Polo del Conocimiento

Pol. Con. (Edición núm. 49) Vol. 5, No 09
Septiembre 2020, pp. 949-968
ISSN: 2550 - 682X
DOI: 10.23857/pc.v5i9.1745

Combinación de métodos: análisis de correspondencia, simple y múltiple bajo el enfoque de correlaciones canónicas. Clases latentes

Combination of methods: correspondence analysis, simple and multiple under the canonical correlation approach. Latent classes

Combinação de métodos: análise de correspondência, simples e múltipla sob a abordagem de correlação canônica. Classes latentes

Yoel Hernández-Navarro [I]
yoel.hernandez@uta.edu.ec
https://orcid.org/0000-0002-0178-4949

Jesús E. Sánchez-García [II]
jesanch64@yahoo.com
https://orcid.org/0000-0001-8137-2882

Correspondencia: yoel.hernandez@uta.edu.ec

Ciencias técnicas y aplicadas
Artículo de investigación

*__Recibido:__ 05 de julio de 2020 *__Aceptado:__ 20 de agosto 2020 * __Publicado:__ 07 de septiembre de 2020

I. Máster en Ciencias Matemáticas Mención Probabilidades y Estadísticas, Licenciado en Educación en la Especialidad de Matemática Computación, Investigador Independiente, Ecuador.
II. Investigador Independiente, Ecuador.

Resumen

Se presentan diversos métodos asociados al análisis de datos categóricos. Se hace una extensa descripción de las posibilidades de vinculación entre ellos, especialmente del Análisis de las Correspondencias con los demás.

Palabras Claves: Análisis de las correspondencias; Análisis de las frecuencias de configuraciones; análisis de clases latentes.

Abstract

Various methods associated with the analysis of categorical data are presented. An extensive description is made of the possibilities of linking them, especially the Analysis of Correspondences with others.

Keywords: Analysis of the correspondences; Analysis of the frequencies of configurations; latent class analysis.

Resumo

Vários métodos associados à análise de dados categóricos são apresentados. É feita uma extensa descrição das possibilidades de vinculá-los, especialmente a Análise de Correspondências com outras.

Palavras-chave: Análise das correspondências; Análise das frequências das configurações; análise de classe latente.

Introducción

El análisis de datos categóricos ha experimentado un extraordinario desarrollo en los últimos años. El creciente proceso de matematización de las investigaciones en ciencias sociales, así como el establecimiento de formas típicas de análisis para datos provenientes de la psicología y la medicina, ha hecho que muchos estadísticos se hayan vuelto hacia el trabajo en este campo, que, hasta hace muy poco, era un coto casi exclusivo de especialistas de otras disciplinas con intereses y conocimientos suficientes como para embarcarse en este tipo de estudio.

Lo anterior se refleja en la gran cantidad de artículos sobre la temática del uso de la estadística en las ciencias sociales que aparece en publicaciones de esas disciplinas, mientras que la mayoría de las revistas de estadística han ignorado todo lo que se ha venido realizando en temáticas que son, por su naturaleza, propias de ella.

Un ejemplo de que los estadísticos han comenzado a trabajar en serio en las temáticas de datos categóricos lo constituye el libro Statistics for the 21st Century (2000), editado por C. R. Rao y GáborJ. Székely, que incluye un conjunto de artículos por renombrados autores.

En este trabajo se pretende hacer una recopilación de los aspectos fundamentales de trabajo con datos categóricos desde el punto de vista de la mejora en la interpretación, de modo que se logre una potenciación de la capacidad de comprensión del analista. Los artículos están dispersos, como se dijo anteriormente, en muchas revistas de otras especialidades; en algunos casos, la terminología es propia y se hace necesario su identificación con los conceptos estadísticos reconocidos. Sin ánimo de que este trabajo sea un "estado del arte", sí se quiso que los interesados tuvieran reunidos los puntos esenciales del desarrollo más reciente y dar, de una forma preliminar, la opinión de los autores acerca del trabajo con los métodos propuestos.

Representación de datos categóricos

En esta sección se presentan las dos formas habituales en que se representan los datos categóricos.

Tablas de contingencia

La forma usual de presentación de datos categóricos es a través de una tabla de contingencia. Con vistas a una mejor comprensión ésta se definirá mediante un ejemplo.

Sea un grupo de individuos a los cuales se les miden dos atributos: el color de los ojos (categorías: Azules, Verdes y Pardos) y tamaño de los mismos (Categorías: normal o grande).

Tabla 1: Ejemplo de tabla de contingencia

Tamaño/Color	Azules	Verdes	Pardos	Total
Normal	10	5	20	**35**
Grande	3	8	6	**17**
Total	**13**	**13**	**26**	**52**

Forma matricial

Dada la tabla de contingencia, se puede definir las categorías de las variables que la conforman como variables mudas (dummy). A las cuales se les da valor 1 si la categoría está presente y 0 si no. Al hacer este procedimiento se convierte esta tabla en una matriz, de modo que en las columnas tenemos las categorías de las variables de la tabla y en las filas, a los individuos de la muestra.

Esta matriz se puede ver como una matriz particionada por columnas. Por ejemplo, en el caso de una tabla de contingencia de doble entrada, se define dos conjuntos de variables.

