Análisis Multivariante, Algunas Aplicaciones en Casos Clínicos

Por

Zuly Mary Briceño

Trabajo aprobado

_____ _____

Coordinador

_____

Barquisimeto, 29 de Marzo de 2010

AGRADECIMIENTO

Al Consejo de desarrollo Científico y Humanístico (CDCHT) y a la Dirección de Extensión Universitaria (DEU) de la Universidad Centroccidental “Lisandro Alvarado” por el apoyo económico para la realización de las investigaciones que sustentan este trabajo.

A todos los miembros del Laboratorio de Bioquímica Nutricional del Decanato de Ciencias Medicas de la Universidad Centroccidental “Lisandro Alvarado”, por todo el apoyo y la colaboración brindada, en especial al Licenciado Jhan Papale, a la Licenciada Yelitza Berne, al Doctor Miguel Ángel Churillo, al Doctor Rafael Bonfante, a la Licenciada Esther Giménez, al Doctor Mario Torres y a la técnica de laboratorio Luli González.

A todos los miembros del Laboratorio de Histopatología de la Sección de Anatomía Microscópica del Departamento de Ciencias Morfológicas del Decanato de Ciencias Médicas de la Universidad Centroccidental “Lisandro Alvarado”, e igualmente a la Sección de Microbiología del Departamento de Medicina Preventiva y Social del Decanato de Ciencias Médicas de la Universidad Centroccidental “Lisandro Alvarado”, por la colaboración brindada.

Resumen

El contenido de este trabajo forma parte de los resultados de las labores de investigación desarrolladas por la autora en el Laboratorio de Bioquímica Nutricional del Decanato de Medicina de la Universidad Centroccidental “Lisandro Alvarado”.

La intención fundamental es dar a conocer algunas de las aplicaciones de las técnicas estadísticas de Análisis Multivariante para resolver parte de los problemas presentados en el laboratorio.

El trabajo esta dividido en 4 capítulos. En el primer capítulo presenta un resumen general y algunas reseñas históricas acerca de las técnicas multivariantes más utilizadas. Cada uno de los restantes capítulos está dedicado a presentar con detalle casos de investigación en Ciencias Biomédicas, en los cuales la autora ha aplicado diversas técnicas del Análisis Multivariante en el análisis de los datos correspondientes a dichas investigaciones.

En el segundo capítulo se describe una aplicación en la cual se utilizaron tres técnicas multivariantes, el Análisis de Segmentación (AS), el Análisis de Cluster y el Análisis de Correspondencia Múltiple (ACM), con el objeto de estudiar la relación existente entre el diagnóstico de Zinc asociado con un grupo de variables nutricionales estudiadas en un grupo de niños menores de 15 años de edad.

En el tercer capitulo se describe una aplicación en la cual se utilizaron dos técnicas multivariantes, el Análisis de Componentes Principales (ACP) y el Análisis de Correspondencia Simple o Análisis de Correspondencia Binaria, con el objeto de estudiar la asociación entre la deficiencia de hierro y la parasitosis intestinal intestinal con un grupo de variables medidas en niños menores de 15 años de una comunidad rural del Estado Lara.

En el cuarto capitulo se describe una aplicación en la cual se utilizó la técnica multivariante denominada Análisis de Regresión Logística, para determinar la relación entre el profesional encargado de la toma de muestras citológicas cérvico vaginales y la calidad de las mismas.

A través de estas aplicaciones se observa, cómo dependiendo del objeto de la investigación y del tipo de variables involucradas en ella, se recurre al uso de distintas técnicas multivariantes que permiten lograr los objetivos planteados al inicio de la investigación.

Tabla de Contenido

Capitulo 1
Breve introducción sobre los métodos multivariantes

Hoy por hoy se reconoce y aprecia la importancia de la Estadística en todas las esferas de la ciencia, hasta el punto de que es utilizada en disciplinas tales como Historia, Literatura y Lingüística, en las cuales la idea de realizar estudios cuantitativos era inconcebibles hasta hace unos pocos años.

Kachigan (1991) define el análisis multivariante como la rama del análisis estadístico que se centra en la investigación simultanea de dos o más características (variables) medidas en un conjunto de objetos. Suele utilizarse el término "multivariante" (del inglés multivariate) para destacar el hecho de que se consideran múltiples variables, y se considera como sinónimo de multivariable y multivariado

El Análisis Multivariante está constituido por un conjunto de métodos y técnicas utilizadas en el estudio del comportamiento simultáneo de varias variables, que permiten obtener una visión de conjunto de fenómenos de la realidad cuya complejidad exige que sean estudiados con técnicas de mayor alcance que las de la estadística univariante o bivariante. Su objetivo fundamental es resumir y sintetizar la información contenida en grandes conjuntos de datos, con el fin de lograr una mejor comprensión del fenómeno en estudio. Lo que fundamentalmente caracteriza al análisis multivariante es el estudio del comportamiento conjunto de las variables y sus relaciones, y no la multiplicidad de ellas.

Hasta épocas relativamente recientes, los métodos multivariantes habían permanecido en el campo meramente teórico. Actualmente, con el uso de los potentes equipos de computación, estos métodos son utilizados en muchas investigaciones científicas, habiéndose comprobado ampliamente su eficacia en el tratamiento de grandes masas de datos. Precisamente, el término "Análisis de Datos" surge en la década de los 60 con la intención de distinguirlo del Análisis Multivariante Clásico basado en modelos y supuestos teóricos, y enfatizando la idea de la descripción de conjuntos numerosos de datos.

Reseña Histórica de la Técnicas Multivariantes

El origen del Análisis Multivariante descansa sobre los conceptos matemáticos desarrollados por matemáticos franceses e italianos del siglo pasado, quienes se dedicaron a estudiar los aspectos del Álgebra Matricial que sirvieron de base para la factorización de una matriz en sus valores y vectores singulares (DVS). Los primeros estudios multivariantes se remontan a las generaciones de los estudios de correlación y regresión realizados a principios de siglo por Francis Galton, Kart Pearson y Charles Spearman, científicos ingleses que trabajaron en psicología y biometría. En particular, a Galton se debe el término "regresión", con el cual se refería a la tendencia de las tallas de los individuos hacia la estatura promedio de la población a la cual pertenecen. Por su parte, Pearson definió al Análisis de Componentes Principales (ACP) como una técnica para hallar rectas y planos de ajuste óptimo a un conjunto de n puntos en un espacio p-dimensional. Fue Sperman quien desarrolló el primer modelo de Análisis de Factores (AF), en el cual se postula que los resultados de cualquier test psicométrico se pueden expresar como una combinación lineal de un factor común a todas las pruebas que incluye el test y de un factor específico para cada prueba. El trabajo desarrollado posteriormente por R. A. Fisher incorpora formalmente el lenguaje algebraico y el punto de vista geométrico a algunas distribuciones probabilísticas, al Análisis de Varianza, al Diseño de Experimentos y al Análisis Discriminante. En particular la ley de distribución normal, que surge con los trabajos de De Moivre, Laplace y Gauss en el siglo XVII, adquiere forma bivariante a finales del siglo XIX con los trabajos de Galton y Pearson, deviniendo en multivariante con

los trabajos de Fisher a principios de este siglo. Estos tres grandes maestros de la estadística hicieron importantes aplicaciones en las áreas de Antropometría, Genética y Biometría.
El establecimiento definitivo de la mayoría de los Métodos Multivariantes se produce alrededor de los años 30 en Estados Unidos con H. Hotelling (1931), S. Wilks (1932) y Bartlett (1939). Hotelling en 1933 plantea el problema del ACP como un procedimiento de reducción de variables, estableciendo que es posible construir un conjunto de nuevas y pocas variables incorrelacionadas, denominadas componentes, que logran resumir la información contenida en las variables originales. Simultáneamente se desarrolla en la India un movimiento que hace aportes fundamentales a los métodos multivariantes, iniciado por P. C. Mahalanobis (1936) y S. N. Roy (1939), y posteriormente profundizado por C. R. Rao (1952-1964) y P. R. Krisnaiah (1971). En 1939 Hotelling discute una interpretación geométrica del ACP en términos de elipsoides de concentración de una distribución normal multivariante. Por esa misma época, sus aportes son complementados con 4 artículos fundamentales de M. Girschilk, R. Fisher, P. Hsu y S. Roy sobre la distribución probabilística de los valores propios de la matriz de varianzas y covarianzas de una muestra procedente de una población normal multivariante. Estas ideas son ampliamente desarrolladas en los textos clásicos de T. W. Anderson (1958) y M. Kendall y A. Stuart (1969).

Thurstone, en 1930 reformula el Análisis Factorial proponiendo un modelo con varios factores comunes e imprimiéndole un sentido geométrico al mismo. Además de los desarrollos del AF y del ACP en la década de los 30, surge el Análisis Discrimininante introducido por Fisher. La función lineal discriminante de Fisher se relaciona con la T^2 de Hotelling introducida por este autor en 1931, así como con la distancia D^2 de Mahalanobis. El Análisis Canónico, que constituye una generalización de la Correlación Múltiple a dos conjuntos de variables, es propuesto por Hotelling en 1935. Más tarde P. Horst, J. Carroll y J. Kettenring extienden este enfoque a varios conjuntos de variables, surgiendo así lo que se conoce como Análisis Multicanónico.

En 1936, C. Eckart y G. Young publican un trabajo que resulta de fundamental importancia en el desarrollo de las técnicas multivariantes. En este artículo se presentan la teoría de

Aproximaciones de Matrices, basada en la descomposición de una matriz en sus valores singulares, cuya algebra y geometría constituyen el soporte matemático de la mayoría de las técnicas de Análisis de Datos.

El Análisis de Correspondencias (AC) tiene su origen en el método de Promedios Recíprocos obtenido por H. Hirchsfeld en 1935. Este método define un procedimiento de optimización para asignar puntuaciones a las modalidades de dos variables categóricas, que relaciona los vectores directores de los espacios de representación óptima de las dos variables mediante las relaciones de doble transición definidas por la DVS. De esta manera la puntuación asignada a la j-ésima modalidad de una de las variables es, salvo un coeficiente, una media ponderada de las puntuaciones de la otra variable. J.P. Benzecri y B. Escofier presentan en 1969 el Análisis de Correspondencia desde una óptica geométrica y multidimensional, cercana a la que Pearson le imprimió al ACP. Estos autores inician los fructíferos trabajos de la denomina Escuela Francesa de Análisis de Datos, que posteriormente han sido continuados entre otros, por L. Lebart, A. Morienat y J.P. Fenelon.

En 1971 K.R. Gabriel desarrolla los principios del Biplot, técnica factorial que se diferencia de las anteriores en que garantiza la representación simultanea de los objetos de estudio y de sus atributos. En esta década se inicia la escuela Sueca de Análisis de Datos, promovida fundamentalmente por K.G. Joreskog y D. Sorbom.

A partir de los años 80 surge la Escuela Holandesa con los trabajos de Van de Geer, Kroonenberg, Jan de Leeuw y el grupo GIFI de la Universidad de Leiden, cuyas investigaciones se han centrado en el estudio y desarrollo de técnicas multivariantes aplicadas a datos categóricos.

Clasificación de los Métodos Multivariantes

Los procedimientos para abordar el conocimiento de los fenómenos reales son muy similares en todas las ramas del quehacer científico. Cuando hay tres o más variables involucradas en el problema, los métodos estadísticos multivariantes permiten analizar simultáneamente las interrelaciones que se producen entre ellas, aún cuando éstas se irán

haciendo tanto más complejas cuanto mayor sea el número de variables a analizar. Si el interés del investigador consiste en estudiar la asociación entre dos conjuntos de variables, donde uno de ellos (variables independientes, explicativas o predictoras) ayuda a predecir o a explicar el comportamiento del otro (variables dependientes, explicativas o respuestas), entonces las técnicas apropiadas para el tratamiento de los datos corresponden a los métodos denominados de dependencia. En el caso que el interés se centre en el estudio de las interrelaciones entre las variables sin distinguir entre sus roles, se utilizan Métodos de Interdependencia, algunos de los cuales se conocen como Métodos de Reducción de la Dimensión y otros como Métodos de Clasificación y Escalamiento.

Para abordar el estudio del comportamiento de las variables y de sus interrelaciones, los métodos multivariantes consideran como elemento fundamental de análisis la variabilidad existente en los datos, buscando explicarla a través de las fuentes que la originan. En el caso de los métodos de dependencia la variabilidad de las variables dependientes es explicada por las independientes, que son variables observables. Usualmente esta explicación no es completa, y por ello se agrega un término de error que capta aquella parte de la variabilidad no recogida por las primeras. En el caso de los Métodos de Interdependencia conocidos como de Reducción de la Dimensión, se supone que un conjunto de variables observables pueden ser explicadas en términos de otro conjunto de variables no observables; las fuentes de variabilidad en los datos se atribuyen a éstas últimas. En relación con los Métodos de Clasificación y escalamiento se utilizan medidas de semejanza entre los objetos de estudio para detectar patrones de agrupación que conducen a la formación de clases homogéneas de objetos, lo que da lugar a particionar la variabilidad total de los datos en dos términos, uno debido a la variabilidad interna de los grupos y otro a la variabilidad entre ellos.

Panorama General de los Métodos Multivariados

Los métodos multivariados son extraordinariamente útiles para ayudar a los investigadores a hacer que tengan sentido conjuntos grandes, complicados y complejos de datos que constan de una gran cantidad de variables medidas en números grandes de unidades

experimentales. La importancia y utilidad de los métodos multivariados aumentan al incrementarse el número de variables que se están midiendo y el número de unidades experimentales que se están evaluando.

A menudo, el objetivo primario de los análisis multivariantes es resumir grandes cantidades de datos por medio de relativamente pocos parámetros. El tema subyacente de muchas técnicas multivariadas es la simplificación.

A menudo, el interés de los análisis multivariados es encontrar relaciones entre:

1) Las variables respuestas.
2) Las unidades experimentales.
3) Tanto las variables respuestas como las unidades experimentales.

Se podría decir que existen relaciones entre las variables respuesta cuando, en realidad, algunas de las variables están midiendo una unidad común. Podrían existir relaciones las unidades experimentales si algunas de ellas son semejantes entre si.

Muchas técnicas multivariadas tienden a ser de naturaleza exploratoria en lugar de confirmatoria. Es decir, muchos métodos multivariados tienden a motivar hipótesis en lugar de probarlas. Considere una situación en la cual un investigador puede tener 50 variables medidas sobre más de 2000 unidades experimentales. Los métodos estadísticos tradicionales suelen exigir que un investigador establezca algunas hipótesis, reúna algunos datos y, a continuación, use estos datos para comprobar o rechazar esas hipótesis. Una situación alternativa que se da frecuentemente es un caso en el cual un investigador dispone de una gran cantidad de datos y se pregunta si pudiera haber una información valiosa en ellos. Las técnicas multivariadas suelen ser útiles para examinar los datos en un intento por saber si hay información que valga la pena y seas valiosa en esos datos.

Técnicas Dirigidas por las Variables y Dirigidas por los Individuos.

Una distinción fundamental entre los métodos multivariados es que algunos se clasifican como “técnicas dirigidas por las variables”, en tanto que otras se clasifican como “técnicas dirigidas por los individuos”.

Las técnicas dirigidas por las variables son aquellas que se enfocan primordialmente en las relaciones que podrían existir entre las variables respuesta que se están midiendo. Algunos ejemplos de este tipo de técnica se encuentran en los análisis realizados sobre las matrices de correlación, el análisis de componentes principales, el análisis por factores, el análisis de regresión y el análisis de correlación canónica.

