\mathcal{R}) \setminus \mathcal{A} = \emptyset$. Of course it is possible to define the set:

$$D_r(\mathcal{A}) = \max\{m \in \mathcal{F} / m \not\leq m, \quad \forall m \ \in \mathcal{A}\}$$

and use $D_r(\mathcal{A}) \setminus \mathcal{R} = D_r(\mathcal{A}) \cap \mathcal{R}^c$ (where $\mathcal{R}^c$ is the set of models in $\mathcal{F}$-$\mathcal{R}$); if one of these models is accepted, $\mathcal{A}$ is should be updated and the new $D_r(\mathcal{A}) \setminus \mathcal{R}$ has fewer elements; again this procedure is repeated until this set is empty.

To increase the computational efficiency ADEST uses both sets $D_a(\mathcal{R}) \setminus \mathcal{A}$ and $D_r(\mathcal{A}) \setminus \mathcal{R}$, and the sequential procedure begins with an initial set of models so that the number of steps is minimized. In addition the highest level of interaction to be included in the final set of acceptable (non weakly) models can be stated beforehand, and again, this accelerates the selection procedure.

The final set of log-linear models selected, with a fixed significance level is independent of the initial set of models used in the sequential procedure, but the computational time is dependent of this initial set, so it is important a good selection as starting values.

Also ADEST has a module to produce automatic aggregation of categories to avoid the problem of low expected frequencies in the chi-square test, taking account of the measurement scale of the variables, the expected frequencies associated with each test, and the information provided by the user to guide the collapsability of the categories at different stages of the model selection.

The ADEST environment is user friendly, avalaible for use on DOS based computers (although there is also a Unix version) and it includes an interface with BMDP categorical data analysis programs.