Para la mejor comprensión de lo que se acaba de explicar, se utilizará nuevamente la tabla (1). Por ejemplo:

Los (10) sujetos que tienen tamaño normal y ojos azules (1 0 0 1 0)

Los (5) sujetos que tienen tamaño normal y ojos verdes (0 1 0 1 0)

Los (20) sujetos que tienen tamaño normal y ojos pardos (0 0 1 1 0)

así sucesivamente.

De este modo, se obtienen 6 perfiles de respuestas y el tamaño de la matriz quedaría

$$\begin{bmatrix} 10010 \\ \vdots \ddots \ddots \ddots \vdots \\ 00101 \end{bmatrix} 52x5$$

En este trabajo se le llamará a este tipo de matrices indicadoras.

Algunos métodos asociados al análisis de tablas de contingencia

El análisis de las correspondencias (AC)

El análisis de las correspondencias (Benzécri, 1973; Greenacre, 1984) es una técnica estadística que se utiliza para representar, desde un punto de vista gráfico, las relaciones de dependencia e independencia de un conjunto de variables categóricas a partir de los datos de una tabla de contingencia. Existen dos tipos de análisis de correspondencias:

Simple: cuando se trabaja con 2 dimensiones.

Múltiple: cuando se trabaja con más de 2 dimensiones.

A continuación, se explicará cada uno de ellos.

Análisis de las correspondencias simples (ACS)

Sean A y B variables categóricas. Se desea analizar la asociación entre ellas. En lo que sigue se utilizará la representación matricial de las tablas de contingencia.

Las respuestas a las preguntas de las 2 variables (A y B) se codifican en las matrices indicadoras Z_1 y Z_2 respectivamente, cuyas columnas son variables dummy. La tabla de contingencia (1) no es más que el producto $Z_1^T Z_2$ de las matrices indicadoras.

Sean los vectores S_1 y S_2 que contienen los valores propuestos para las categorías de las dos variables. A partir de aquí, queda claro que los vectores Z_1S_1 y Z_2S_2 contienen las respuestas individuales cuantificadas.

La media centrada de las respuestas cuantificadas se puede escribir como:

$$1^T Z_1 S_1 = 1^T Z_2 S_2 = 0$$

Además, la covarianza S12 entre las dos variables y las varianzas v_1^2 y v_2^2 se obtienen;

$$S_{12} = \left(\frac{1}{n}\right) s_1^T Z_1^T Z_2 s_2 = s_1^T P_{12} s_2$$

$$v_1^2 = \left(\frac{1}{n}\right) s_1^T Z_1^T Z_1 s_1 = s_1^T D_1 s_1 \qquad \text{y} \qquad v_2^2 = \left(\frac{1}{n}\right) s_2^T Z_2^T Z_2 s_2 = s_2^T D_2 s_2$$

Donde $P_{12} = \left(\frac{1}{n}\right) Z_1^T Z_2$ es la matriz de correspondencia que contiene las frecuencias relativas y D_1 y D_2 son las matrices diagonales de las frecuencias relativas marginales (masas) de las dos variables. El coeficiente de correlación es igual a:

$$r = \frac{S_{12}}{v_1 v_2} = \frac{s_1^T P_{12} s_2}{\sqrt{s_1^T D_1 s_1 s_2^T D_2 s_2}}$$

La ecuación anterior se obtuvo para los valores dados a los atributos. Sin embargo, se puede suponer que estos valores son desconocidos y se transforma el problema en la búsqueda de las escalas que maximizan la correlación. Está claro que ese problema no es más que el objetivo del análisis de las correlaciones canónicas (ACC) . En la sección siguiente se presentará el desarrollo del ACS a partir de esta noción.

A la luz de las consideraciones anteriores, se incorporan al problema las condiciones de identificación propias del ACC que son: utilizar las variables estandarizadas (media cero y varianza 1).

$$\left(\frac{1}{n}\right) 1^T Z_1 s_1 = \left(\frac{1}{n}\right) 1^T Z_2 s_2 = 0 \qquad \text{y} \qquad s_1^T D_1 s_1 = s_2^T D_2 s_2 = 1$$

con estas condiciones se muestra que la solución óptima coincide con las coordenadas estándar de las categorías de respuestas sobre la primera dimensión principal del (ACS) de la tabla original.

El análisis de las correspondencias simples con un enfoque de correlaciones canónicas

Goodman (2000) hace una presentación abarcadora de este enfoque y en lo que sigue se tratarán sus aspectos fundamentales. Este trabajo está restringido a las tablas de contingencia de doble entrada.

Sea la tabla I x J, para cada casilla se cumple que:

$$\pi_{ij} = \pi_{i.}\pi_{.j} \qquad (3.1)$$

donde π_i y π_j son las distribuciones marginales de las filas y las columnas.

La ecuación (3.1) es el modelo con independencia estadística, entre las clasificaciones de las filas y las clasificaciones de las columnas en la tabla de contingencia.