Las técnicas dirigidas por los individuos son las que se interesan principalmente en las relaciones que podrían existir entre las unidades experimentales o individuos que se están midiendo, o en ambos. Algunos ejemplos de este tipo de técnica se encuentran en el análisis discriminante, el análisis por agrupación y el análisis multivariado de la varianza (MANOVA: multivariate analysis of variance)

Técnicas Dirigidas a Crear Nuevas Variables

Con bastante frecuencia es de utilidad crear nuevas variables para cada unidad experimental, de modo que se puedan comparar entre sí con más facilidad.

Muchos métodos multivariados ayudan a los investigadores a crear nuevas variables que tengan propiedades deseables. Algunas de las técnicas multivariadas que crean nuevas variables son el análisis de componentes principales, el análisis por factores, el análisis de correlación canónica, el análisis discriminante canónico y el análisis de variables canónicas.

Análisis de Componentes Principales (PAC: principal components análiyis)

Esta técnica tiene por objeto transformar un conjunto de variables, denominadas variables originales, en un nuevo conjunto de variables denominadas *componentes principales*. Estas últimas se caracterizan por estar incorrelacionadas.

El análisis de componentes principales permite pasar a un nuevo conjunto de variables, las componentes principales, que gozan de la ventaja de estar incorrelacionadas entre sí y que, además, pueden ordenarse de acuerdo con la información que llevan incorporada.
Como medida de la cantidad de información incorporada en una componente se utiliza su varianza. Es decir, cuanto mayor sea su varianza mayor es la información que lleva incorporada dicha componente. Por esta razón se selecciona como primera componente aquella que tenga mayor varianza mientras que, por el contrario, la última es la de menor varianza.

En general, la extracción de componentes principales se efectúa sobre variables tipificadas para evitar problemas derivados de escala, aunque también se puede aplica sobre variables expresadas en desviaciones respecto a la media. El nuevo conjunto de variables que se obtiene por el método de componentes principales es igual en número al de variables originales. Es importante destacar que la suma de sus varianzas es igual a la suma de las varianzas de las variables originales. Las diferencias entre ambos conjuntos de variables estriba en que las componentes están incorrelacionadas entre sí. Cuando las variables originales están muy correlacionadas entre sí, la mayor parte de su variabilidad se puede explicar con muy pocas componentes.

Si las variables originales estuvieran completamente incorrelacionadas entre sí desde el inicio, entonces el análisis de componentes principales carecería por completo de interés, ya que en ese caso las componentes principales coincidirán con las variables originales.

El ACP se puede hacer sobre una matriz de varianza covarianza de las muestras o una matriz de correlación.

Análisis por Factores (FA: factor analysis)

El análisis por factores es una técnica que se emplea frecuentemente para crear nuevas variables que resuman toda la información de la que podría disponerse de las variables

originales. El análisis por factores también se usa para estudiar las relaciones que podrían existir entre las variables medidas en un conjunto de datos. Semejante al PCA, el FA es una técnica dirigida por las variables.

Un objetivo básico del FA es determinar si las variables respuesta exhiben patrones de relaciones entre sí, tales que esas variables se puedan dividir en subconjuntos de modo que las variables en un subconjunto estén fuertemente correlacionadas con cada una de las otras y que las variables en subconjuntos diferentes tengan bajas correlaciones entre sí. Por tanto el FA se usa con frecuencia para estudiar la estructura de correlación de las variables en un conjunto de datos. Una semejanza entre FA y PCA es que aquel también se puede usar para crear nuevas variables que no estén correlacionadas entre sí. Esas variables se llaman *clasificación de factores.*

Una de las ventajas que tiene el FA sobre el PCA, es que las variables creadas por el FA son mucho más fáciles de interpretar que las creadas por el PCA.

Análisis Discriminante (DA: discrinant analysis)

Es una técnica multivariante que se usa principalmente para clasificar individuos o unidades experimentales en dos o más poblaciones definidas de manera única. El objeto de la técnica es desarrollar una regla discriminante que clasifique las unidades experimentales en una de varias categorías posibles. El investigador debe tener una muestra aleatoria de unidades experimentales de cada grupo posible de clasificación. Entonces, el DA proporciona los métodos que permitirán a los investigadores establecer reglas que se puedan emplear para clasificar otras unidades experimentales en uno de los grupos de clasificación.

Análisis Discriminante Canónico (CDA: canonical discriminant analysis)

Es una técnica multivariante con la que se crean nuevas variables que contienen toda la información útil para la discriminación de la que se dispone en las variables originales. A

menudo, estas nuevas variables conducen a reglas más sencillas para clasificar las unidades experimentales en los diferentes grupos.

Análisis de Regresión Logística

Esta técnica multivariante se usa para modelar la probabilidad de que una unidad experimental caiga en un grupo particular, con base en la información medida en la propia unidad. Estos modelos se pueden usar con fines se discriminación.

Este método se considera en situaciones en las que las variables predictoras no estén distribuidas normalmente y en las que algunas o todas esas variables sean discretas o categóricas. La regresión logística es semejante a la regresión múltiple. la diferencia principal es que, en la logística la variable dependiente suele ser binaria, en tanto que en la múltiple, esa variable dependiente es continua.

Análisis de Correspondencia Binaria

Es una técnica multivariante cuyo objetivo consiste en explicar la asociación existente entre dos variables cualitativas, utilizando dispositivos gráficos construidos a manera de diagramas de dispersión los cuales se denominan planos factoriales. Sobre los planos se representan simultáneamente los perfiles de las categorías de las variables incluidas en el estudio, que definen respectivamente las filas y las columnas de una tabla de contingencia.

Las representaciones obtenidas aproximan con la mayor fidelidad posible los aspectos más importantes de la información contenida en las tablas de perfiles.

Análisis de Cluster (CA: cluster analysis)

Esta técnica es semejante al análisis discriminante en el sentido de que se usa para clasificar individuos o unidades experimentales en subgrupos definidos de manera única. Este análisis se puede emplear cuando el investigador cuenta con muestras aleatorias previamente obtenidas de cada uno de los subgrupos definidos de manera única. El análisis por agrupación trata de los problemas de clasificación cuando no se sabe de antemano de cuales subgrupos se originan las observaciones.

Análisis de Segmentación

Es una técnica multivariante cuya finalidad es formar grupos de individuos, definidos por valores de variables independientes (predictores) que particularmente se consideran categóricas, que sean bien diferenciados entre si con respecto al perfil de una variable dependiente (respuesta). Es por tanto una técnica de agrupación.

El Análisis de Segmentación se diferencia del Análisis de Cluster en quer esta últimas tecnica no distingue entre variables dependientes y variables independientes sino que todas cumplen el mismo papel.

Análisis de Correspondencia Múltiple

Es una técnica multivariante de interdependencia recientemente desarrollada que faciliota tanto la reducción dimensional de una clasificación de objetos (por ejemplo, productos, personas, etc.) sobre un conjunto de atributos y el mapa perceptual de objetos relativos a estos atributos. Los investigadores se enfrentan constantemente za la necesidad de cuantificar datos cualitativos que encuentran en variables nominales. El análisis de correspondencia difiere de otras técnicas de interdependencia discutidas antes en su capacidad para acomodar tanto datos no métricos como relaciones no lineales.

Análisis Multivariado de la Varianza (MANOVA: multivariate analysis of variance)

Es una técnica multivariante que generaliza el análisis univariado de la varianza (ANOVA). Es usada para comparar las medias de varias poblaciones en una sola variable medida.

Cuando se miden varias variables en cada unidad experimental, podría producirse un ANOVA sobre cada variable medida, usando una variable a la vez, el MANOVA puede ayudarnos a comparar varias poblaciones al considerar, simultáneamente todas las variables medidas y no una a la vez.

Debe realizarse un MANOVA siempre que se están comparando entre sí dos o más poblaciones diferentes sobre un número grande de variables respuesta. Si un MANOVA muestra diferencias significativas entre las medias de las poblaciones, entonces el investigador puede confiar en que verdaderamente existen diferencias reales. En este caso, resulta razonable considerar el análisis de una variable a la vez para detectar dónde ocurren en realidad las diferencias. Si el MANOVA no revela diferencias significativas entre las medias de las poblaciones, entonces el investigador debe tener precaución extrema al interpretarse como “positivos falsos”

Análisis de Variables Canónicas (CVA: canonical variates analysis)

El análisis de variables canónicas es un método en el que se crean nuevas variables en conjunción con los análisis multivariados de la varianza. Estas nuevas variables son útiles porque ayudan a los investigadores a determinar en dónde ocurren las diferencias importantes entre las medias de las poblaciones, cuando se están comparando poblaciones sobre muchas variables diferentes mediante el uso simultáneo de todas las variables medidas. En ocasiones, las variables canónicas pueden sugerir diferencias importantes que, de lo contrario, podrían pasarse por alto.

Análisis de Correlación Canónica.

Es una técnica multivariante la cual es generalización de la correlación múltiple en los problemas de regresión. Para aplicarla se requiere que las variables respuesta se dividan en dos grupos. La asignación de las variables en estos dos grupos siempre debe motivarse por la naturaleza de las variables respuesta y nunca por una inspección de los datos. Por ejemplo, una asignación legítima sería aquella en la que las variables en uno de los grupos sean fáciles de obtener y no caras para medirse, mientras que las que se encuentren en el otro grupo sean difíciles de obtener y no caras para medirse.

Una cuestión básica que se espera responder con el análisis de correlación canónica es si se puede usar las variables que se encuentran en uno de los grupos para predecir las variables en el otro. Cuando se puede, entonces este análisis intenta resumir las relaciones entre los dos conjuntos de variables, mediante la creación de nuevas variables a partir de cada uno de los dos grupos de variables originales.

Capitulo 2

Bases matemáticas de los métodos multivariantes

Cuando solamente se considera una variable, las medidas resumen más comúnmente utilizadas para describir el comportamiento de los datos son la media, la varianza y las medidas de asimetría y curtosis, cuyo cálculo no es afectado por el volumen de los datos. Cuando se trata de múltiples variables las medidas estadísticas utilizadas, muchas de ellas definidas como generalizaciones de sus contrapartes univariantes se organizan sobre arreglos matriciales cuya obtención requiere cálculos laboriosos que se hacen más complejos cuanto mayor es el número de variables involucradas. Los datos multidimensionales tienen características muy específicas cuyo tratamiento requiere procedimientos especiales que pueden entender más fácilmente a partir de operaciones matriciales.

Existen procedimientos algebraicos que permiten abordar de manera simplificada el cálculo de medidas resumen comúnmente utilizadas para caracterizar muestras de poblaciones multivariantes, así como para producir representaciones gráficas aproximadas que ilustren visualmente los principales aspectos de la información contenida en los arreglos de datos asociados.

La información sobre la que se explican los métodos multivariantes se organiza sobre una matriz de datos X con n filas y p columnas, las filas describen a los individuos (unidades

estadística) y las columnas quedan definidas por un conjunto de p variables que caracterizan a los primeros. Por consiguiente, el elemento genérico de la matriz que denotaremos mediante x_{ij} representa el valor de la variable j observado sobre el i-ésimo individuo. Una estructura de datos como la descrita es generada por un estudio transversal.

La forma expandida de la matriz X es la siguiente:

Variable j

$$X_{np} = \begin{pmatrix} x_{11} & x_{12\Lambda\Lambda} & x_{1p} \\ M & M & M \\ x_{i1} & x_{i2} & x_{ip} \\ M & M & M \\ x_{n1} & x_{n2} & x_{np} \end{pmatrix} \longrightarrow \text{Individuo i}$$

El i-ésimo vector fila de este arreglo contiene las observaciones correspondientes al individuo i en cada una de las p variables, y se denotara mediante:

$$X_i^t = (x_{i1} \quad x_{i2} \quad x_{i3} \quad \ldots. \quad x_{ip})$$

El j-ésimo vector columna describe la información de la variable j medida sobre los n individuos, y se denotará mediante:

$$X^j = \begin{pmatrix} x \\ x_{2j} \\ M \\ x_{ij} \\ M \\ x_{nj} \end{pmatrix}$$

Matrices Grammian

Las matrices de la forma X^tX y de la forma XX^t. denominadas matrices Grammian, son de particular interés en estadística, debido a que sobre ellas se organiza información fundamental para el análisis de relaciones entre variables y del parecido entre individuos respectivamente. En particular, la matriz de varianzas y covarianzas y la matriz de correlaciones puede considerase como ejemplos por excelencia de este tipo de matrices.

Análisis General

El análisis general comprende una serie de herramientas del álgebra matricial que constituyen el núcleo matemático común de las principales técnicas del análisis factorial de datos multivariantes.

Como ya se ha dicho, la información básica de referencia para el análisis está constituida por una matriz de datos X_{nxp}, cuyas filas quedan descritas por individuos caracterizados de acuerdo con un conjunto de p variables. El análisis tiene como propósito central extraer las aspectos más importantes de la información contenida en la matriz de datos, relacionados específicamente con semejanzas entre individuos y patrones en la estructura de varianzas e íntercorrelaciones entre las variables. Esta información sirve en primero lugar como referencia inicial para sugerir agrupamientos de individuos y en segundo lugar, para determinar las direcciones principales de la estructura de relaciones entre variables a partir de las cuales es posible explicar las diferencias o el parecido entre los individuos.

Ajuste en R^p

La nube de puntos en R^p (nube de individuos) que denotaremos por N(I), queda definida por el conjunto de las n filas X_1, $X_{2,}$ .....X_n de la matriz de datos, dotadas de una distancia (la euclidea) que se utiliza para evaluar semejanzas entre ellas. El ajuste, consiste en encontrar

el subespacio de R^p de dimensión q (q<p), con direcciones ortonormalizadas, que produzcan una óptima a la nibe N(I) en el sentido de los mínimos cuadrados. En otras palabras, el subespacio debe garantizar globalmente el mayor parecido entre los datos originales y sus representaciones sobre ese subespacio. El procedimiento a seguir para la construcción del espacio de representación, que de ahora en adelante denominaremos "subespacio de mejor ajuste", se desarrolla en etapas sucesivas. En cada etapa se obtiene un vector normalizado que define una nueva dirección del subespacio, que debe ser ortogonal respecto de los previamente hallados.

En la primera etapa hallaremos el subespacio unidimensional (la recta) de mejor ajuste. El vector director de esa recta será denotado por v y para efectos prácticos, exigiremos que sea normalizado. En términos estadísticos se trata de hallar un espacio unidimensional de representación de datos con una dirección normalizada, que minimice la suma de cuadrados de errores que se cometen al aproximar los X_i mediante sus estimaciones mínimos cuadráticas sobre el subespacio.

