Análisis Multivariado de Datos: Aproximación Didáctica

Giovanni **Sanabria** Brenes
Instituto Tecnológico de Costa Rica - Universidad de Costa Rica
Costa Rica
gsanabriab@yahoo.com

Félix **Núñez** Vanegas
Instituto Tecnológico de Costa Rica - Universidad de Costa Rica
Costa Rica
fnunez@itcr.ac.cr

Resumen

Con el objetivo de facilitar la comprensión de las principales técnicas del Análisis Multivariado de datos se diseñaron varias propuestas didácticas que abordaran el problema de su enseñanza, las cuales estuvieron basadas en una serie de situaciones problema y fueron validadas en un curso de capacitación dirigido a investigadores del Instituto Tecnológico de Costa Rica. Se espera que tal trabajo sirva de ayuda y referencia a aquellos investigadores que tengan que aplicar dichas técnicas y no cuentan con los conocimientos necesarios para tales efectos.

Palabras clave: Didáctica, multivariado, componentes principales, correspondencias múltiples, análisis discriminante, clasificación jerárquica.

1. Definición del problema

El auge que en los últimos años han tenido los ordenadores, ha provocado que las técnicas del análisis multivariado de datos sean, hoy por hoy, herramientas útiles en las investigaciones, tanto cuantitativas como cualitativas. Se aplican en disciplinas como la Sociología, Biología, Farmacia entre otras.

Dicho análisis podría revelar relaciones entre los datos que el investigador no había tomado en cuenta, lo cual implicaría proponer nuevas hipótesis y de pronto desechar otras.

No obstante, por su naturaleza, para algunos investigadores las técnicas del Análisis Multivariado de Datos no son fáciles de entender y más difícil todavía es su correcta aplicación.

Debido a lo anterior, se presentó ante la Vicerrectoría de Investigación y Extensión del Instituto Tecnológico de Costa Rica el proyecto de investigación "Estudio de métodos de análisis multivariado de datos" cuyo objetivo general fue desarrollar una metodología para abordar las

principales técnicas del Análisis de Datos desde una perspectiva didáctica tales como Análisis en Componentes Principales (ACP), Análisis Factorial de Correspondencias (AFC), Análisis Factorial de Correspondencias Múltiples (ACM), Análisis Factorial Discriminante (AFD) y Clasificación Jerárquica Ascendente (CJA). Dicho proyecto fue aprobado en agosto de 2007 y finalizó en junio de 2009.

Los resultados de dicha investigación están dirigidos a investigadores principiantes y al sector docente y estudiantil de las universidades con conocimientos básicos de estadística descriptiva (análisis univariado).

El presente documento describe el proceso utilizado en el desarrollo de la metodología, su validación y los principales resultados obtenidos.

2. Objetivos

- Que el investigador comprenda y aplique el Análisis en Componentes Principales para sintetizar y describir la relación entre las modalidades de dos variables cualitativas.
- Que el investigador comprenda y aplique el Análisis en Factorial de Correspondencias para sintetizar y describir la información de una tabla de datos individuos × dos variables cualitativas.
- Que el investigador comprenda y aplique el Análisis Factorial de Correspondencias Múltiples para sintetizar y describir la información de una tabla de datos individuos × variables cualitativas.
- Que el investigador comprenda y aplique el Análisis Factorial Discriminante para discriminar clases o grupos de individuos dados a priori por las modalidades de una variable cualitativa (variable a explicar) con la ayuda de variables cuantitativas (predictores), con el fin de describir los grupos y asignar nuevos individuos a alguno de los grupos establecidos.

3. Marco de referencia

3.1 Teorías didácticas

Como toda intención didáctica debe estar apoyada en concepciones sobre la enseñanza y el aprendizaje, nuestra propuesta de proponer una didáctica para el análisis multivariado de datos, tomó como base la Teoría de Situaciones, de Brousseau (1986) y los trabajos realizados por Batanero (2001) y Batanero & Godino (2001) para la didáctica de la estadística y el análisis de datos. El lector que lo desea puede consultar Núñez & Sanabria (2009) donde se justifica y desarrolla esta didáctica.