¿Cómo analizar la dependencia? Goodman (2000) propone el siguiente esquema:

$$\pi_{ij} = \pi_{i.}\pi_{.j}\left(1 + \sum_{m=1}^{M} \rho_m x_{im} y_{jm}\right) \qquad (3.2)$$

Donde M= min (I,J) - 1, y los puntajes de las filas x_{im} (m=l, …, M) y los de las columnas y_{jm}, (m= 1, ..., M) son los coeficientes de las combinaciones lineales que satisfacen las siguientes condiciones:

$$\left.\begin{array}{ll} \sum_{i=1}^{I} x_{im} P_{i.} = 0, & \sum_{j=1}^{J} y_{jm} P_{.j} = 0 \\ \sum_{i=1}^{I} x_{im} P^2{}_{i.} = 0, & \sum_{j=1}^{J} y^2{}_{jm} P_{.j} = 0 \\ \sum_{i=1}^{I} x_{im} x_{im'} P_{i.} = 0, & \sum_{j=1}^{J} y_{jm} y_{jm'} P_{.j} = 0 \end{array}\right\} \qquad (3.2)$$

para m = m'. Los coeficientes x_{im} y y_{jm} en el modelo (3.2) son los puntajes estandarizados para las categorías de las filas (i = 1,...,I) y las categorías de las columnas (j = 1,...,J), respectivamente, correspondientes a la m-ésima componente (m = 1,...,M) en el término derecho del modelo (3.2). Los puntajes de filas diferentes están incorrelacionados; lo mismo ocurre con los de las columnas. El parámetro en el modelo (3.2) es la medida de la correlación entre los puntajes de las filas y las columnas (x_{im} y y_{jm}), que se calcula.

$$\sum_{i=1}^{I}\sum_{j=1}^{J} x_{im}\, y_{jm}\, P_{ij} = \rho_m \qquad (3.3)$$

para m = 1, …, M, en el modelo (3.2). Los parámetros de la correlación se ordenan como $1 \geq \rho_1 \geq \rho_2 \geq \cdots \geq \rho_m \geq 0$

Los puntajes de las filas y las columnas, x_{i1}, y_{j1}, en el modelo (3.2) son los puntajes estandarizados que maximizan la correlación ρ_1 y así sucesivamente.

El modelo (3.2) es una representación de la asociación en tablas de contingencias por la vía del análisis de correlaciones canónicas. En este contexto, el coeficiente ρ_1, es la primera correlación canónica entre las combinaciones lineales $\sum_i x_{i1}X_i$ y $\sum_j x_{j1}Y_j$, donde x_{i1} y y_{j1} son los coeficientes asociados a la correlación canónica y X_i y Y_j son las variables mudas de la matriz construida en el sentido del acápite anterior.

Definición de una medida de no independencia

Goodman (2000) define λ_{ij}, como la contingencia de Pearson, de la forma siguiente:

$$\lambda_{ij} = \frac{(P_{ij} - P_{i.}P_{.j})}{(P_{i.}P_{.j})} \qquad (3.4)$$

El modelo (1.4) satisface las siguientes condiciones:

$$\sum_{i=1}^{I}\lambda_{ij}P_{i.} = 0 \;\; (para\; j = 1, \ldots, J), \quad \sum_{j=1}^{J}\lambda_{ij}P_{ij} = 0 \;\; (para\; i = 1, \ldots, I) \qquad (3.4)$$

Si se despeja el término $\sum_{m=1}^{M}\rho_m x_{im}y_{jm}$ en (3,2) se llega a:

$$\lambda_{ij} = \sum_{m=1}^{M}\rho_m x_{im}y_{jm} \quad (3.5)$$

Con lo que se ve que la contingencia de Pearson mide también la asociación entre las variables de la tabla de contingencia.

Relación con características básicas del ACS

Goodman (2000) define la contingencia cuadrática media como una medida global de la asociación en una tabla de contingencia de la forma siguiente:

$$\lambda^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \lambda_{ij}^2 P_{i.} P_{.j} \qquad (3.6)$$

A partir de esta definición y con la aplicación de las propiedades de (3.2) se tiene:

$$\lambda^2 = \sum_{m=1}^{M} \rho_m^2 \qquad (3.7)$$

Esta relación es fundamental para la justificación del enfoque, porque (3.7) no es más que la inercia total del análisis de las correspondencias.

De igual forma, siguiendo las ideas de van der Heijden et al. (1999) se tiene al sumar los elementos de la tabla de contingencia según (3.2):

$$\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(\pi_{ij} - \pi_{i.}\pi_{.j}\right)^2}{\pi_{i.}\pi_{.j}} = \sum_{m=1}^{M-1} \rho_m^2 \qquad (3.8)$$

Que establece una importante relación entre la X2 Y la inercia total asociada al análisis de las correspondencias.

El Análisis de las Correspondencias Múltiples (ACM)

Después de haber visto lo anterior puede pasarse ahora a una posible generalización del AC al caso multivariado.

Sean P variables categóricas, cuantificadas en Z_1, Z_2, ... ,Z_p matrices indicadoras. El problema consiste en buscar los valores escalas s_1, s_2, ... , s_p para las variables de modo que se maximicen un conjunto de medidas de correlación.

Al igual que en el ACS (2 variables) , la medida seleccionada es la suma de correlaciones al cuadrado de las puntuaciones individuales Z_1s_1, Z_2s_2, … , Z_ps_p con la suma de puntajes Z_i, donde Z y s son las concatenaciones de Z q' s y s q' s respectivamente.