La estimación mínimo cuadrática de X_i queda determinada por su proyección ortogonal, ya que $\overline{X_i} = pxi$ es el vector sobre la recta que minimiza la distancia del punto al subespacio:

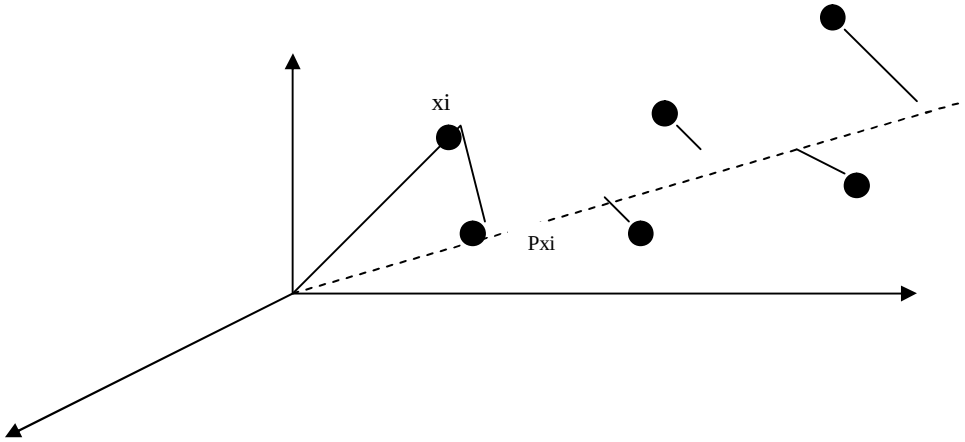
$$\min_{xi} \| X_i - \overline{X_i} \|^2 = d^2(X_i, Pxi)$$

La traducción matemática de este planteamiento remite entonces a resolver un problema de mínimos condicionados:

$$\min_{v^t v=1} \sum d^2(xi, Pxi)$$

Siendo Pxi la proyección ortogonal de xi sobre la recta.

Gráficamente, en el caso de R^3:



Minimizar globalmente la suma de cuadrados de errores cometidos al aproximar los xi mediante su proyección ortogonal sobre la recta de mejor ajuste, es equivalente a maximizar la suma:

$$\sum d^2(\theta, Pxi) = \sum v^t xixi^t v = v^t \sum xixi^t v = v^t X^t Xv$$

Para maximizar esta forma cuadrática, bajo la condición de que v sea un vector normalizado, construimos el Lagrangiano:

$$L = v^t X^t Xv - \lambda(v^t v - 1)$$

Obteniéndose la ecuación:

$$X^t Xv = \lambda v$$

Lo que determina que el vector buscado v es un autovector de la matriz X^tX asociado con el autovalor λ. Esto quiere decir que una condición necesaria para que v defina un máximo en la forma cuadrática bajo consideración, es que sea autovector de X^tX.

El vector director de la recta que mejor se ajusta a la nube de puntos es el autovector de X^tX, que se denota comúnmente v^1, asociado con su mayor autovalor, que se denota comúnmente λ_1.

Realizando un análisis similar al anterior, ahora en un subespacio de dimensión 2, se observa que las direcciones del plano de mejor ajuste a la nube de puntos en R^p están definidas por los vectores v^1 y v^2, autovectores ortonormalizados de la matriz X^tX asociados con sus dos mayores autovalores λ_1 y λ_1.

Al iterar q veces el procedimiento se obtiene que, una base ortonormalizada del subespacio de dimensión q que mejor se ajusta a la nube de puntos fila en R^p, esta constituida por los q autovectores $v^1, v^2, \ldots v^q$ correspondientes a los q mayores autovalores $\lambda_1, \lambda_2, \ldots \lambda_q$ de la matriz X^tX.

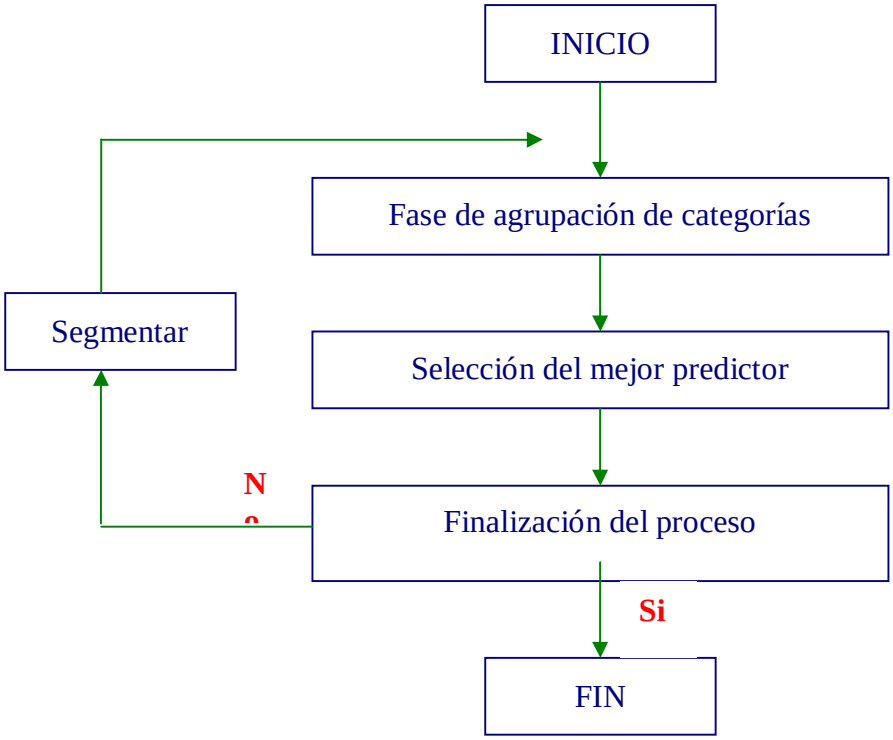
Un análisis equivalente se puede hacer con las columnas, obteniéndose un subespacio de dimensión R^n. El análisis se hace sobre la matriz XX^t .

A continuación se dará un breve resumen sobre consideraciones matemáticas que se deben tomar en cuenta al aplicar algunas de las técnicas multivariantes utilizadas en nuestras aplicaciones.

Recordemos que el Análisis de Segmentación (AS) es una técnica de análisis estadístico multivariante cuya finalidad es formar grupos de individuos, definidos por valores de variables independientes (predictores), que particularmente consideraremos categóricas, que sean bien diferenciados entre sí con respecto al perfil de una variable dependiente (respuesta). Es por tanto una técnica de agrupación.

El AS pertenece a una familia de métodos originalmente denominados **AID** (Automatic Interaction Detection) cuyo objetivo básico era detectar la existencia de interacción en un modelo de predicción. Actualmente sin embargo, se utilizan con fines exploratorios y descriptivos.

Algoritmo general de segmentación



Existen algunos algoritmos **AID** (Automatic Interaction Detection) comúnmente utilizados entre los que se tienen los siguientes:

XAID (eXtended AID)

Es una extensión del algoritmo de Morgan y Sonquist que utiliza el estadístico F del Análisis de la Varianza.

THAID (Theta AID)

Produce segmentaciones binarias utilizando como criterio la maximización del número de observaciones en cada categoría modal.

CHAID (Chi-square AID)

Utiliza el contraste Chi-cuadrado de independencia en las diferentes fases del proceso.

Tomando en cuenta que el modelo esta formado por un grupo de predoctores para efecto de establecer las agrupaciones permitidas, se pueden considerar varios **tipos de predictores**, según la naturaleza de sus categorías:

Predictores Monótonos

Se dice que un predictor es monótono si sus categorías pertenecen a una escala ordinal. Esto implica que solamente categorías contiguas en la escala pueden ser agrupadas para formar una sola.

Si consideramos un predictor monótono con c categorías iniciales, el total de agrupaciones posibles en d categorías es:

c – 1;

d - 1

Predictores Libres

Se dice que un predictor es libre si sus categorías pertenecen a una escala nominal. Esto implica que se permite la agrupación de cualesquiera categorías.

Si consideramos un predictor libre con c categorías iniciales, el total de agrupaciones posibles en d categorías es:

$$\sum_{i=0}^{d-1}(-1)^i \frac{(d-1)^c}{i!(d-1)!}$$

Predictores Flotantes

En este caso todas las categorías del predictor pertenecen a una escala ordinal menos una de ellas (que denominaremos flotante) que no concuerda con el resto, o cuya posición en la escala ordinal es desconocida. Con la excepción de la categoría flotante, se permite la agrupación solamente para categorías contiguas. La categoría flotante puede quedar sola o combinada con cualquier otra categoría o grupo de categorías.

Reglas de finalización

Si no se pusiesen otras limitaciones al proceso, éste terminaría solamente cuando no hubiese predictores significativos en ninguno de los grupos. En ese caso, probablemente el

estadístico chi-cuadrado se obtendría a partir de tablas poco ocupadas, con la problemática que esto conlleva.

El proceso de segmentación se limitar mediante la introducción de ciertos controles, a los cuales denominaremos filtros. Los posibles filtros a utilizar son los siguientes:

Significación de Categoría (SC)

Es el nivel de significación utilizado en la fase de agrupación de categorías. Para verificar si dos categorías tienen un perfil similar, es decir, no son significativamente diferentes, se compara su significación con SC.

En CHAID esto se lleva a cabo cruzando la variable dependiente con las dos categorías del predictor bajo consideración, se calcula el estadístico chi-cuadrado y se compara su valor p correspondiente con SC.

Significación de Predictor (SP)

Es el nivel de significación utilizado en la fase de selección del mejor predictor. Para verificar si un predictor es significativo, se compara su significación con SP.

En CHAID esto se lleva a cabo cruzando la variable dependiente con el predictor ya agrupado, se calcula el estadístico chi-cuadrado y se compara su valor p correspondiente con SP.

Filtros de Asociación (FA)

Se establece una asociación mínima entre la variable dependiente y el predictor para considerarlo como potencial candidato para realizar la segmentación.

Si un determinado coeficiente de asociación entre la variable dependiente y el predictor es menor que FA, éste es descartado.

Una posibilidad sería utilizar el coeficiente de Pawlik, que expresa el valor del coeficiente de contingencia (CC) como porcentaje del valor máximo:

$$CP = \left(\frac{CC}{\sqrt{\frac{r-1}{r}}} * 100 \right) \%$$

siendo r el mínimo entre el número de filas y el de columnas.

Tamaño Antes (TA)

Se establece un tamaño mínimo para que un grupo pueda segmentarse. Esto quiere decir que si un grupo determinado Gh tiene menos de TA individuos, no se segmenta y se declara como terminal.

Tamaño Después (TD)

Se establece una tamaño mínimo para que un subgrupo sea formado. Por lo tanto, si alguno de los grupos formados en la segmentación de Gh, digamos Ghj, tiene menos de TD individuos, la segmentación es descartada.

Filtro de Nivel (FN)

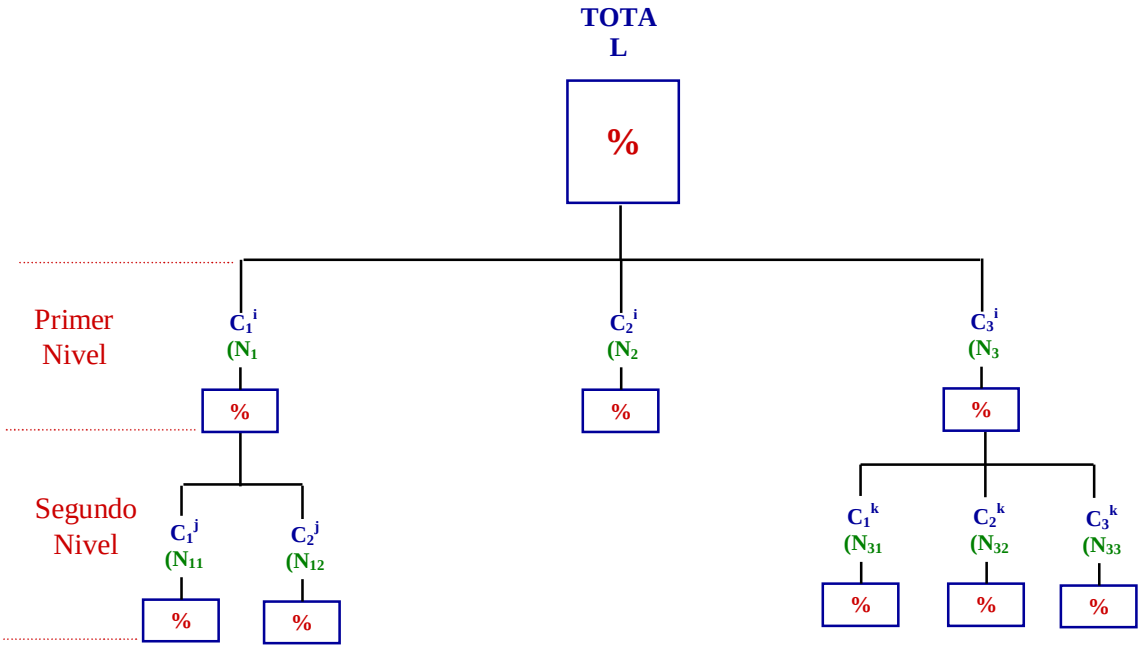
Se establece un máximo número de niveles de segmentación. Una segmentación con un solo nivel resulta útil pero demasiado simple.

Por otro lado, una segmentación con muchos niveles puede resultar compleja y difícil de interpretar.

El árbol de segmentación

El resultado del proceso de segmentación suele representarse en un diagrama de árbol en el cual se muestran en forma resumida los diferentes grupos formados en cada nivel de segmentación, el predictor responsable de la partición, el tamaño del grupo y una descriptiva de la variable dependiente.

Cada "nodo" del árbol representa una segmentación, y en él se indica el predictor Xi que la produce. Cada "rama" del árbol representa a uno de los subgrupos y en ella se indica la categoría Chi que define al grupo, y entre paréntesis el tamaño del grupo. Dentro del rectángulo se indica la descriptiva correspondiente. Los niveles quedan reflejados en cada "franja" horizontal del árbol.



Análisis Factorial de Correspondencias Simples y Multiple

El análisis factorial de correspondencias simples contempla el posible agrupamiento de sujetos (grupos) y de modalidades según el comportamiento que los I sujetos (grupos)

manifiesten en las J características o Modalidades normalmente de una sola variable discreta.

El análisis factorial de correspondencias múltiples se interesa por las interrelaciones entre varias modalidades de distintas variables de forma que pueda conocerse si algunas actúan de manera semejante y a la vez distinta de otro conjunto de ellas. Algo así como lo que sucede con las variables cuantitativas mediante el análisis factorial de componentes principales buscando los ejes o factores latentes.

El punto de partida básico del análisis de correspondencias múltiples puede ser una tabla de datos binarios (tabla de unos y ceros) como la que sigue:

VARIABLES							
V1			V2			V3	
1	0	0	1	0	0	1	0
0	1	0	0	1	0	1	0
0	0	1	1	0	0	0	1
0	1	0	0	0	1	1	0
1	0	0	0	1	0	0	1
0	0	1	0	1	0	0	1

Esta tabla consta de 6 filas (sujetos) y 3 variables. La variable 1 (V1) tiene 3 modalidades de respuesta, la variable 2 (V2) lo mismo y la variable 3 (V3) tiene dos modalidades. El sujeto 4 ha respondido o elegido la modalidad dos de la variable V1, la modalidad tres de la variable V2 y la modalidad uno de la variable V3.

A partir de la matriz Z, correspondiente a la de la tabla anterior, puede calcularse la matriz de frecuencias relativas P de la forma:

$$f_{ij} = \frac{Z_{ij}}{I * V} \qquad f_{i+} = \frac{V}{I * V} = \frac{1}{I} \qquad f_{+j} = \frac{n_j}{I * V}$$

Siendo I el número de sujetos, V el numero de variables, y nj la frecuencia absoluta de la modalidad j

Si se realiza el cálculo de $Z^t * Z$ resultará una matriz B simétrica llamada TABLA DE BURT.

$$B = \sum_{i=1}^{I} Z_{ij} Z_{ik} = n_{jk}$$

siendo j, k cualquiera de las J modalidades.