Las ideas expuestas en Núñez & Sanabria (2009) están plasmadas en el presente trabajo y vienen a ser un aporte significativo desde el polo del estudiante, aunque desde luego es también un esfuerzo para el profesor interesado en una metodología para el establecimiento de tales conocimientos, y por ello es un aporte desde el polo pedagógico.

3.2. Técnicas de análisis de datos consideradas

Seguidamente, se presenta un breve resumen de las técnicas consideradas. El lector interesado en una descripción más amplia puede consultar Sanabria & Núñez (2009) y si desea abordar el desarrollo teórico de estos métodos puede consultar Trejos (1998) y Trejos (2004).

Análisis en Componentes Principales (ACP)

Este análisis se utiliza en tablas de individuos × variables cuantitativas y consiste en hallar un número menor de variables nuevas (componentes principales) que conserve la mayor cantidad de información de los datos.

Las componentes principales establecen planos y círculos de correlaciones donde se proyectan respectivamente los individuos y las variables respectivamente. Estos gráficos son como radiografías de la estructura de los datos y permiten describir la información contenida en la tabla. Los programas computacionales del mercado tales como SPSS, PIMAD, entre otros, realizan el ACP a partir de la tabla de datos y permiten visualizar los diferentes planos y círculos de correlaciones, así como la cantidad de información (inercia) en estos gráficos.

Análisis Factorial de Correspondencias (AFC)

Este análisis permite establecer la relación entre las modalidades de dos variables cualitativas. El AFC es un Análisis en Componentes Principales muy particular aplicado a la tabla de perfiles fila (obtenida a partir de la tabla original).

Esto permite obtener, a partir de las componentes principales, planos donde se proyectan las modalidades de las variables y se debe entender que la cercanía de modalidades es vista como dependencia y no como correlación (concepto solo para variables cuantitativas). Los programas computacionales del mercado realizan el AFC a partir de la tabla de contingencia o la tabla original.

Análisis Factorial de Correspondencias Múltiples (ACM)

Esta técnica permite analizar los datos de una tabla de la forma: individuos × variables cualitativas, específicamente permite describir las relaciones entre las modalidades de las variables.

El ACM consiste en realizar un AFC muy particular sobre esta tabla de contingencia entre dos grandes variables cualitativas:

1. La variable individuos. Sus modalidades son los individuos

2. La variable modalidades. Sus modalidades son las modalidades de las variables originales

Esto permite obtener planos donde se proyectan las modalidades de las variables y se debe entender la cercanía de modalidades como dependencia y no como correlación (concepto solo para variables cuantitativas). Sobre estos planos se pueden proyectar los individuos para ver su relación con las modalidades. Los programas computacionales del mercado realizan el AFC a partir de la tabla original.

Análisis Factorial Discriminante (AFD)

Esta técnica de análisis de datos, en primera instancia, busca discriminar las clases o

grupos de individuos dados a priori por las modalidades de una variable cualitativa (variable a explicar) con la ayuda de variables cuantitativas (predictores). Cada clase de individuos es formada por todos aquellos que eligieron la misma modalidad de la variable cualitativa. Así la tabla de datos a analizar es de la forma individuos × (variables cualitativas- variable cualitativa).

Se pretende construir la caracterización de las clases de individuos con base en la información aportada por los predictores. Así se obtienen nuevas variables independientes, llamadas funciones discriminantes, las cuales son combinación lineal de las variables cuantitativas y discriminan lo mejor posible la separación de las clases de individuos y por ende las modalidades de la variable cualitativa. Los objetivos de este análisis son:

1. Objetivo descriptivo. Las funciones discriminantes describen la mejor separación de las clases de individuos determinadas por las modalidades de la variable cualitativa. Así, se pueden describir las características que potencian la separación.