Pol. Con. (Edición núm. 49) Vol. 5, No 09, septiembre 2020, pp. 949-968, ISSN: 2550 - 682X

$S^TD\ s = l$ es la identificación general.

$$D = \left(\frac{1}{P}\right) diag(D_1, D_2, \dots, D_p)$$

la varianza individual $S^T_p\, D_p\, s_p$ no necesariamente es 1 en la solución final como en el caso de P= 2

Para alcanzar la solución se tienen dos vías:

1. AC a la matriz súper indicadora Z = [Z, …, Z_p]. Las dimensiones de esta matriz son n x J , donde $J = \sum_{p=1}^{P} J_P$

La matriz de correspondencias es $\left(\frac{1}{P_n}\right) Z$, la matriz de las masas de las filas es $\left(\frac{1}{n}\right) I$ y la matriz columna de las masas es D.

Luego la DVS para calcular la solución del AC de Z (no centrada) es:

$$\sqrt{n}\frac{Z}{P_n} D^{-1/2} = UGV^T \; donde \; U^TU = V^TV = I \quad (1.9)$$

Análisis centrado:

$$\sqrt{n}\left(\frac{Z}{P_n} - \frac{1}{n} 11^T D\right) D^{-1/2}$$

donde $\left(\frac{1}{n}\right)$ 1 es el vector de las masas de filas y 1^TD es el vector de las masas de las columnas de la matriz indicadora denotado por C^Ten el ACS).

2. AC aplicado a la matriz dé Burt (Benzécri, 1992). Para la definición de la matriz de Burt se trabajará con 3 variables categóricas con I, J y K categorías, respectivamente. La generalización a más variables es inmediata. La matriz de Burt tiene la forma siguiente:
    - Los bloques diagonales son matrices diagonales, uno para cada variable; sus elementos en la diagonal son del tipo $\pi_{i..}, \pi_{.j.}, \pi_{..k}$

- Los bloques no diagonales son matrices que contiene las sumas marginales de dos variables cada vez, por ejemplo, para el caso de I y J:

$$\begin{pmatrix} \pi_{11.} & \cdots & \pi_{1J.} \\ \vdots & \ddots & \vdots \\ \pi_{i1.} & \cdots & \pi_{IJ.} \end{pmatrix}_{IxJ}$$

La matriz de Burt es cuadrada y su dimensión es J x J . La aplicación- del ACS a la matriz de Burt produce una descomposición de cada una de las submatrices:

$$\pi_{ij.} = \pi_{i..}\pi_{.j.}\left(1 + \sum_{m=1}^{M-1} \rho_m x_{im} y_{jm}\right)$$

$$\pi_{i.k} = \pi_{i..}\pi_{..k}\left(1 + \sum_{m=1}^{M-1} \rho_m x_{im} z_{km}\right)$$

$$\pi_{.jk} = \pi_{.j.}\pi_{..k}\left(1 + \sum_{m=1}^{M-1} \rho_m y_{jm} z_{km}\right)$$

donde

$$\rho \_m>0$$

$$\sum_{i=1}^{I} \pi_{i..} x_{im} = \sum_{j=1}^{J} \pi_{.j.} y_{jm} = \sum_{k=1}^{K} \pi_{..k} z_{km} = 0$$

$$\sum_{i=1}^{I} \pi_{i..} x_{im} x_{im'} + \sum_{j=1}^{J} \pi_{.j.} y_{jm} y_{jm'} + \sum_{k=1}^{K} \pi_{..k} z_{km} z_{km'} = \delta^{mm'}$$

(La simultaneidad se aprecia en . la aparición reiterada de las coordenadas).

El análisis de las correspondencias múltiples con un enfoque de correlaciones canónicas

En analogía a lo que se presentó con respecto al ACS, también existe una manera de definir el ACM mediante una generalización del ACC. En lo que sigue se presenta a grandes rasgos las ideas de Tenenhaus & Young (1985).

La idea básica es la aplicación de la generalización del análisis de correlaciones canónicas (Horst, 1961). Para ello se considera cada partición de la matriz de variables dummy como una submatriz y se plantea maximizar la suma de las correlaciones al cuadrado entre los datos rescalados y el escalamiento de los sujetos. Esto es:

$$Max\ \frac{1}{P}\sum_{h=1}^{m}\sum_{j=1}^{p} cor^2(Z_j\phi_j^h, \psi^h)$$

donde Z_j es la submatriz indicadora correspondiente a la variable categórica j. ϕ^h, h=1, … , m; normalizados (es el valor del vector dividido entre la raíz cuadrada del valor propio correspondiente) e incorrelacionados y ψ^h, h=1, … , m; normalizados e incorrelacionados.

En Tenenhaus & Young (1985) se demuestra que la solución óptima del problema anterior es precisamente, los factores del análisis de las correspondencias múltiples y que el valor en el óptimo es:

$$\sum_{h=1}^{m}\lambda_h$$

En ese mismo trabajo, Tenenhaus & Young (1985) demuestran que el análisis de las correspondencias múltiples de variables binarias es equivalente al análisis de componentes principales de las matrices indicadoras normalizadas.

Este resultado es importante desde el punto de vista de la obtención de los valores de los factores, ya que el análisis de componentes principales es de fácil realización.

El análisis de clases latentes

El análisis de clases latentes (Lazarsfeld, 1950; Lazarsfeld y Henry, 1968) es un método de análisis factorial, cuya característica más importante es que la variable latente es nominal u ordinal, por lo que su efecto es el dé clasificar los individuos en clases. A continuación, se dan los elementos esenciales del método sobre la base del análisis para una tabla de doble entrada. La generalización a más categorías no ofrece ninguna dificultad.