En esta matriz B los efectivos de la diagonal son las frecuencias absolutas de las 8 modalidades y el resto son las frecuencias absolutas resultantes del cruce entre si de las 8 modalidades.

Tabla de Burt (B)

		VARIABLES							
		V1			V2			V3	
Modalidades		1	2	3	1	2	3	1	2
V1	1	2	0	0	1	1	0	1	1
	2	0	2	0	0	1	1	2	0
	3	0	0	2	1	1	0	0	2
V2	1	1	0	1	2	0	0	1	1
	2	1	1	1	0	3	0	1	2
	3	0	1	0	0	0	1	1	0
V3	1	1	2	0	1	1	1	3	0
	2	1	0	2	1	2	0	0	3

La tabla de Burt puede ser definida como un cruce de J modalidades entre sí o lo que es lo mismo, como un cruce de V variables, cada una con su número particular de modalidades.

A partir de este cruce de modalidades precisamente, es decir, a partir de sus similitudes o diferencias conjuntas podrán lograrse es objetivo del análisis de correspondencias múltiples que no es otro que encontrar ejes o factores alrededor de los cuales se aglutinen algunas de tales modalidades.

Por supuesto que la tabla de Burt, como ocurre en el caso del Análisis de Correspondencias Simples, admite la existencia de modalidades activas y de modalidades suplementarias. Las primeras son las que entran a formar parte del análisis mientras que las segundas pueden conocerse su pertenencia a uno u otro factor una vez realizado el análisis.

En el análisis factorial de correspondencias múltiples se siguen los pasos y el desarrollo matemático expuestos para analizar el comportamiento de las J modalidades en la matriz Z de I filas, según el Análisis de Correspondencia simples lo que en definitiva se acaba consiguiendo es un análisis de las J modalidades entre sí.

Consideremos la matriz U a diagonalizar como:

$$U_{jk} = \sum_{i=1}^{I} \frac{\frac{Z_{ij}}{I*V}\frac{Z_{ik}}{I*V}}{\frac{V}{I*V}\sqrt{\frac{n_j}{I*V}}\sqrt{\frac{n_k}{I*V}}} = \frac{\sum_{i=1}^{I} Z_{ij}Z_{ik}}{V*\sqrt{n_j n_k}} = \frac{B}{V*\sqrt{n_j n_k}} = \frac{n_{jk}}{V*\sqrt{n_j n_k}}$$

Por consiguiente con la diagonalización de esta matriz U calculada a partir de la matriz de Burt puede lograrse analizar las interrelaciones entre las J modalidades.

La inercia total que debe explicarse será:

$$\text{Inercia Total} = \sum_{j=1}^{J}\left(d_{ji}^2 \frac{n_j}{I*V}\right) = \frac{J}{V} - 1$$

Que coincide con la traza de la matriz U menos 1, al despreciarse el primer valor propio igual a la unidad y que se deja de tener en cuenta lo mismo que sus autovectores correspondientes.

Al ser B y U matrices simétricas de dimensión J x J, los vectores propios, las coordenadas, las contribuciones absolutas y relativas son iguales para filas y columnas.

Consideremos que u_{jf} es el vector propio en cada factor y λ_f es el valor propio asociado al anterior en cada factor.

Coordenadas de las modalidades en los distintos F factores.

$$coord_{jf} = \frac{\sqrt{\lambda_f}}{\sqrt{\frac{n_j}{I*V}}} u_{jf}$$

Contribuciones absolutas de las modalidades de los F factores

$$C.Ab_{jf} = u_{jf}^2$$

Contribuciones relativas de las modalidades de los F factores

$$C.Re_{jf} = \frac{(coord_{jf})^2}{(dist_j)^2} = \frac{(coord_{jf})^2}{\frac{I}{n_j} - 1}$$

Distancia al centro de gravedad

$$(dist_j)^2 = \frac{I}{n_j} - 1$$

Ésta se deduce teniendo en cuenta que para cada modalidad el centro de gravedad es la frecuencia marginal relativa a cada sujeto.

Cuando interese conocer la posición o coordenadas de alguna modalidad suplementaria en cada factor la forma de calcularla es similar, como todo lo anteriormente expuesto, a como se realiza en el Análisis de Correspondencias Simples.

En la interpretación de los resultados del Análisis de Correspondientes múltiples al igual que ocurre en el Análisis de Correspondencias Simples y el análisis factorial con variables cuantitativas, la identificación, contenido, sentido,… que se le atribuye a cada factor depende de la subjetividad del investigador.

Lo que estos análisis ofrecen es tan sólo la ubicación de una serie de modalidades que, en el caso de estar más o menos agrupados pueden indicar un comportamiento similar y distinto al de otro conjunto de variables o modalidades que, también agrupadas, estén sin embargo, lejos del subgrupo anterior.

Regresión Logistica

La regresión logística es una de las herramientas estadísticas con mejor capacidad para el análisis de datos en investigación clínica y epidemiología, de ahí su amplia utilización. El objetivo primordial que resuelve esta técnica es el de modelar cómo influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores y el valor o nivel de los mismos.
También puede ser usada para estimar la probabilidad de aparición de cada una de las posibilidades de un suceso con más de dos categorías (politómico).

De todos es sabido que este tipo de situaciones se aborda mediante técnicas de regresión. Sin embargo, la metodología de la regresión lineal no es aplicable ya que ahora la variable respuesta sólo presenta dos valores (nos centraremos en el caso dicotómico), como puede ser presencia/ausencia de hipertensión.

Si clasificamos el valor de la variable respuesta como 0 cuando no se presenta el suceso (ausencia de hipertensión por ejemplo) y con el valor 1 cuando sí está presente (paciente hipertenso por ejemplo), y buscamos cuantificar la posible relación entre la presencia de

hipertensión y la cantidad media de sal consumida al día como posible factor de riesgo, podríamos caer en la tentación de utilizar una regresión lineal:

$$Hipertensión = a + b \cdot [Consumo\_sal]$$

y estimar, a partir de nuestros datos, por el procedimiento habitual de mínimos cuadrados, los coeficientes *a* y *b* de la ecuación. Sin embargo, y aunque esto es posible matemáticamente, nos conduce a la obtención de resultados absurdos, ya que cuando se calcule la función obtenida para diferentes valores de consumo de sal se obtendrá resultados que, en general, serán diferentes de 0 y 1, los únicos realmente posibles en este caso, ya que esa restricción no se impone en la regresión lineal, en la que la respuesta puede en principio tomar cualquier valor.

Si utilizamos cómo variable dependiente la probabilidad ***p*** de que un paciente padezca hipertensión y construimos la siguiente función:

$$\ln \frac{p}{1-p}$$

ahora sí tenemos una variable que puede tomar cualquier valor, por lo que podemos plantearnos el buscar para ella una ecuación de regresión tradicional:

$$\ln \frac{p}{1-p} = a + b \cdot [consumo\_sal]$$

que se puede convertir con una pequeña manipulación algebraica en

$$\Pr.HTA = \frac{1}{1 + e^{\left(-a - b \cdot [consumo\_sal]\right)}}$$

Y este es precisamente el tipo de ecuación que se conoce como modelo logístico, donde el número de factores puede ser más de uno, así en el exponente que figura en el denominador de la ecuación podríamos tener:

b1.consumo_sal + *b2*.edad + *b3*.sexo + *b4*.fumador

Los coeficientes del modelo logístico como cuantificadores de riesgo

Una de las características que hacen tan interesante la regresión logística es la relación que éstos guardan con un parámetro de cuantificación de riesgo conocido en la literatura como "**odds ratio**" (aunque puede tener traducción al castellano, renunciamos a ello para evitar confusión ya que siempre se utiliza la terminología inglesa).

El odds asociado a un suceso es el cociente entre la probabilidad de que ocurra frente a la probabilidad de que no ocurra:

$$odds = \frac{p}{1-p}$$

Siendo ***p*** la probabilidad del suceso. Así, por ejemplo, podemos calcular el odds de presencia de hipertensión cuando el consumo diario de sal es igual o superior a una cierta cantidad, que en realidad determina cuántas veces es más probable que haya hipertensión a que no la haya en esa situación. Igualmente podríamos calcular el odds de presencia de hipertensión cuando el consumo de sal es inferior a esa cantidad. Si dividimos el primer odds entre el segundo, hemos calculado un cociente de odds, esto es un odds ratio, que de alguna manera cuantifica cuánto más probable es la aparición de hipertensión cuando se consume mucha sal (primer odds) respecto a cuando se consume poca. La noción que se está midiendo es parecida a la que encontramos en lo que se denomina **riesgo relativo** que corresponde al cociente de la probabilidad de que aparezca un suceso cuando está presente el factor respecto a cuando no lo está. De hecho cuando la prevalencia del suceso es baja (< 20 %) el valor del odds ratio y el riesgo relativo es muy parecido, pero no es así cuando el suceso es bastante común, hecho que a menudo se ignora y será objeto de un comentario más extenso en un nuevo artículo.

Si en la ecuación de regresión tenemos un factor dicotómico, como puede ser por ejemplo si el sujeto es no fumador, el coeficiente ***b*** de la ecuación para ese factor está directamente relacionado con el odds ratio ***OR*** de ser fumador respecto a no serlo

$$OR = \exp(b)$$

es decir que *exp(b)* es una medida que cuantifica el riesgo que representa poseer el factor correspondiente respecto a no poseerlo, suponiendo que el resto de variables del modelo permanecen constantes.

Cuando la variable es numérica, como puede ser por ejemplo la edad, o el índice de masa corporal, es una medida que cuantifica el cambio en el riesgo cuando se pasa de un valor del factor a otro, permaneciendo constantes el resto de variables. Así el odds ratio que supone pasar de la edad *X1* a la edad *X2*, siendo *b* el coeficiente correspondiente a la edad en el modelo logístico es:

$$OR = \exp[b \cdot (X2 - X1)]$$

Nótese que se trata de un modelo en el que el aumento o disminución del riesgo al pasar de un valor a otro del factor es proporcional al cambio, es decir a la diferencia entre los dos valores, pero no al punto de partida, quiere esto decir que el cambio en el riesgo, con el modelo logístico, es el mismo cuando pasamos de 40 a 50 años que cuando pasamos de 80 a 90.

Cuando el coeficiente *b* de la variable es positivo obtendremos un odds ratio mayor que 1 y corresponde por tanto a un factor de riesgo. Por el contrario, si *b* es negativo el odds ratio será menor que 1 y se trata de un factor de protección.

Las variables cualitativas en el modelo logístico

Puesto que la metodología empleada para la estimación del modelo logístico se basa en la utilización de variables cuantitativas, al igual que en cualquier otro procedimiento de regresión, es incorrecto que en él intervengan variables cualitativas, ya sean nominales u ordinales.

La solución a este problema es crear tantas variables dicotómicas como número de respuestas - 1. Estas nuevas variables, artificialmente creadas, reciben en la literatura

anglosajona el nombre de "*dummy*", traduciéndose en español con diferentes denominaciones como pueden ser variables internas, indicadoras, o variables diseño.

CAPITULO 3

Aplicación de técnicas multivariantes para estudiar los niveles de zinc y un grupo de variables nutricionales

Introducción

En este capitulo se describe la aplicación de tres técnicas multivariantes. Estas técnicas se aplicaron con el propósito de estudiar la relación existente entre el diagnóstico de Zinc y un grupo de variables nutricionales, estudiadas en un grupo de niños menores de 15 años de edad, pertenecientes a una comunidad rural del Estado Lara.

Las técnicas multivariantes utilizadas fueron el Análisis de Segmentación (AS), el Análisis de Cluster y el Análisis de Correspondencia Múltiple (ACM). La muestra estuvo conformada por un total de 342 niños. La recolección de los datos y el análisis de los mismos se realizaron desde enero de 2005 hasta abril de 2007. A través del Análisis de Segmentación se observó que la variable que mejor explica el diagnóstico de zinc es el grupo etario, los grupos formados por los valores de dicha variable son bien diferenciados en cuanto a las categorías normal y descompensado. A través del Análisis de Cluster se agruparon los niños en seis grupos. A través del ACM se observaron algunas agrupaciones interesantes de modalidades para el conjunto de datos estudiado. Se observó que los niños con una edad comprendida entre dos y seis años de edad, cuya madre tiene un nivel de educación secundaria, presentan valores de cobre normal y de zinc normal. Por otro lado se observa que los niños que tienen una madre que sabe leer presentan una modalidad

“circunferencia de brazo normal’; se observó además que niños que tienen una madre analfabeta presentan una modalidad “zinc” deficiente.

Se obtuvo previamente el consentimiento informado por escrito de todos los representantes de los niños que participaron en el estudio.

Análisis Estadísticos Aplicados a los Datos

Fueron analizados en los niños un total de tres variables antropométricas: Peso, Talla, y Circunferencia de Brazo; un grupo de indicadores relacionados con el estado nutricional: Peso para la Talla, Talla para la Edad y Peso para la Edad. Se utilizaron variables para registrar el estado nutricional del niño en cinco categorías, el estado nutricional del niño en tres categorías, los valores de zinc, los valores de cobre, el diagnóstico de zinc en dos categorías, el diagnóstico de cobre en dos categorías, la edad de los niños en años, la edad de los niños por grupos etarios y el nivel de instrucción de la madre. En la Tabla 1 se resumen las frecuencias de cada una de las modalidades de las variables categóricas incluidas en el análisis y en la tabla 2 se describe el valor promedio, desviación típica de las variables continuas.

Se realizó un Análisis de Segmentación con el objeto de investigar las variables que mejor explicaban el diagnóstico de zinc presentado por los niños. El análisis generó el árbol de segmentación presentado en la Figura 1.

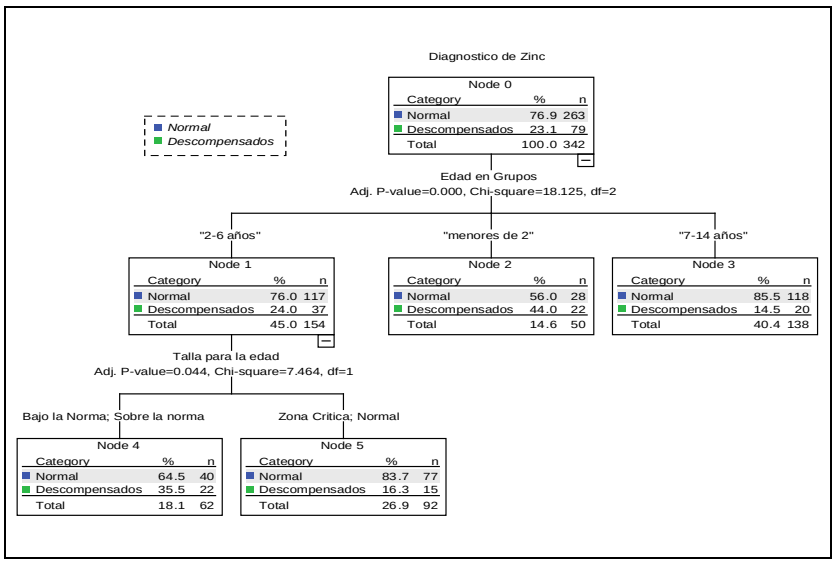
Para la realización del análisis se tomó como variable dependiente la variable diagnóstico de zinc y como variables independientes las variables diagnóstico de cobre, edad en grupos, nivel de instrucción de la madre, diagnóstico nutricional en tres categorías, peso para la edad, talla para la edad y peso para la talla. El análisis excluyó varias variables dejando únicamente como variables dependientes las variables edad en grupos y talla para la edad.