2. Objetivo decisional. Si para un nuevo individuo se conoce el valor de los predictores, las funciones discriminantes ayudarán a predecir la modalidad de la variable cualitativa a la cual pertenece.

Clasificación Jerárquica Ascendente (CJA)

La Clasificación Jerárquica Ascendente, a partir de una tabla de datos de la forma individuos× variables busca reconocer subgrupos de individuos homogéneos de acuerdo a una medida de semejanza, estableciendo diferentes particiones del grupo original de individuos según el grado de semejanza, lo cual se puede representar mediante un árbol. Note que a diferencia del Análisis Discriminante, en este análisis no se tienen subgrupos a priori, sino se busca establecerlos. Las medidas de semejanza utilizadas por el CJA dependen del tipo de variables involucradas en el análisis.

4. Metodología

El desarrollo del proyecto siguió las siguientes etapas:

Etapa I: Estudio de las técnicas del Análisis Multivariado de Datos descritas anteriormente. Para cada técnica se estudió: justificación matemática, aplicación e interpretación de resultados.

Etapa II: Estudio de algunas propuestas para abordar el problema de la enseñanza – aprendizaje de la matemática en general:

- Teoría de Situaciones, Brousseau (1986)
- La Transposición Didáctica, Chevallard (1998)
- Teoría de Campos Conceptuales, Vérganud (1990)
- Didáctica de la estadística y del análisis de datos en particular, Propuestas desarrolladas por Batanero & Godino (2001)

Etapa III: Ejemplificación de las técnicas estudiadas. Con base en las sugerencias obtenidas en la etapa anterior, se definieron problemas (con fines didácticos) cuya solución requieren la aplicación de una técnica estudiada. Esta etapa conlleva, para cada técnica: definir uno o varios

problemas, recolectar datos para cada problema, analizar los datos recolectados e interpretarlos.

Etapa IV: Elaboración de una aproximación didáctica para la enseñanza del Análisis Multivariado de Datos. A partir de las etapas anteriores surgieron algunas consideraciones didácticas para la enseñanza del análisis multivariado de datos expuestas en Núñez & Sanabria (2009), las cuales son muy importantes en el desarrollo del proyecto.

Etapa V: Elaboración de una propuesta metodológica para abordar las principales técnicas del Análisis de Datos desde una perspectiva didáctica. Con base en una posición didáctica (etapa IV) y el saber "sabio" (etapa I) se establecieron los objetivos, conocimientos previos, pautas metodológicas y actividades a considerar en la enseñanza de cada una de las técnicas propuestas.

Etapa VI: Desarrollo de propuestas guías para la enseñanza de cada una de las técnicas consideradas a partir de la metodología establecida.

Etapa VII: Validación de la metodología planteada por medio de un curso de capacitación dirigido a investigadores del Instituto Tecnologíco de Costa Rica (ITCR, junio de 2009).

5. Resultados y conclusiones

La Teoría de Situaciones permitió plantear propuestas para abordar el problema de la enseñanza de las principales técnicas del análisis multivariado de datos: ACP, AFC, ACM, AFD y CJA. A través de la propuesta de varios grupos de situaciones problema se establecieron los principales conceptos.

El estudio realizado nos permitió hablar de una incipiente y necesaria Didáctica para el Análisis Multivariado de Datos en el I Encuentro de Didáctica de la Estadística, la Probabilidad y el Análisis de datos (Núñez & Sanabria, 2009). Esta primera aproximación es base de las propuestas realizadas.