Considérese una tabla de contingencia H con I filas y J columnas, de modo que en sus casillas se tengan las frecuencias relativas. Esto es:

$$\Pi_{ixj} = (\pi_{ij})$$

Si se considera válido el modelo de T clases latentes para una tabla de contingencia, se puede escribir:

$$\pi_{ij} = \sum_{t=1}^{T} \pi_t^Q\ \pi_{it}^{\bar{A}\,Q}\ \pi_{jt}^{\bar{B}\,Q} \qquad (3.10)$$

donde π_t^Q, es la probabilidad de que una observación caiga en la clase latente t (también se le llama el tamaño de la clase t); $\pi_{it}^{\bar{A}\,Q}$ y $\pi_{jt}^{\bar{B}\,Q}$son las probabilidades condicionales que dan la probabilidad de estar en la categoría i o j respectivamente dado que la observación está en la clase latente t.

La generalización del ACL a más de dos variables manifiestas es directa. La exposición de la misma se hará con el ejemplo de 3 variables.

El ACL para tres clases no es más que:

$$\pi_{ijk} = \sum_{t=1}^{T} \pi_t^Q \; \pi_{it}^{\bar{A}\,Q} \; \pi_{jt}^{\bar{B}\,Q} \; \pi_{kt}^{\bar{C}\,Q} \qquad (3.11)$$

y las restricciones:

$$\sum_{t=1}^{T} \pi_t^Q = \sum_{i=1}^{I} \pi_{it}^{\bar{A}\,Q} = \sum_{j=1}^{J} \pi_{jt}^{\bar{B}\,Q} = \sum_{k=1}^{K} \pi_{kt}^{\bar{C}\,Q}$$

con A, B y C las variables manifiestas y Q la variable latente.

De igual forma que en el caso bivariado, en este también se pueden rescalar los parámetros con la misma interpretación que ya se analizó.

El análisis de las frecuencias de las configuraciones

El análisis de las frecuencias de las configuraciones (AFC) (Lienert, 1969, Krauth & Lienert, 1973) es un método en el que se buscan las combinaciones de rasgos o síntomas que se presentan con una mayor frecuencia que la esperada. El concepto básico dentro del método lo constituye el "tipo", esto es: la casilla de la tabla de contingencia que muestra una frecuencia significativamente mayor que la esperada.

Para lograr una comprensión más íntegra del método, se explicará a partir de una tabla de contingencia de cuatro entradas; la generalización a mayores dimensiones es inmediata.

Sea Π una tabla de contingencia de cuatro categorías: A, B, C y D, con I, J, K, L atributos, respectivamente. Las I x J x K x L configuraciones se. denotan por {i, j, k, L} con i = 1,..., I; j = 1,...,J; k = 1,...,K y L = 1,...,L. La probabilidad asociada a cada configuración se denota por $\pi_{ijkl.}$ Aquí es necesario precisar el término "esperado" ya que es esencial para la cabal comprensión del concepto fundamental de "tipo".

Pol. Con. (Edición núm. 49) Vol. 5, No 09, septiembre 2020, pp. 949-968, ISSN: 2550 - 682X

Se dice que una cierta configuración {i, j, k, L} es un tipo si se cumple que $\pi_{ijkl} = \pi_{i...}\pi_{.j..}\pi_{..k.}\pi_{...l}$ manera de probar lo anterior es a través del contraste de la hipótesis de independencia local.

Vinculación de métodos

La vinculación de los métodos se hace a través del análisis de las correspondencias en sus dos variantes: simple y múltiple. La idea básica es encontrar la semejanza del análisis de clases latentes y el análisis de las frecuencias de las configuraciones, respectivamente, con estas variantes del AC. En esta sección se siguen las ideas generales dadas por van der Heijden et al (1999).

El análisis de las frecuencias de las configuraciones y el análisis de las correspondencias

La relación entre el AFC y el AC, en general, no presenta ninguna dificultad. Una vez determinados los tipos según la forma usual, se pasa al análisis de las correspondencias múltiples. En este sentido se tienen en cuenta dos aspectos:

- El análisis usual de las categorías de las variables implicadas a partir del gráfico (bidimensional o tridimensional, según sea el caso)
- El análisis de las configuraciones, que no es más que la transformación de las casillas al espacio factorial determinado por el ACM.

El análisis de clases latentes y el análisis de las correspondencias simple

En la sección anterior se presentó el análisis de las correspondencias simples con el enfoque del análisis de las correlaciones canónicas. En esta se verá qué, bajo ciertas condiciones, existe semejanza entre este y el ACL.

Tanto el ACS como el ACL pueden considerarse como métodos que dan una descomposición de rango reducido de la matriz, en el sentido siguiente: si se considera Π como una matriz, ambos métodos lo que persiguen es dar una representación reducida de la matriz.

Por ejemplo, en el caso presentado anteriormente, (3.10) define una matriz de rango R. Se pueden ver los siguientes casos:

(a) Si R = min(I, J) , se tiene que Π es de rango completo y el modelo propuesto es el saturado.