En el árbol se observa que el perfil del diagnóstico de zinc en la población estudiada es de 76,9 % de niños normales y 23,1 de niños descompensados. La variable que mejor explica el diagnóstico de zinc es el grupo etario, los grupos formados por los valores de dicha variable son bien diferenciados en cuanto a las categorías normal y descompensado.

Mayoritariamente los niños entre dos y seis años presentan un diagnóstico de zinc normal, así como también los niños menores de dos años y los niños entre 7 y 14 años de edad.

En el árbol se observa una agrupación de categorías. Se unieron las categorías bajo la norma y sobre la norma de la variable talla para la edad, indicando que para los niños entre dos y seis años de edad, ambas categorías no hacen diferencia en cuanto al diagnóstico de zinc, bajo el mismo criterio se unieron las categorías zona crítica y normal de la variable talla para la edad.

Figura 1. Árbol de Segmentación. Salida generada por el programa SPSS

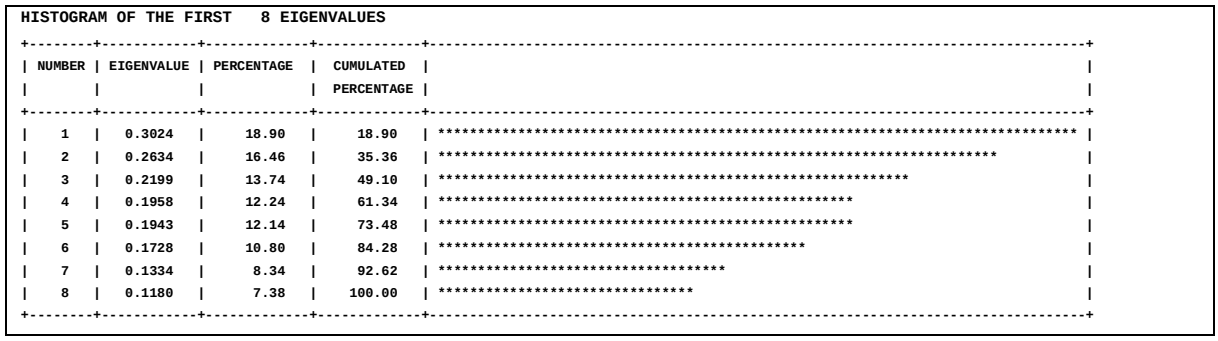


Se realizó un análisis simultáneo de las categorías de algunas de las variables estudiadas a través de un Análisis de Correspondencia Múltiple. Para la realización del análisis se tomó como variable ilustrativa la variable diagnóstico de zinc y como variables nominales activas las variables diagnóstico nutricional en cinco categorías, circunferencia de brazo, diagnóstico de cobre, grupo etario del niño y grado de instrucción de la madre. Los resultados más importantes del análisis son mostrados en las Figuras 2 , 3 y 4.

En la figura 2 se observa el histograma de los ocho primeros valores propios vinculados al análisis factorial. Se observa que el primer autovalor explica 18.9 % del total de la variabilidad presente en los datos, el segundo autovalor explica el 16.46 % del total de la

variabilidad presente en los datos, el tercer autovalor explica el 13.74 % y el cuarto autovalor explica el 12.24 %. Los cuatro juntos explican el 61.34 % de la variabilidad total presente en los datos. Para explicar más de un cincuenta por ciento de la variabilidad total se necesita tomar cuatro autovalores.

Figura 2. Histograma de los Autovalores. Salida generada por el programa SPAD

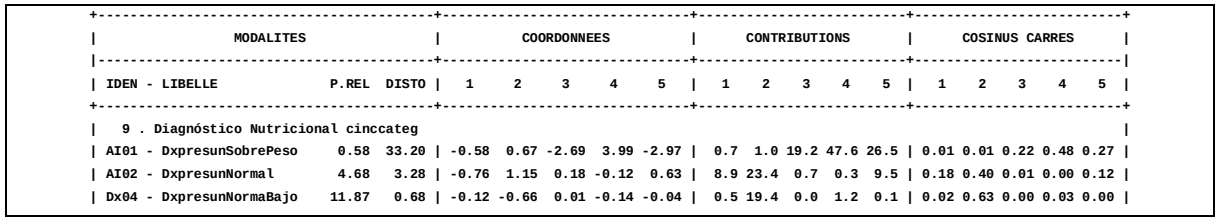


HISTOGRAM OF THE FIRST 8 EIGENVALUES

NUMBER	EIGENVALUE	PERCENTAGE	CUMULATED PERCENTAGE	
1	0.3024	18.90	18.90	**********
2	0.2634	16.46	35.36	*********
3	0.2199	13.74	49.10	*******
4	0.1958	12.24	61.34	*******
5	0.1943	12.14	73.48	*******
6	0.1728	10.80	84.28	******
7	0.1334	8.34	92.62	*****
8	0.1180	7.38	100.00	*****

En la figura 3 se observa la tabla de coordenadas, contribuciones y cosenos cuadrados de las modalidades activas consideradas. Los valores bajo la columna que lleva por título “Cosinus Carres” nos indican que la mayoría de las categorías no están muy bien representadas en el eje factorial, la gran mayoría esta por debajo del 20 % a excepción de las modalidad desnutrido de la variable diagnóstico nutricional en cinco categorías y las modalidades bajo la norma y normal de la variable circunferencia del brazo. Los valores bajo la columna que lleva por título ”Contribution” nos indican que las modalidades que más contribuyen a la variabilidad del primer eje es la modalidad desnutrido de la variable diagnóstico nutricional y la modalidad bajo la norma de la variable circunferencia de brazo. El resto de de las contribuciones de las modalidades es bastante bajo.

Figura 3. Coordenadas, Contribuciones y Cosenos Cuadrados de las Modalidades Activas.
Salida generada por el programa SPAD



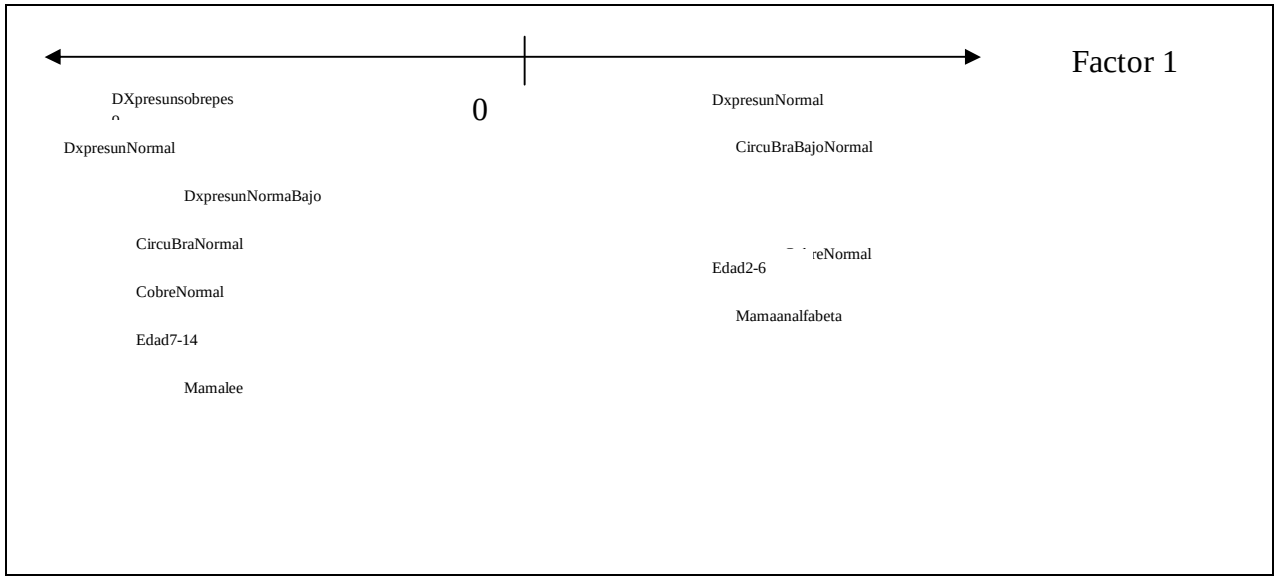
MODALITES			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDEN - LIBELLE	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
9 . Diagnóstico Nutricional cinccateg																	
AI01 - DxpresunSobrePeso	0.58	33.20	-0.58	0.67	-2.69	3.99	-2.97	0.7	1.0	19.2	47.6	26.5	0.01	0.01	0.22	0.48	0.27
AI02 - DxpresunNormal	4.68	3.28	-0.76	1.15	0.18	-0.12	0.63	8.9	23.4	0.7	0.3	9.5	0.18	0.40	0.01	0.00	0.12
Dx04 - DxpresunNormaBajo	11.87	0.68	-0.12	-0.66	0.01	-0.14	-0.04	0.5	19.4	0.0	1.2	0.1	0.02	0.63	0.00	0.03	0.00

| Dx05 - DxpresunDesnutrido	2.87	5.98 |	1.85	0.70	0.23	-0.03	-0.26 |	32.3	5.4	0.7	0.0	1.0 |	0.57	0.08	0.01	0.00	0.01 |
+----------CONTRIBUTION CUMULEE =								42.4	49.2	20.6	49.1	37.2 +----------+					
| 11 . Circunferencuia brazo																	
| AK01 - CircufBraBajoNorma	5.38	2.72 |	1.22	-0.39	-0.15	-0.32	-0.23 |	26.6	3.1	0.5	2.8	1.5 |	0.55	0.06	0.01	0.04	0.02 |
| AK02 - CircuBraNormal	14.62	0.37 |	-0.45	0.14	0.05	0.12	0.08 |	9.8	1.1	0.2	1.0	0.5 |	0.55	0.06	0.01	0.04	0.02 |
+----------CONTRIBUTION CUMULEE =								36.4	4.2	0.7	3.9	2.0 +----------+					
| 15 . Diagnóstico de cobre																	
| AO01 - CobreNormal	19.24	0.04 |	-0.02	0.06	-0.08	-0.13	-0.09 |	0.0	0.2	0.5	1.5	0.9 |	0.01	0.08	0.16	0.40	0.23 |
| AO02 - CobreDeficiente	0.76	25.31 |	0.54	-1.43	2.00	3.17	2.40 |	0.7	5.9	13.9	39.1	22.5 |	0.01	0.08	0.16	0.40	0.23 |
+----------CONTRIBUTION CUMULEE =								0.8	6.2	14.4	40.6	23.4 +----------+					
| 16 . Grupo etario del niño																	
| AP01 - Menore2años	2.92	5.84 |	0.76	1.69	-0.30	0.13	0.42 |	5.5	31.8	1.2	0.2	2.6 |	0.10	0.49	0.02	0.00	0.03 |
| AP02 - Edad2-6	9.01	1.22 |	0.07	-0.14	0.81	0.15	-0.43 |	0.1	0.7	26.9	1.0	8.5 |	0.00	0.02	0.54	0.02	0.15 |
| in03 - Edad7-14	8.07	1.48 |	-0.35	-0.46	-0.80	-0.21	0.33 |	3.3	6.4	23.2	1.9	4.4 |	0.08	0.14	0.43	0.03	0.07 |
+----------CONTRIBUTION CUMULEE =								9.0	38.9	51.3	3.2	15.5 +----------+					
| 17 . Grado de instruccion de la madre																	
| AR01 - MamaAnlafabeta	11.81	0.69 |	0.35	-0.12	-0.31	0.15	0.38 |	4.7	0.6	5.3	1.3	9.0 |	0.17	0.02	0.14	0.03	0.21 |
| in04 - Mamalee	8.19	1.44 |	-0.50	0.17	0.45	-0.21	-0.55 |	6.8	0.9	7.7	1.9	13.0 |	0.17	0.02	0.14	0.03	0.21 |
+----------CONTRIBUTION CUMULEE =								11.5	1.6	13.0	3.2	21.9 +----------+					

Continuación de la tabla de Coordenadas, Contribuciones y Cosenos Cuadrados de las Modalidades Activas.

Salida generada por el programa SPAD

Los valores bajo la columna que lleva por título “Cordonnees” nos dan una ubicación de las modalidades en cada uno de los factores. Por ejemplo para el primer factor se puede observar la siguiente ubicación de las modalidades.



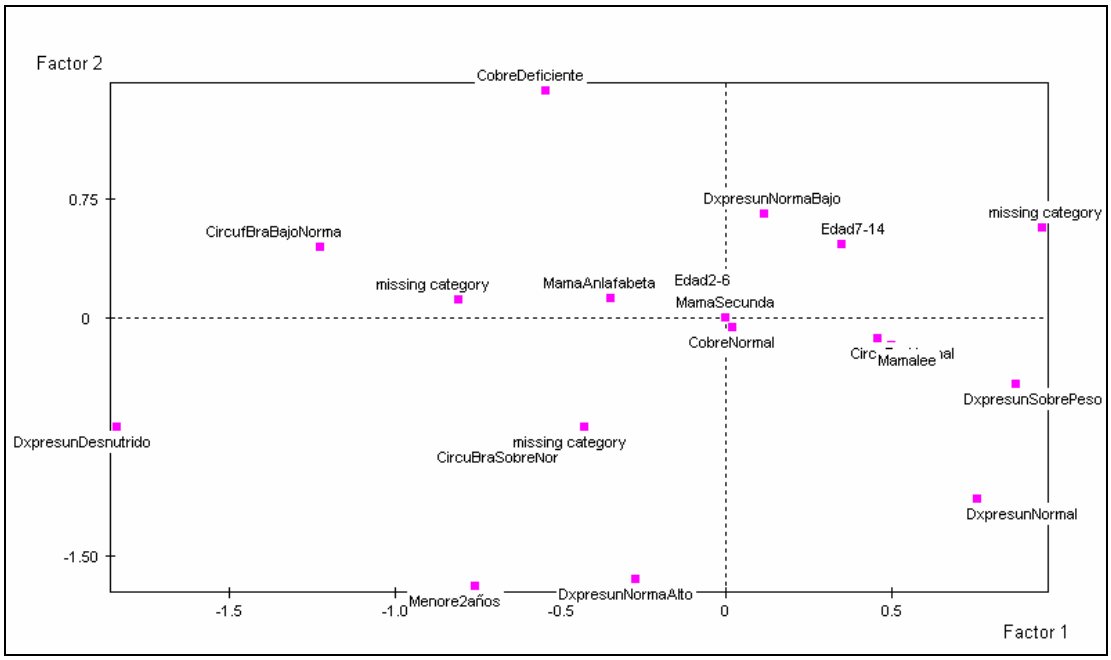
Fuente: Investigación primaria

El primer factor distingue modalidades asociadas con la nutrición del niño. Es un factor del estado de nutrición. De un extremo se ubican modalidades que describen condiciones

anormales en la nutrición de los niños como serían las modalidades sobrepeso, normal y normal bajo de la variables diagnóstico nutricional en cinco categorías, la modalidad normal de la variable circunferencia de brazo, la modalidad normal de la variable diagnóstico de cobre, la modalidad 7-14 de la variable edad y la modalidad “mama lee”. Del otro extremo del factor se ubican las modalidades desnutrido de la variable de la variable diagnóstico nutricional en cinco categorías, la modalidad bajo la norma de la variable circunferencia de brazo, las modalidades menores de dos años y edad entre dos y seis de la variable grupo etario y la modalidad “mamá analfabeta”. Es importante destacar que en el análisis está incluida una variable del entorno social del niño, la modalidad madre que lee y madre analfabeta.