Así, para cada una de las técnicas se elaboró una propuesta que contempló situaciones problema, en su mayoría contextualizadas a una tabla de datos, descripción de los conceptos involucrados, explicación teórica de las técnicas y el uso de software. Dichas propuestas han sido expuestas en diferentes congresos:

- "Una propuesta para la comprensión del Análisis en Componentes principales". I Encuentro Latinoamericano de Educación Estadística (ELEE), del 4 y 5 de Julio de 2008 en el Instituto Tecnológico y de Estudios Superiores de Monterrey, México.
- "Una propuesta para la compresión de las técnicas de análisis multivariado: Análisis Factorial de Correspondencias (AFC) y Análisis Factorial de Correspondencias Múltiples (ACM)". I Congreso Internacional de Computación y Matemática, del 21 al 23 de agosto de 2008.
- "Una propuesta para la comprensión de la Clasificación Jerárquica Ascendente (CJA)". Primer Encuentro Nacional en la Enseñanza de la Probabilidad y la Estadística (1° ENEPE), Puebla, México, del 16 al 18 de junio del 2010.
- "Una propuesta para la compresión del Análisis Factorial Discriminante (AFD)". II Congreso Internacional de Computación y Matemática, del 26, 27 y 28 de agosto del 2010.

En cada una de estas exposiciones, algunos miembros del público manifestaron su interés por la propuesta realizada y la necesidad de realizar trabajos similares sobre la Didáctica del Análisis Multivariado de Datos.

Por otra lado, cada una de las propuestas realizadas fue validada por un grupo de diez investigadores del ITCR que asistieron al curso "Introducción al Análisis Multivariado de Datos" que se impartió, los días 24, 25, 29 y 30 de junio del 2009. La metodología utilizada en este curso fue la elaborada en el proyecto de investigación y estaba dirigido a profesores del ITCR que hubieran desarrollado al menos un proyecto de investigación (aprobado por la Vicerrectoría de Investigación y Extensión del ITCR) y desearan tener conocimientos básicos de las principales técnicas de Análisis Multivariado de Datos

La participación activa en la clase resolviendo las situaciones problema que se plantearon en el desarrollo de la misma, nos indicaba la pertinencia o reformulación de tales situaciones, las cuales buscaban que los estudiantes investigadores al resolverlas, obtuvieran el concepto que pretendíamos establecer. A través de la observación por parte de los expositores, se detectaron algunas dificultades y aciertos en la resolución de las situaciones. Algunas dificultades fueron superadas y se consideran como parte medular del proceso de aprendizaje. Los indicadores utilizados para valores la pertinencia y reformulación de las distintas situaciones propuestas, fueron la comprensión del problema y las dificultades para resolverlos. La dificultad detectada en algunas situaciones problema por parte de los investigadores estudiantes, nos llevó a la tarea de dividir el problema en pequeños sub-problemas.

Por otro lado, los investigadores participantes consideraron importante que se desarrollen propuestas metodológicas para el abordaje del aprendizaje de las técnicas estudiadas. Además indicaron que la metodología empleada les permitió asimilar los conceptos básicos del curso.

Como dato, indicar que el resultado de la evaluación del desempeño del curso aplicada por el Departamento de Recursos Humanos del ITCR fue de 90, y en la siguiente tabla se detallan los rubros evaluados:

Tabla 1

Evaluación del desempeño del curso "Introducción al Análisis Multivariado de Datos"

Facilitador	91%
Material y equipo	93%
Practicas y evacuación	85%
Aspectos generales	90%
Promedio	90%

Los participantes externaron a su vez algunos comentarios sobre el curso dado con dicha metodología

- Excelentes instrucciones, muy clara las exposiciones y con mucha propiedad.
- Excelente curso, sin embargo se le debe dar más tiempo para profundizar.

- Mejorar el software empleado como el SPSS o MINITAB.

La principal limitación que presentó el proyecto fue el software utilizado. Dado que los costos de los programas computacionales especializados en análisis multivariado de datos son elevados, se optó por utilizar el software gratuito PIMAD 3.0 desarrollado por el Dr. Oldemar Rodríguez Rojas. Pese a que este programa en el manejo de los datos es poco flexible, no significó un obstáculo para desarrollar en su totalidad la metodología del proyecto y no tuvimos que lidiar con el problema de una licencia. Las propuestas desarrolladas son fácilmente adaptables a otros programas computacionales.

El trabajo desarrollado viene a ser un aporte ubicado en Didáctica del Análisis Multivariado de Datos, en gestación apenas, y permite una mayor comprensión de las técnicas consideradas, para que los académicos puedan utilizarlos en su quehacer profesional.