(b) Si R = 1, se tiene el modelo de independencia

(c) Si 1 < R < min(I, J) , (3.10) nos da una matriz de rango reducido

Con vistas al establecimiento de la relación entre ambos métodos, es necesario aplicar una transformación (Goodman, 1974) que da pie, en el marco de las ciencias sociales, a lo que se

conoce por el nombre de análisis de presupuesto latente. La transformación se da a continuación:

$$\frac{\pi_{ij}}{\pi_{i.}}\sum_{t=1}^{T}\pi_{it}^{A\,\bar{Q}}\,\pi_{jt}^{\bar{B}\,Q} \quad con \quad \pi_{it}^{A\,\bar{Q}}=\frac{\pi_{t}^{Q}\pi_{it}^{\bar{A}\,Q}}{\pi_{i.}}$$

Esta transformación juega un papel importante en el establecimiento de la vinculación entre los dos análisis, por eso a continuación se presentan algunas de sus características más importantes: Es conocido que los modelos de análisis de clases latentes se pueden presentar de varias maneras y que una de las más conocidas es el enfoque a través de modelos loglineales. En este contexto, se tiene que las variables manifiestas A y B son independientes condicionadas por Q, debido a las reglas de colapsibilidad para modelos loglineales (véase Agresti, 2002). De las diversas formas de estudiar la dependencia para tablas de 3 entradas, la más conveniente para el fin que se persigue es la siguiente:

- Si se define independencia para elementos del tipo π_{ij} [1] como sigue: $\frac{\pi_{i.t}}{\pi_{..t}}=\pi_{i..}$, entonces es fácil estudiarla mediante la comparación de las condicionales con las marginales, antes expuestas, ya que es fácil ver que el miembro izquierdo no es más que $\pi_{it}^{A\,\bar{Q}}$. Esto es lo usual en el ACL.
- Con lo anterior queda claro que los parámetros rescalados tienen una interpretación en el sentido de la masa de la categoría i que corresponde a la clase t.

Una vez visto que tiene sentido trabajar con los parámetros rescalados, se presenta a continuación la forma en que se produce la vinculación entre el ACS y el ACL.

Es conocido (Benzécri, 1973, Greenacre, 1991) que el ACS se tiene siempre que el número de factores es igual al rango de la matriz Π, con lo que se logra una descomposición de ésta. A partir de aquí se puede establecer una relación con lo que se mencionó anteriormente acerca de los modelos de clases latentes y el rango de Π:

(a) El número de factores del ACS coincide con el de clases latentes

(b) El número de factores del ACS es 1 y el ACL =1

(c) El número de factores será igual al rango, pero no siempre se logrará el mismo número de clases latentes

Además, de Leeuw & van der Heijden (1991) demuestran un caso adicional:

[1] En Mood et al. (1974) se dan las tres formas de definir independencia para elementos del tipo marginal en tablas de 3 entradas.

(d) Si R = 2, ACS implica ACL y los modelos son equivalentes.

Por lo que se aprecia en los puntos anteriores, existe coincidencia entre ambos métodos en algunos casos, siempre que se utilice para estimar los parámetros el mismo método de estimación. Si se supone que las estimaciones de ambos métodos son iguales se encuentra la siguiente relación muy interesante para los fines del análisis de datos:

$$\pi_{ij} = \sum_{t=1}^{T} \pi_t^Q \, \pi_{it}^{\bar{A}\,Q} \pi_{jt}^{\bar{B}\,Q} = \pi_{i.}\pi_{.j} \sum_{t=1}^{T} \pi_{it}^{A\,\bar{Q}} \, \pi_{jt}^{B\,\bar{Q}} \left(\pi_t^Q\right)^{-1} \quad (4.1)$$

Al comparar (3.2) con (4.1) se tiene la siguiente relación:

$$\left(1 + \sum_{m=1}^{M} \rho_m \, x_{im} y_{jm}\right) = \pi_{i.}\pi_{.j} \sum_{t=1}^{T} \pi_{it}^{A\,\bar{Q}} \, \pi_{jt}^{B\,\bar{Q}} \left(\pi_t^Q\right)^{-1} \quad (4.2)$$

De (4.2) se aprecia una relación lineal entre ambas expresiones, esto es: Existen matrices F y G de transformación con dimensiones R x R que dan lo siguiente:

- Sea X, de dimensión I x R, la matriz que contiene los X_{im} más la primera columna de 1; sea igualmente Y, de dimensión J x R, la matriz que contiene los y_{jm} más la primera columna igual a 1. Se tiene entonces

$$X = \Pi_i \, F$$
$$Y = \Pi_j \, G$$

donde Π_i y Π_j son matrices de I x R y J x R, respectivamente, tales que: $\Pi_i = \left(\pi_{it}^{A\,\bar{Q}}\right)$ y $\Pi_j = \left(\pi_{jt}^{B\,\bar{Q}}\right)$.

Claro que esta relación sólo se cumple bajo la hipótesis de igualdad de los parámetros ajustados.

El análisis de clases latentes y el análisis de las correspondencias múltiples

La relación el ACL y el ACM se hace muy claro cuando (3.11) se suma en i, j y k, respectivamente. Una vez hecho esto, se obtiene:

$$\pi_{ij.} = \sum_{t=1}^{T} \pi_t^Q \pi_{it}^{\bar{A}\,Q} \pi_{jt}^{\bar{B}\,Q}$$
$$\pi_{i.k} = \sum_{t=1}^{T} \pi_t^Q \pi_{it}^{\bar{A}\,Q} \pi_{kt}^{\bar{C}\,Q}$$
$$\pi_{.jk} = \sum_{t=1}^{T} \pi_t^Q \pi_{jt}^{\bar{B}\,Q} \pi_{kt}^{\bar{C}\,Q}$$

Como se ve, el ACL da tres matrices con márgenes bivariados de rango reducido T, además, estas ecuaciones tienen parámetros en común. De aquí es fácil establecer relaciones semejantes al caso bivariado. Si bien el procedimiento es muy semejante, no se puede llegar al establecimiento de la relación que se obtuvo para rango 2, ya que el recíproco no se cumple, esto es: varias tablas de doble entrada no implican una tabla de orden superior.