A pesar que las modalidades en general, no están bien representadas en los factores, analizaremos el gráfico perceptual generado por el análisis. En el gráfico se observa algunas agrupaciones interesantes de modalidades. Se puede deducir del grafico que los niños con una edad comprendida entre dos y seis años de edad cuya madre tiene un nivel de educación secundaria, presentan valores de cobre normal y de zinc normal. Por otro lado se observa que los niños que tienen una madre que sabe leer presentan una modalidad circunferencia de brazo normal; se observó además que niños que tienen una madre analfabeta presentan una modalidad zinc deficiente. Al parecer hay una variable socio cultural den entorno del niño que esta influenciando el resto de las variables consideradas en el estudio.

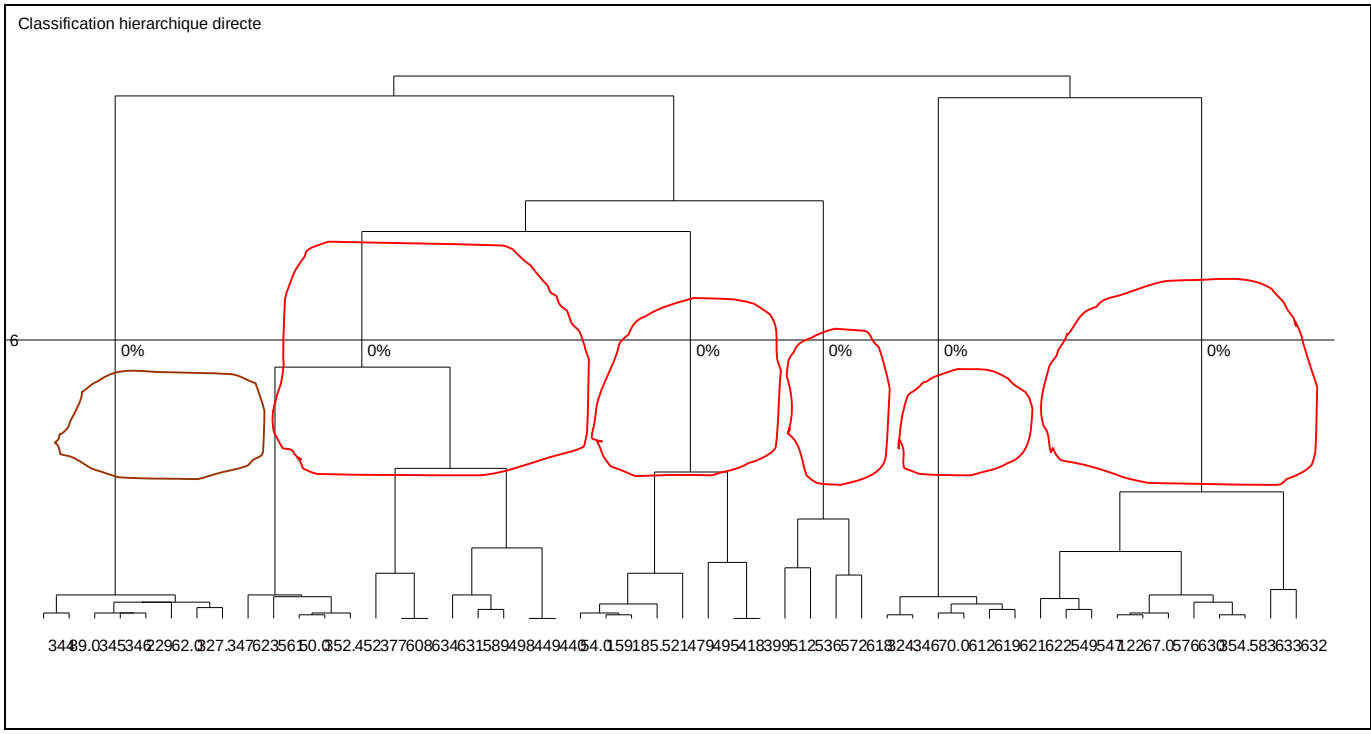
Figura 4. Gráfico Perceptual generado por el Análisis de Correspondencia Múltiple. Salida generada por el programa SPAD



Con el propósito de agrupar individuos similares entre sí, se aplicó un análisis de Cluster. Para esto se tomaron en cuenta las variables diagnóstico de zinc, diagnóstico nutricional en cinco categorías, circunferencia de brazo, diagnóstico de cobre, grupo etario del niño y grado de instrucción de la madre. Los resultados más importantes del análisis son mostrados en las Figuras 5 y 6.

En la Figura 5 se observa el Histograma de Índices de Nivel obtenido en el análisis. Dicho histograma nos da información sobre el incremento en la variabilidad dentro de la partición que se produce al fusionar los grupos. A partir de la partición 677 se observa un incremento importante en la variabilidad dentro de las clases formadas hasta el momento. Basado en la información obtenida a través del histograma se toma la decisión de agrupar los elementos en 6 clases.

Figura 5. Dendrograma. Salida generada por el programa SPAD



En la figura 6 se observa el Dendograma. Utilizando como criterio considerar los grupos tales que, la fusión siguiente va a unir individuos muy distintos, se deberían considerar seis grupos. El primer grupo contienen el 4 % de los objetos incluidos en el análisis, el segundo grupo contiene el 34 % de los objetos, el tercer grupo contiene el 28 % de los objetos, el cuarto grupo contiene el 17 % de los objetos, el quinto grupo contiene el 3 % de los objetos y el sexto grupo contiene el 14 % de los objetos.

Se ejecutó el método PARTI-DECLA del programa estadístico SPADcon el fin de estudiar de manera más detalla la estructura interna de cada clase, el cual generó las salidas mostradas en las Figuras 7 y 8.

Figura 6. Histograma de Índices de Nivel. Salida generada por el programa SPAD

```
HIERARCHICAL CLUSTER ANALYSIS (NEAREST NEIGHBORS)
ON THE FIRST    4 FACTORIAL AXES
DESCRIPTION OF THE   50 NODES WITH HIGHEST INDEX
NUM. FIRST  LAST  COUNT   WEIGHT     INDEX   HISTOGRAM OF LEVEL INDEXES
 634     3   590     3      3.00   0.00395   *
 635   555   535    10     10.00   0.00480   *
 636   582   568     7      7.00   0.00495   *
 637   604   583     9      9.00   0.00550   *
 638   632   614    17     17.00   0.00583   *
 639   618   581    15     15.00   0.00598   *
 640   608   610    15     15.00   0.00624   *
 641   606   607    27     27.00   0.00661   *
 642   596   589     8      8.00   0.00664   *
 643   635   613    24     24.00   0.00684   *
 644   636   622    11     11.00   0.00696   *
 645   616   609    17     17.00   0.00750   *
 646   637   588    13     13.00   0.00805   *
 647   612   641    39     39.00   0.00817   *
 648   640   633    24     24.00   0.00823   *
 649   619   625    18     18.00   0.00836   *
 650   621   594    15     15.00   0.00888   *
 651   628   631    30     30.00   0.00909   *
 652   623   601    22     22.00   0.00920   *
 653   629   624    10     10.00   0.00975   **
 654   603   626    13     13.00   0.01028   **
 655   571   645    22     22.00   0.01130   **
 656   586   620    13     13.00   0.01181   **
 657   642   574    10     10.00   0.01193   **
 658    68    11     2      2.00   0.01228   **
 659   634   605     5      5.00   0.01372   **
 660   654   627    27     27.00   0.01373   **
 661   597   650    29     29.00   0.01640   **
 662   558    87     3      3.00   0.01775   **
 663   652   639    37     37.00   0.01863   **
 664   655   638    39     39.00   0.01980   ***
 665   648   615    26     26.00   0.02171   ***
 666   651   647    69     69.00   0.02294   ***
 667   665   656    39     39.00   0.02496   ***
 668   664   643    63     63.00   0.03282   ****
 669   646   659    18     18.00   0.03545   ****
 670   630   649    23     23.00   0.03702   ****
 671   644   657    21     21.00   0.04287   *****
 672   663   661    66     66.00   0.04498   *****
 673   653   670    33     33.00   0.05203   ******
 674   672   669    84     84.00   0.07773   *********
 675   666   673   102    102.00   0.08237   *********
 676   675   658   104    104.00   0.10654   ************
 677   667   660    66     66.00   0.12783   **************
 678   676   674   188    188.00   0.25234   ***************************
 679   668   678   251    251.00   0.31093   *********************************
 680   671   662    24     24.00   0.39723   *****************************************
 681   680    33    25     25.00   0.45899   ************************************************
 682   677   679   317    317.00   0.62675   ********************************************************************
 683   681   682   342    342.00   0.77026
*******************************************************************************
SUM OF LEVEL INDEXES =    4.00000
```

En la Figura 7 se observa información sobre la descomposición de la inercia en cada uno de los seis grupos formados, la primera clase contiene 3 % de la inercia total, la segunda clase

contiene 22 % de la inercia total, la tercera clase contiene 10 % de la inercia total, la cuarta clase contiene 7 % de la inercia total, la quinta clase contiene 1 % de la inercia total y la sexta clase contiene 11 % de la inercia total. Se observa que el porcentaje de inercia de cada grupo es directamente proporcional al tamaño del mismo. En la figura también se observa el número de individuos clasificados en cada grupo antes y después de aplicar el algoritmo de centros móviles, el cual es el mismo.

Figura 7. Descomposición de la inercia en los grupos. Salida generada por el programa SPAD

DECOMPOSITION DE L'INERTIE
CALCULEE SUR 7 AXES.

INERTIES	INERTIES AVANT	INERTIES APRES	EFFECTIFS AVANT	EFFECTIFS APRES	POIDS AVANT	POIDS APRES	DISTANCES AVANT	DISTANCES APRES
INTER-CLASSES	0.9149	0.9193						
INTRA-CLASSE								
CLASSE 1 / 6	0.0315	0.0315	13	13	13.00	13.00	5.1311	5.1311
CLASSE 2 / 6	0.2292	0.1956	117	111	117.00	111.00	0.2965	0.3297
CLASSE 3 / 6	0.1012	0.1245	97	101	97.00	101.00	0.4125	0.3908
CLASSE 4 / 6	0.0730	0.0790	58	60	58.00	60.00	0.7602	0.7374
CLASSE 5 / 6	0.0189	0.0189	10	10	10.00	10.00	6.6635	6.6635
CLASSE 6 / 6	0.1132	0.1132	47	47	47.00	47.00	1.2928	1.2928
TOTALE	1.4820	1.4819						

En la figura 8 se observa información sobre la caracterización de cada una de las seis clases formadas. La primera clase se caracteriza por incluir el 3.8 % del total de los individuos, el 3.8 % de los niños en el conjunto de datos tienen la modalidad cobre Deficiente de la variable diagnóstico de cobre, el 96, 2 % de los niños incluidos en el estudio presentan la modalidad cobre normal de la variable diagnóstico de cobre, el 100 % de los niños en esta clase presentan la modalidad cobre deficiente de la variable diagnóstico de cobre y el 100 % de los niños con la modalidad cobre deficiente están en esta clase. La segunda clase incluye el 32.46 % de todos los niños incluidos en el estudio. El 59.06 % de los niños presentes en el estudio, tienen la modalidad normal bajo de la variable diagnóstico nutricional cinco categorías. El 77.48 % de los niños en la clase presentan la modalidad normal bajo de la variable diagnóstico nutricional cinco categorías. El 96.20 % de los niños presentes en el estudio, tienen la modalidad normal de la variable diagnóstico de Cobre.

Figura 8. Caracterización de las Clases generada por el programa SPAD

```
CLASSE  1 /  6
-------------------------------------------------------------------------------------------------------------------------------
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                                                  IDEN  POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
                                3.80  CLASSE  1 /  6                                                           aa1a     13
  9.99  0.000  100.00  100.00   3.80  CobreDeficiente      Diagnóstico de cobre                                AO02     13
 -9.99  0.000    0.00    0.00  96.20  CobreNormal          Diagnóstico de cobre                                AO01    329
CLASSE  2 /  6
-------------------------------------------------------------------------------------------------------------------------------
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                                                  IDEN  POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------------------------------------------
                               32.46  CLASSE  2 /  6                                                           aa2a    111
 15.51  0.000   72.08  100.00  45.03  Edad2-6              Grupo etario del niño                               AP02    154
  4.79  0.000   42.57   77.48  59.06  DxpresunNormaBajo    Diagnóstico Nutricional cinccateg                   Dx04    202
  2.55  0.005   33.74  100.00  96.20  CobreNormal          Diagnóstico de cobre                                AO01    329
  2.38  0.009   48.94   20.72  13.74  DxpresunDesnutrido   Diagnóstico Nutricional cinccateg                   Dx05     47
 -2.55  0.005    0.00    0.00   3.80  CobreDeficiente      Diagnóstico de cobre                                AO02     13
 -6.15  0.000    0.00    0.00  14.62  Menore2años          Grupo etario del niño                               AP01     50
 -8.11  0.000    0.00    0.00  22.81  DxpresunNormal       Diagnóstico Nutricional cinccateg                   AI02     78
-11.96  0.000    0.00    0.00  40.35  Edad7-14             Grupo etario del niño                               in03    138
-------------------------------------------------------------------------------------------------------------------------------
CLASSE  3 /  6
-------------------------------------------------------------------------------------------------------------------------------
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                                                  IDEN  POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
                               29.53  CLASSE  3 /  6                                                           aa3a    101
 15.70  0.000   73.19  100.00  40.35  Edad7-14             Grupo etario del niño                               in03    138
  8.81  0.000   46.53   93.07  59.06  DxpresunNormaBajo    Diagnóstico Nutricional cinccateg                   Dx04    202
  2.34  0.010   30.70  100.00  96.20  CobreNormal          Diagnóstico de cobre                                AO01    329
 -2.34  0.010    0.00    0.00   3.80  CobreDeficiente      Diagnóstico de cobre                                AO02     13
 -5.76  0.000    0.00    0.00  14.62  Menore2años          Grupo etario del niño                               AP01     50
 -7.61  0.000    0.00    0.00  22.81  DxpresunNormal       Diagnóstico Nutricional cinccateg                   AI02     78
-12.18  0.000    0.00    0.00  45.03  Edad2-6              Grupo etario del niño                               AP02    154
-------------------------------------------------------------------------------------------------------------------------------
CLASSE  4 /  6
-------------------------------------------------------------------------------------------------------------------------------
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                                                  IDEN  POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------------------------------------------
                               17.54  CLASSE  4 /  6                                                           aa4a     60
 14.53  0.000   75.64   98.33  22.81  DxpresunNormal       Diagnóstico Nutricional cinccateg                   AI02     78
  5.22  0.000   23.67   96.67  71.64  CircuBraNormal       Circunferencuia brazo                               AK02    245
  3.03  0.001   20.91   91.67  76.90  ZinNormal            Diagnóstico de Zinc                                 AN01    263
 -3.03  0.001    6.33    8.33  23.10  ZinDeficient         Diagnóstico de Zinc                                 AN02     79
 -3.87  0.000    0.00    0.00  13.74  DxpresunDesnutrido   Diagnóstico Nutricional cinccateg                   Dx05     47
 -4.05  0.000    0.00    0.00  14.62  Menore2años          Grupo etario del niño                               AP01     50
 -4.75  0.000    2.30    3.33  25.44  CircufBraBajoNorma   Circunferencuia brazo                               AK01     87
-10.93  0.000    0.00    0.00  59.06  DxpresunNormaBajo    Diagnóstico Nutricional cinccateg                   Dx04    202
-------------------------------------------------------------------------------------------------------------------------------
CLASSE  5 /  6
-------------------------------------------------------------------------------------------------------------------------------
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                                                  IDEN  POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------------------------------------------
                                2.92  CLASSE  5 /  6                                                           aa5a     10
  7.64  0.000  100.00   80.00   2.34  DxpresunSobrePeso    Diagnóstico Nutricional cinccateg                   AI01      8
 -3.70  0.000    0.00    0.00  59.06  DxpresunNormaBajo    Diagnóstico Nutricional cinccateg                   Dx04    202
-------------------------------------------------------------------------------------------------------------------------------
CLASSE  6 /  6
-------------------------------------------------------------------------------------------------------------------------------
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                                                  IDEN  POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------------------------------------------
                               13.74  CLASSE  6 /  6                                                           aa6a     47
 99.99  0.000   94.00  100.00  14.62  Menore2años          Grupo etario del niño                               AP01     50
  3.36  0.000   31.91   31.91  13.74  DxpresunDesnutrido   Diagnóstico Nutricional cinccateg                   Dx05     47
  3.07  0.001   25.32   42.55  23.10  ZinDeficient         Diagnóstico de Zinc                                 AN02     79
  2.44  0.007   23.08   38.30  22.81  DxpresunNormal       Diagnóstico Nutricional cinccateg                   AI02     78
 -3.07  0.001   10.27   57.45  76.90  ZinNormal            Diagnóstico de Zinc                                 AN01    263
 -4.86  0.000    5.94   25.53  59.06  DxpresunNormaBajo    Diagnóstico Nutricional cinccateg                   Dx04    202
 -6.91  0.000    0.00    0.00  40.35  Edad7-14             Grupo etario del niño                               in03    138
 -7.50  0.000    0.00    0.00  45.03  Edad2-6              Grupo etario del niño                               AP02    154
```