Referencias y bibliografía

Batanero, Carmen. 2001. Didáctica de la estadística. Grupo de Investigación y Educación Estadística, Departamento de Didáctica de la Matemática, Universidad de Granada. España. Servicio de Reprografía de la Facultad de Ciencias, Granada, España.

Batanero, Carmen; Godino, Juan. 2001. Análisis de datos y su didáctica. Grupo de Investigación y Educación Estadística, Departamento de Didáctica de la Matemática, Universidad de Granada. España. Servicio de Reprografía de la Facultad de Ciencias, Granada, España.

Brousseau, G. (1986). Fundamentos y Métodos de la Didáctica de las Matemáticas. Traducción al castellano del artículo "Fondements et méthodes de la didactiques des mathématiques" publicado en la revista Recherches en Didactique des Mathématiques, 7(2):33-115, y realizada por Julia Centeno, Begoña Melendo y Jesús Murillo.

Chevallard, Y. (1998). La transposición didáctica. Del Saber Sabio Al Saber Enseñado. Tercera edición, Aique editor.

Núñez, F, Sanabria, G. (2009). Didáctica del Análisis Multivariado de Datos. En Escuela de Matemática, Instituto Tecnológico de Costa Rica. *Memorias I Encuentro de Didáctica de la Estadística, la Probabilidad y el Análisis de datos (I EDEPA), 2, 3 y 4 de diciembre de 2009.* Cartago, Costa Rica.

Sanabria, G. Núñez, F. (2008). Una propuesta para la comprensión del Análisis en Componentes principales. En Instituto Tecnológico y de Estudios Superiores de Monterrey. *Memorias del I Encuentro Latinoamericano de Educación Estadística (ELEE), del 4 y 5 de Julio de 2008.* Monterrey, México.

Sanabria, G. Núñez, F. (2008). Una propuesta para la compresión de las técnicas de análisis multivariado: Análisis Factorial de Correspondencias (AFC) y Análisis Factorial de Correspondencias Múltiples (ACM). En Escuela de Matemática, Universidad Nacional de Costa Rica. *Memorias I Congreso Internacional de Computación y Matemática, del 21 al 23 de agosto del 2008.* Heredia, Costa Rica.

Sanabria, G. Núñez, F. (2009). *Informe del proyecto Estudio de Análisis Multivariado de Datos.* Costa Rica: Instituto Tecnológico de Costa Rica.

Sanabria, G. Núñez, F. (2010). Una propuesta para la comprensión de la Clasificación Jerárquica Ascendente (CJA). En Facultad de Ciencias Físico Matemáticas, Benemérita Universidad

Autónoma de Puebla. *Memorias Primer Encuentro Nacional en la Enseñanza de la Probabilidad y la Estadística (1° ENEPE), del 16 al 18 de junio del 2010*. Puebla, México.

Sanabria, G. Núñez, F. (2010). Una propuesta para la compresión del Análisis Factorial Discriminante (AFD). En Escuela de Matemática, Universidad Nacional de Costa Rica. *Memorias II Congreso Internacional de Computación y Matemática, del 26, 27 y 28 de agosto del 2010*. Heredia, Costa Rica.

Trejos, J. (2004). *Notas del curso de Análisis de Datos Multivariados, correspondiente al programa de Maestría en Matemática con énfasis en Matemática Educativa*. Costa Rica: Universidad de Costa Rica.

Trejos, J. (1998). *Introducción al análisis de datos*. Costa Rica: Programa de investigación en modelos y análisis de datos, Escuela de matemática, Universidad de Costa Rica.

Trejos, J. (2000). *Manual del Usuario. PIMAD. Versión 3.0*. Costa Rica: Universidad de Costa Rica.

Vérgnaud, Gerard. 1990. "La thèorie des champs conceptuals", Rècherches en Didactique des Mathémathiques 10 (23): 133-170.