De igual forma que antes, se puede poner:

$$1 + \sum_{m=1}^{M-1} \rho_m x_{im} y_{jm} = \sum_{t=1}^{T} \pi_{it}^{A\,\bar{Q}} \pi_{jt}^{B\,\bar{Q}} \left(\pi_t^Q\right)^{-1}$$
$$1 + \sum_{m=1}^{M-1} \rho_m x_{im} z_{km} = \sum_{t=1}^{T} \pi_{it}^{A\,\bar{Q}} \pi_{kt}^{C\,\bar{Q}} \left(\pi_t^Q\right)^{-1}$$
$$1 + \sum_{m=1}^{M-1} \rho_m y_{jm} z_{km} = \sum_{t=1}^{T} \pi_{it}^{A\,\bar{Q}} \pi_{jt}^{B\,\bar{Q}} \left(\pi_t^Q\right)^{-1}$$

Nuevamente se tiene la validez a partir de la repetición de los elementos. Si se supone que existe algún caso en el que se da la igualdad, como se hizo con el bivariado, se puede llegar a una expresión matricial con las correspondientes matrices de transformación.

Representación gráfica

Precisamente en la representación gráfica es donde mejor se aprecia la vinculación de los métodos que se explicó en el acápite anterior.

Básicamente existen tres formas de representación, a saber:

- Gráfico unidimensional
- Gráfico bidimensional (Scatterplot)

Pol. Con. (Edición núm. 49) Vol. 5, No 09, septiembre 2020, pp. 949-968, ISSN: 2550 - 682X

- Gráfico ternario

Los dos primeros no necesitan explicación, ya que son los más comúnmente usados en Estadística. El tercero, aunque es menos conocido, tampoco es exclusivo de los métodos que se analizan en el presente trabajo, ya que son la forma clásica para la representación de los resultados en las superficies de respuestas para modelos de mezcla (Montgomery, 1991). En el contexto del análisis de clases latentes, este gráfico se utiliza' para la representación de los parámetros rescalados que, al sumar uno, pueden ubicarse dentro de un simplex.

Realmente, el AC es el método que tiene como una parte consustancial un gráfico. De ahí que en lo que sigue se tratarán en detalle las características del mismo.

En el AC se distinguen dos tipos de gráficos:

- **Gráfico asimétrico:** En el caso del ACS, las filas y las columnas se presentan de forma independiente, con escalas distintas. En el ACM, se representan los primeros factores también de manera independiente.
- **Gráfico simétrico:** Para el ACS, las filas y las columnas se reproducen sobre el mismo gráfico, con una misma escala. De igual forma se procede con el ACM, en el que se presentan por pares.

Como el objetivo de este trabajo es la vinculación con los otros métodos de análisis de datos categóricos expuestos anteriormente, se continuará sólo con el simétrico, ya que el asimétrico no es útil para estos fines.

Greenacre (2006) considera el gráfico simétrico como una opción conveniente debido a que la representación de los puntos de filas y columnas se hace con la misma escala. Esta forma es conocida también como el "escalamiento francés o de Benzécri" y es el preferido por la escuela francesa de análisis de datos.

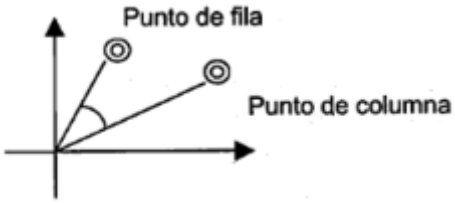
En este contexto se proponen las siguientes reglas de análisis:

Para interpretar el gráfico, se deben considerar las posiciones relativas a un eje de puntos perteneciente a la misma nube. Dos puntos cercanos en el gráfico tendrán un perfil similar.

Interpretación angular entre los puntos pertenecientes a la nube diferente.

Se puede interpretar el ángulo entre los puntos de las filas y las columnas tomando el origen de coordenadas como el vértice.

Figura 1: Relación angular en el ACS



Siguiendo algunas reglas:

a) Si el ángulo entre los puntos es agudo (< 90°) la correlación entre las dos características, es alta.

b) Si el ángulo entre los puntos es obtuso (> 90°) la correlación entre las dos características, es baja (o negativa).

c) Si el ángulo es recto, los puntos no interactúan o no hay correlación entre ellos.