Tabla 1. Características de las variables categóricas

Variable	Categorías	Frecuencia de la categoría
Sexo	masculino	53.2
	femenino	46.8
Peso para la talla	Normal	83.6
	Bajo la Norma	14.0
	Sobre la norma	2.3
Talla para la edad	Bajo la Norma	39.2
	Normal	34.2
	Sobre la norma	1.5
	Zona Critica	25.1
Peso para la edad	Bajo la norma	53.8
	Normal	45.3
	Sobre la norma	.6
Diagnóstico nutricional presuntivo en cinco categorías	Sobrepeso	2.3
	Normal	22.8
	Normal Alto	1.5
	Normal Bajo	59.1
	Desnutrido	13.7
Diag nut presunt tres categorías	Sobrepeso	2.3
	Normal	83.3
	desnutrición	13.7
Circunferencia del brazo para la edad	Bajo la norma	25.4
	Normal	71.6
	Sobre la Norma	1.5
Diagnóstico de Zinc	Normal	76.9
	Descompensados	23.1
Diagnóstico de Cobre	Normal	96.2
	Descompensado	3.8
Edad en Grupos	"menores de 2"	14.6
	"2-6 años"	45.0
	"7-14 años"	40.4
Nivel de instrucción de la madre	Analfabeta	201
	No es analfabeta	139

Tabla 1. Características de las variables continuas

Variable	Mínimo valor	Máximo Valor	Valor promedio	Desviación Estándar
valor de cobre	.27	2.48	1.3046	.28031
Edad en años	.25	14.92	6.1875	3.78934
Peso en Kilos	3.100	117.000	18.65123	10.742005
Talla en Centímetros	1.14	144.50	96.4153	32.40757
Valor de zinc	.29	4.20	.8453	.26221

Conclusiones

A través del Análisis de Segmentación se observó que la variable que mejor explica el diagnóstico de zinc es el grupo etario, los grupos formados por los valores de dicha variable son bien diferenciados en cuanto a las categorías normal y descompensado. A través del Análisis de Cluster se agruparon los niños en seis grupos. A través del ACM se observaron algunas agrupaciones interesantes de modalidades para el conjunto de datos estudiado. Se observó que los niños con una edad comprendida entre dos y seis años de edad, cuya madre tiene un nivel de educación secundaria, presentan valores de cobre normal y de zinc normal. Por otro lado se observa que los niños que tienen una madre que sabe leer presentan una modalidad “circunferencia de brazo normal’; se observó además que niños que tienen una madre analfabeta presentan una modalidad “zinc” deficiente.

CAPITULO 4

Aplicación de técnicas multivariantes para estudiar la deficiencia de hierro y la parasitosis intestinal, en un grupo de niños

Introducción

En este capitulo se estudia la aplicación del Análisis de Componentes Principales y Análisis de Correspondencia Binaria para estudiar la deficiencia de hierro y la parasitosis intestinal en niños menores de 15 años de la comunidad rural La Bucarita, ubicada en el estado Lara-Venezuela.

El tipo de investigación fue descriptivo transversal. La muestra fue aleatoria estratificada de acuerdo a grupos etarios. Se recolectaron muestras sanguineas en los niños, a través de las cuales se midieron un grupo de variables químicas que determinaron los niños anémicos y con deficiencias de hierro. Por otro lado, se recolectaron muestras de heces para estudiar la parasitosis intestinal presente en los niños. Según miembros del Laboratorio de Bioquímica Nutricional de la UCLA, para la determinación de hemoglobina se usó un Coulter-ACT-8; mientras que para la determinación de ferritina sérica, el método de ELISA y para la determinación de hierro sérico y capacidad de fijación de hierro total se utilizó un espectrofotómetro de absorción atómica con horno de grafito. Los resultados muestran que el 17,2% de los individuos estudiados presentaron anemia y el 31,9 % deficiencia de hierro, de los cuales el 33,59% era anémico y el 66,41 % no anémico. Un 79,01 % de los niños y niñas estudiados presentaron parasistosis intestinal, siendo los parásitos mas frecuentes el

Áscaris Lumbricoides y el Trichuris Trichiura, con un 51,66% y 42,82% respectivamente. El grupo etario menor a dos años fue el más afectado en todos los parámetros, excepto en la parasitosis intestinal donde el grupo 7-14 fue el más afectado. A través de los análisis estadísticos aplicados al conjunto de datos se observó que los niños que presentan deficiencias en los niveles de hierro no presentan parasitosis intestinales.

La comunidad de La Bucarita es una comunidad rural ubicada en el Municipio Andrés Eloy Blanco del Estado Lara la cual carece de los servicios de suministro de agua potable, eliminación de basura y deposición de excretas, por lo que sus habitantes consumen el agua proveniente de las quebradas y ríos y defecan en el suelo; factores que facilitan, conjuntamente con las condiciones climáticas del municipio, la infestación parasitaria de sus habitantes. De aquí la importancia de determinar, en la población de La Bucarita, la prevalecía de anemia, deficiencia de hierro y parasitosis intestinal en los niños menores de 15 años.

Muestra poblacional

El censo de la población estudiada se obtuvo a través del Comité de Salud de esta comunidad. La población objeto estuvo formada por 1200 niños menores de 15 años. El muestreo se hizo por estratificación de acuerdo a los grupos de etarios. El tamaño de la muestra definitiva fue de 401, de los cuales 31 eran menores de 2 años, 166 individuos con edades entre 2-6 años y 204 entre 7-14 años. El consentimiento escrito para participar en el estudio se les solicitó a sus respectivos padres, siendo concedidos por todos ellos.

Para la agrupación en edades se usaron las escalas que utiliza el Ministerio de Salud y Desarrollo Social (MSDS) de la República Bolivariana de Venezuela y el Sistema de Vigilancia Alimentaría Nutricional (SISVAN) del Instituto Nacional de Nutrición (INN), que los agrupa en < 2 años, 2-6 años y de 7-14 años.

Análisis Estadísticos

El análisis estadístico de los datos se realizó con la ayuda de los programas estadísticos SAS y SPAD. Las variables incluidas en el estudio fueron: hemoglobina, hematocrito, glóbulos blancos, presencia de parásito intestinal y tipo de parásito intestinal, presencia de anemia, deficiencia de hierro.

La cantidad de niños a los cuales se les practicó extracción de sangre fue de 401, sin embargo 61 de esos niños no llevaron las muestras de heces, razón por la cual no se les pudo aplicar las pruebas para detectar parasitosis intestinales. Esto trajo como consecuencia que los análisis en los cuales se estudiaban variables relacionadas con los parásitos se realizaron sobre los 340 niños con muestras de heces, y a lo largo del estudio se va observar diferencias entre los tamaños de muestras.

En cuanto a la parasitosis intestinal, el 79,01% de los niños estudiados resultaron positivos de los cuales, el 38,13% resultaron monoparasitados y el 40,88% poliparasitados. Los parásitos predominantes fueron: Áscaris Lumbricoides, Trichuris Trichiura y Blastocistis Hominis con 51,66%, 42,82 % y 16,02 % respectivamente (Figura 1).

Es importante destacar que en los niños con parasitosis intestinal, fue muy común la presencia de más de un parasito intestinal, tal como lo muestra la figura 2.

Figura 1: Distribución porcentual de parasitosis, según grupo etario y población total.

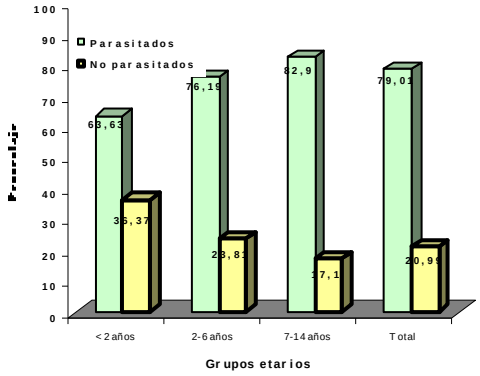
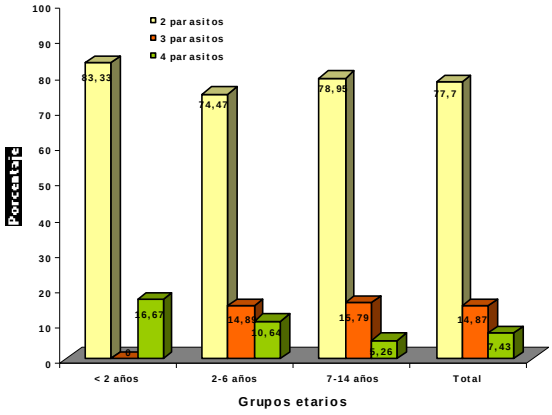


Figura 2. Distribución porcentual de poliparasitosis, según grupo etario y población total, de la población.



El análisis de los datos se inició aplicando la técnica multivariante denominada Análisis de Componentes Principales con el objeto de reducir la dimensión del número de variables. El análisis generó las salidas presentadas en las Figuras 3, 4 5.

En la salida presentada en la Figura 3 se muestra la matriz de correlaciones entre las variables incluidas en el análisis, se observa que a excepción de la variable edad, el resto de las variables entre sí presentan una correlación positiva, las correlaciones más altas se observan en las variables hematocrito con la variable edad y hemoglobina con edad.

Figura 3. Matriz de Correlaciones, generadas por el programa SPAD

```
MATRICE DES CORRELATIONS
     |   hist   edad   hb     hto    vcm
ferr
-----+----------------------------------------
-
hist |   1.00
edad |  -0.12   1.00
hb   |   0.05   0.54   1.00
hto  |   0.02   0.58   0.92   1.00
vcm  |   0.00   0.40   0.53   0.53   1.00
ferr |   0.02  -0.12  -0.09  -0.10  -0.06
1.00
-----+----------------------------------------
-
```

En la salida presentada en la figura 4 se observa que el primer autovalor explica 46.5 % del total de la variabilidad presente en los datos, el segundo autovalor explica el 17.49 % del total de la variabilidad presente en los datos y el tercer autovalor explica el 16.18 %. Los tres juntos explican el 80.23 % de la variabilidad total presente en los datos. Al aplicar el criterio del codo seleccionamos tres de ellos.

En la salida presentada en la Figura 4 se suministra información que nos ayuda a calcular el porcentaje de variación de cada variable explicada por cada componente. Nos muestra por ejemplo que el porcentaje de variabilidad de la variable edad explicada por la primera componente es aproximadamente 54.7 %.

En la salida presentada en la Figura 4 se suministra información que nos ayuda a calcular el porcentaje de variabilidad captada por cada factor que es explicada por cada variable. Si tomamos como punto de corte el valor 0.2, para observar cuales son las variables que más aportan a la formación del factor vemos que las variables que se destacan son edad del tipo, hemoglobina y hematocrito.

Figura 4. Histograma de Valores propios. Salida generada por el programa SPAD

HISTOGRAMME DES 6 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	2.7937	46.56	46.56	**********
2	1.0494	17.49	64.05	****
3	0.9711	16.18	80.23	****
4	0.5976	9.96	90.19	**
5	0.5065	8.44	98.64	**
6	0.0818	1.36	100.00	*

Se realizó un Análisis Correspondencia Binaria para establecer el grado de asociación entre la variable que registró la presencia o ausencia de parasitosis intestinal en los niños y la variable que registró la deficiencia de hierro presentada por el grupo de estudiados. Fueron excluidos de la base de datos aquellos niños que no se les aplicó exámenes de heces, quedando la base de datos reducida a 340 casos.

En la Figura 6 se presenta la tabla de distribución de frecuencias generada por el software estadístico, en ella se observa que de un total de 340 niños, 75 presentaron deficiencias de hierro de los cuales 18 niños no tienen parásitos y 57 niños si los tienen. Por otro lado 265 niños no presentaron deficiencias de hierro de los cuales 51 no tienen parásitos y 214 si los tienen, claramente se observa que dentro del grupo de niños que presentaron parásitos aproximadamente el 76 por ciento presenta parásitos.

Figura 6. Tabla de Frecuencias generada por el programa SAS

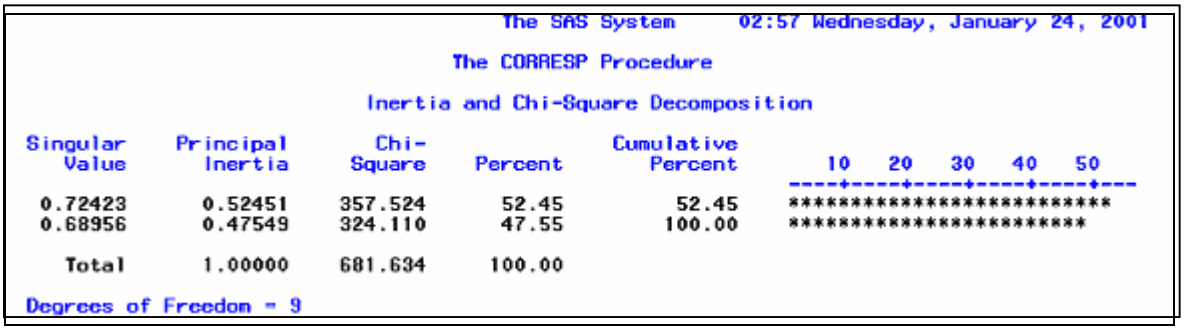
Burt Table

	Defi hierro	No defi Hierro	Niño No tiene paras	Niño tiene parasito
Defi hierro	75	0	18	57
No defi Hierro	0	265	51	214
Niño No tiene paras	18	51	69	0
Niño tiene parasito	57	214	0	271

En la Figura 7 se presenta la salida generada por el programa estadístico SAS, en donde se describe la inercia y la descomposición Chi cuadrado para el grupo de variables analizadas. Se observa que el primer eje captura el 52.45 por ciento de asociación entre las variables parasitosis intestinal y hierro en el grupo de niños y el segundo 47.55 por ciento de

asociación entre ambas variables. Los dos ejes capturan el 100 por ciento de asociación entre las variables.