Contrariamente a lo expuesto por Greenacre y la escuela francesa, Goodman (2000) plantea en su exhaustivo trabajo sobre el análisis de tablas de contingencia de dos entradas, contenido en Statistics for the 21st Century, la necesidad de hacer un rescalamiento de las coordenadas para lograr una representación que dé una interpretación geométrica directa. Para ello, define una familia de transformaciones del tipo siguiente:

$$x_{im}^{*} = x_{im}\rho_m^{\gamma} = \frac{\rho_m x_{im}}{\rho_m^{\delta}} \qquad y_{jm}^{*} = y_{jm}\rho_m^{\delta} = \frac{\rho_m y_{jm}}{\rho_m^{\gamma}}$$

Donde γ+δ=1. En el caso de interés para el análisis de las correspondencias, se toman δ= γ=0.5 y se tiene una formulación simplificada del modelo de dependencia (3.2). De aquí se tiene que la contingencia de Pearson se puede expresar como el producto escalar de los dos vectores rescalados, con lo que se obtiene la posibilidad de una interpretación geométrica directa.

Conclusiones y recomendaciones

- La combinación de métodos para el análisis de datos categóricos mejora considerablemente las posibilidades de interpretación y son una ayuda eficaz para el analista.

- Si bien, en el trabajo se presentan algunos aspectos de cómo se realiza la vinculación entre el AC y el ACL, es necesario continuar con esa investigación, para lograr una comprensión más cabal de las relaciones entre ambos tipos de análisis.

Referencias

1. AGRESTI, A. (2002): Categorical Data Analysis (2nd Edition), Wiley, Nueva York

2. BENZÉCRI, J.-P et collaborateurs (1973) : L'Analyse des Donnés. L'Analyse des Correspondences, Dunoá, París

3. BENZÉCRI, J.-P. (1992): Correspondence Analysis Handbook, Marcel Dekker, Inc., Nueva York, 665 + xi pp.

4. FERNÁNDEZ, R. S. M. (2011). Análisis de correspondencias simples y múltiples. Universidad Autónoma de Madrid: Facultad de Ciencias Económicas y Empresariales.

5. GABRIEL, K. R. (1971): The biplot graphic display of matrices with application to principal component analysis, Biometrika 58(3), 453-467 pp.

6. GONZALEZ, D. A. (2006): Dos enfoques para el análisis de clases latentes ordinales. En: Revista de la Facultad de Matemática y Computación de la Universidad de La Habana. Cuba. Vol. 27, Núm. 1.

7. GONZALEZ, D. A. (2006): Algunas consideraciones prácticas acerca de la estimación de parámetros en el modelo clásico de clases latentes. En: Revista de la Facultad de Matemática y Computación de la Universidad de La Habana. Cuba. Vol. 27, Núm. 1.

8. GOODMAN, L.A. (1997): Statistical Methods, Graphical Displays, and Tukey's Ladder of Re-expression in the Analysis of Nonindependence in Contingency Tables: Correspondence. Analysis, Association Analysis, and the Midway View of Nonindependence, en: BRILLINGER, D., FERNHOLZ, L.T. & MORGENTHALER, S.: The Practice of Data Analysis: Essays in Honor of John W. Tukey, Princeton, Nueva Jersey, Princeton University Press, pp0. 101-132

9. GOODMAN, L. A. (2000) : The Analysis of Cross-Classified Data: Notes on a Century of Progress in Contingency Table Analysis, and Some Comments on Its Prehistory and Its Future, Marcel Dekker, Inc., New York, 231 + i pp.

10. GREENACRE, M. (1984): Theory and Applications of Correspondence Analysis, Academic Press, Londres

11. GREENACRE, M. (2006) : Tying up the loose ends in simple correspondence analysis, Economic Working Paper 940.

12. GREENACRE, M. (2005): From correspondence analysis to múltiple and joint correspondence analysis.

13. JAMBU, M. (1991) : Exploratory and Multivariate Data Analysis, Academic Press, Inc., Boston, 474 + xv pp.

14. LAUTSCH, E. y PLICHTA, M.M. (2003): Configural Frecuency Analysis (CFA), Múltiple Correspondence Analysis (MCA) and Latent Class Analysis (LCA): An empirical comparison, Psychology Science 45(2), 298-323 pp.

15. LAZARSFELD, P.F. (1950): The logical and mathematical foundation of latent structure analysis. En: STOUFFER, S. A.et al. (Eds.): The American Soldier, Vol. IV, Measurement and Prediction, Princeton

16. LAZARSFELD, P.F. y HENRY, N. W. (1968): Latent Structure Analysis, Houghton Mifflin, Boston

17. LEEUW, J. & Van der HEIJDEN, P.G.M. (1991): Reduced rank models for contingency tablas, Biometrika, 78, pp. 229-232

18. MOOD, A.M., GRAYBILL, R.A. y BOES, D.C. (1974): Introduction to the Theory of Statistics (3rd Edition), MacGraw Hill, Londres

19. MONTGOMERY, D, C. (1991) : Design and Analysis of Experiments, Third Edition, John Wiley and Sons, Nueva York, 649 + xvii pp.

20. RAO, C.R. y SZÉKELY, G.J. (Eds.) (2000) : Statistics for the 21st Century. Methodologies for Applications of the Future, Marcel Dekker, Inc., Nueva York

21. Van der HEIJDEN, GILULA, P. G. y van der ARK, L. A. (1999) : An extended study into the relationships between correspondence analysis and latent class analysis, 40 + vii pp.

22. VERMUNT, J. K. (1997): LEM 1.0: A general programa for the analysis of categorical data. Tilburg: Tilburg university Von EYE, A. y NIEDERMEIER, K.E. (1999) : Statistical Analysis of Longitudinal Categorical Data in the Social and Behavioral Sciences, Lawrence Er