Figura 7. Inercia y descomposición Chi cuadrados generada por el programa SAS



The SAS System 02:57 Wednesday, January 24, 2001

The CORRESP Procedure

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	10 20 30 40 50
0.72423	0.52451	357.524	52.45	52.45	**********
0.68956	0.47549	324.110	47.55	100.00	**********
Total	1.00000	681.634	100.00		

Degrees of Freedom = 9

La salida presentada en la Figura 8 suministra información sobre la contribución de cada una de las variables a la formación del factor. Se observa que las variables deficiencia de hierro, se ubican en sectores opuestos y esto influenciado por la variable presencia de parasitosis intestinal en el niño, por tanto la distribución en la deficiencia de hierro en los niños que presentan parásitos es diferente en comparación con los niños que presentan parásitos.

Figura 8 Aportes de las variables a la formación de los ejes generadas por el programa SPAD. Salida generada por el programa SPAD

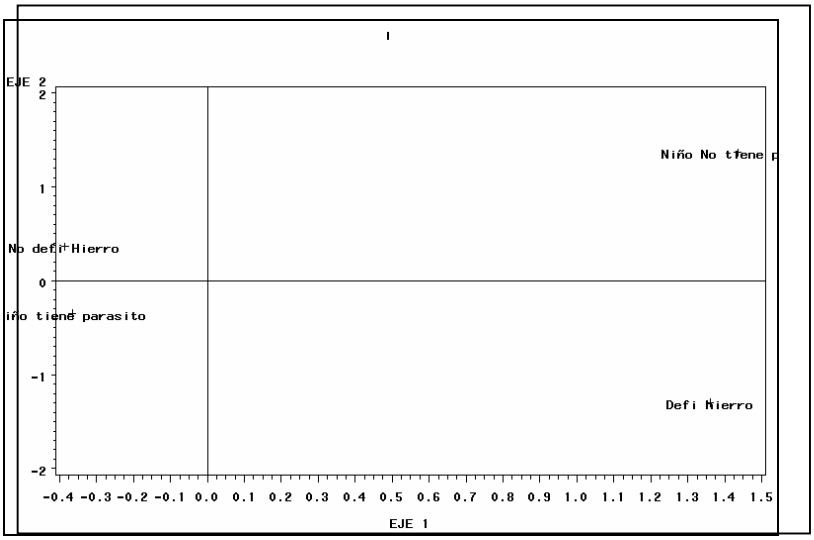
COORDONNEES DES VARIABLES SUR LES AXES 1 A 5
VARIABLES ACTIVES

VARIABLES	COORDONNEES					CORRELATIONS VARIABLE-FACTEUR					ANCIENS AXES UNITAIRES				
IDEN - LIBELLE COURT	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
hist - historia	0.02	-0.88	-0.43	-0.07	0.17	0.02	-0.88	-0.43	-0.07	0.17	0.01	-0.86	-0.44	-0.09	0.24
edad - edadniño	-0.74	0.22	0.06	-0.30	0.55	-0.74	0.22	0.06	-0.30	0.55	-0.45	0.21	0.07	-0.39	0.77
hb - hemoglobina	-0.92	-0.12	0.02	-0.14	-0.30	-0.92	-0.12	0.02	-0.14	-0.30	-0.55	-0.12	0.02	-0.18	-0.42
hto - hematocrito	-0.93	-0.09	0.03	-0.16	-0.25	-0.93	-0.09	0.03	-0.16	-0.25	-0.56	-0.09	0.03	-0.20	-0.35
vcm - volumencorpus	-0.71	-0.08	0.08	0.68	0.15	-0.71	-0.08	0.08	0.68	0.15	-0.43	-0.08	0.08	0.87	0.21
ferr - ferritina	0.17	-0.44	0.88	-0.06	0.04	0.17	-0.44	0.88	-0.06	0.04	0.10	-0.43	0.89	-0.08	0.06

Las conclusiones obtenidas anteriormente se corroboran con el gráfico perceptual presentado en el Gráfico 1**,** generado por el programa, vemos claramente como se distribuyen las categorías presencia de parásitos en los niños y deficiencia de hierro, observamos que al parecer los niños que presentan deficiencias en los niveles de hierro no presentan parasitosis intestinales, y los niños que no presentan deficiencias de hierro tiene

parásitos y esto tal vez sea motivado por la fase etaria de los infantes o se deba al hecho que los principales parásitos que se presentaron en los niños no afectan los niveles de hierro en los mismos.

Grafico 1. Gráfico Perceptual generado por el Análisis de Correspondencia Múltiple.
Salida generada por el programa SAS



Conclusiones

A través de los análisis estadísticos aplicados al conjunto de datos se observó que los niños que presentan deficiencias en los niveles de hierro no presentan parasitosis intestinales y esto tal vez se deba al tipo de parasito presentado en la mayoría de los niños.

Capitulo 5
Aplicación de la técnica multivariante Regresión Logística, en una investigación sobre citologías cérvico–vaginales

En este capítulo se estudia la aplicación del Análisis de Regresión Logística en una investigación sobre citologías cervico–vaginales, con el fin de determinar la relación entre el profesional encargado de la toma de muestras citológicas cérvico vaginales y la calidad de las mismas. Se recolectaron 230 citologías tomadas en diez comunidades rurales del Municipio Andrés Eloy Blanco del Estado Lara, durante el período 2002–2004. Las muestras fueron tomadas por 246 Bachilleres de Enfermería, 529 por bachilleres de Medicina, 150 por Médicos No Venezolanos y 52 muestras fueron tomadas por Médicos Venezolanos.

Para lograr describir la relación antes mencionada, se construyó un modelo de regresión logística, en el cual la variable dependiente dicotómica fue calmuest (calidad de la muestra citológica). A través del modelo construido con los datos se estimó que es aproximadamente 5 veces más probable que un médico No Venezolano tome una muestra citológica insatisfactoria comparada con las probabilidades de que un Médico Venezolano lo haga. Todo lo anterior con respecto a la muestra estudiada.

La Citología Cérvico–Vaginal (CCV) es el método de tamizaje o herramienta fundamental de elección para la pesquisa, detección temprana de lesiones precancerosas y del diagnóstico precoz del cáncer de cuello uterino.

Materiales y Métodos

El presente estudio es de carácter prospectivo longitudinal descriptivo, el universo formado por las mujeres en edad fértil que habitan en las comunidades rurales de Bojó, Cerro Blanco, El Caspito, El Portachuelo, La Bucarita, La Cruz, La Escalera, Miracuy, Monte Carmelo y San Antonio de Guache pertenecientes al Municipio Andrés Eloy Blanco del Estado Lara. La muestra es de carácter no probabilística o por conveniencia, observacional, opinática o accidental, constituida por todas aquellas mujeres que acudieron durante el período 2002–2004 a los ambulatorios rurales de esas comunidades a tomarse la citología cérvico vaginal (CCV).

Análisis Estadístico

En el Análisis de Regresión Logística aplicado al conjunto de datos, la variable dependiente dicotómica fue calmuest (calidad de la muestra citológica) la cual tomó el valor 1 si la muestra fue tomada insatisfactoriamente y 0 caso contrario. En la tabla 1 se muestran las variables incluidas como predictoras y su operacionalización.

Tabla 1. Variables incluidas como predictoras y su operacionalización

Variable	Operacionalización
"Br.Enfe"	0. (No tomo la muestra)
	1. (Si toma la muestra)
"Br.Medi"	0. (No tomo la muestra)
	1. (Si toma la muestra)
"Med.NoVenezolano"	0. (No tomo la muestra)
	1. (Si tomo la muestra)
"Med.Ven"	0. (No tomo la muestra)
	0. (Si tomo la muestra)

Fuente: Investigación Primaria

En la tabla 2 se muestran los resultados del modelo de regresión logística ajustado a los datos, Todas las categorías resultaron estadísticamente significativas ($p<0.001$), la categoría "medico venezolano" fue usada como celda de referencia. Se observa la significación de cada una de las categorías de la variable que representó el responsable de la toma de la muestra, todos los p-valores asociados a cada variable son menores que 0.05 por tanto todas son estadísticamente significativos al 95% de confianza, indicando esto que las variables "Bachiller de enfermería", "Bachiller de medicina", "MedNoVenezolano" están relacionadas con la probabilidad de obtener una muestra citolólgica insatisfactoria. Por otro lado todos los intervalos de confianza calculados para el exponencial β de cada variable no contiene el 1, indicando que las variables tienen una influencia significativa en la ocurrencia del suceso. (muestras citológicas insatisfechas).

Tabla 2. Valores Estimados de los coeficientes del modelo (β), Error Estándar (SE), Estadístico Wald (Wald) , grados de libertad (gl), p-valores (sig) y Odds Ratios (Exp(B))

Variables incluidas en el modelo	β	S.E.	Wald	g.l.	Significación	Exp(B)	Intervalo de confianza (95%) para Exp(B)	
							Límite Inferior	Límite Superior
Br.Enfermería	1.705	0.612	7.748	1	.005	5.500	1.656	18.267
Br.Medicina	1.619	0.603	7.205	1	.007	5.050	1.548	16.476
Med.NoVenezolano	1.712	0.624	7.536	1	.006	5.538	1.632	18.797
Constante	-2.793	0.595	22.058	1	.000	0.061		

Fuente: Investigación Primaria

Al analizar los resultados para la variable "Br. Enfermería" el valor positivo del coeficiente β indica que el riesgo que se corre de que una muestra tomada por un Bachiller de enfermería resulte insatisfactoria aumenta comparada con el riesgos que se corre cuando la

muestra citológica es tomada por un Medico Venezolano, observándose el mismo comportamiento para el resto de las variables, ya que todas tiene un valor positivo del coeficiente β.

Según los valores estimados por el modelo ajustado a los datos, la probabilidad de obtener una muestra insatisfactoria se puede estimar a través de la siguiente ecuación:

$$p = \frac{1}{1 + e^{-(-2.793 + 1.705\, BrEnferm\ + 1.619\, BrMedi\ + 1.712\, MedNoVen\ )}} \qquad (1)$$

A través de esta ecuación podemos estimar, por ejemplo, que la probabilidad de que la persona encargada de tomar la muestra citológica, la tome adecuadamente o satisfactoriamente.

Por otro lado los valores de la tabla 2, son utilizados para calcular los riesgos relativos asociados al modelo.

Conclusiones

A través de esta ecuación (1) podemos estimar que la probabilidad de que un Bachiller de Enfermería tome una muestra insatisfactoria es aproximadamente 0.2519, la probabilidad de que un Bachiller de Medicina tome una muestra citológica insatisfactoria es aproximadamente 0.2361, la probabilidad de que un Medico no Venezolano tome una muestra insatisfactoria es aproximadamente 0.2533 y la probabilidad de que un Médico Venezolano tome una muestra insatisfactoria es aproximadamente 0.057, por tanto es aproximadamente 5 veces más probable que un médico No Venezolano tome una muestra citológica insatisfactoria comparada con la probabilidad de que un Médico Venezolano lo haga. Todo lo anterior con respecto a la muestra estudiada. De los valores de la tabla 2 se estima, que los riesgos que se corren de que un muestra citológica tomada por un Médico No venezolano, un Bachiller de Medicina y un Bachiller de Enfermería son aproximadamente los mismos: 34%, 30 % y 33% respectivamente, mientras que los riesgos

que se corren de que una muestra citológica tomada por un Médico Venezolano resulte insatisfactoria es de aproximadamente el 6 %.

En cuanto al ajuste del modelo se observó que el p_valor asociado al estadístico Chi Cuadrado para el modelo (0.005) es menor que 0.1, por lo tanto al nivel de significación 0.01 se rechaza la hipótesis nula de que los parámetros asociados a las tres variables del modelo son nulos.

Programas de Análisis Estadístico utilizados para el desarrollo de las aplicaciones presentadas previamente.

El manejo de las técnicas de análisis multivariante se ha simplificado sustancialmente con el uso de programas de análisis estadístico para computadoras.

Entre los programas estadísticos utilizados para el procesamiento, de los datos de las aplicaciones presentadas en el presente trabajo se tienen SPSS (Statistical Package for the Social Sciences), SAS (Statistical Analisys System) y SPAD (Sisteme Portable pour L'Analyse de Données),

Para la aplicación presentada en el capitulo 2 fueron utilizados los programas SPAD y SPSS, para la aplicación presentada en el capitulo 3 fue utilizado el programa SAS y para la aplicación presentada en el capitulo 4 fue utilizado el programa SPSS.

Todos los programas utilizados están bajo la licencia de la Universidad Central de Venezuela y la Universidad de los Andes.

Referencias Bibliograficas

Berné Yelitza (2006). *'Evaluación Nutricional de una población rural menor de 15 años del Municipio Andrés Eloy Blanco del Estado Lara'. Trabajo de Ascenso. Universidad Centroccidental "Lisandro Alvarado"*. Decanato de Medicina. Barquisimeto (Venezuela).

Becue Bertaut Mónica, *'Manual de Introducción a los métodos Factoriales y Clasificación con SPAD'*.

Dallas E., *'Métodos Multivariados Aplicados al Análisis de Datos'*, Internacional Thomson Editores. 2000.

Daniel Peña, (2002) 'Análisis de Datos Multivariantes`, McGraw-Hill, Interamericana, 2002.

Daniel Pérez, (2004). *'Técnicas de Análisis Multivariante de Datos, Aplicaciones con SPSS'*. McGraw-Hill, Interamericana, 2002.

Dellan Rodríguez Graciela del Valle. (2004) *'Evaluación Nutricional de una población rural menor de 15 años del Municipio Andrés Eloy Blanco del Estado Lara'*. Trabajo de Ascenso. Universidad Centroccidental "Lisandro Alvarado". Decanato de Medicina. Barquisimeto (Venezuela).

Greenacre Michael J. *'Theory and Aplications of Correspondence Análisis'*. Academi Press editors , 1984.

Greenacre Michael J. *'Correspondence Análisisin Practice'*, primera edición. Academi Press editors 1984.

Idelfonso G. Esteban. *'Métodos Multivariantes para la Investigación Comercial', Editorial Ariel'*, s.a., 1989.

Johnson, R. A.; Wichern, D. W. (2001).*'Applied Multivariate Statistical Analysis'*,Prentice Hall. 5th edition,

Joseph F. Hair, Jr, Rolph E. Anderson, Ronald L. Tatham, Willliam C. Black,*'Análisis Multivariante, Hambridge M. Human Zinc defiicnecy'*. J Nutr 2000:130:1344S-1349S.

Merino Antonio Pardo.*' SPSS 11, Guía para el Análisis de datos'*, Prentice Hall. 2002.

Perez Cesar, *'El Sistema Estadístco SAS'*. McGraw-Hill, Interamericana, 2001.