Sistema para el análisis estadístico de técnicas multivariadas del rendimiento académico de los estudiantes de una institución de enseñanza superior*

Jorge Chue Gallardo
jchue@ulima.edu.pe
Emma Barreno Vereau
ebarreno@ulima.edu.pe
Rosa Millones Rivalles
rbmillon@ulima.edu.pe

Universidad de Lima

Resumen

Este trabajo de investigación presenta diferentes técnicas multivariadas aplicadas a las calificaciones obtenidas por los alumnos en las diferentes asignaturas de un plan de estudios, así como algunas variables socioeconómicas para detectar las posibles relaciones y comportamientos irregulares de alumnos, secciones y profesores de una institución de educación superior. El objetivo es alcanzar una gestión educativa eficiente, oportuna y confiable. Las técnicas multivariadas utilizadas son: análisis de correspondencias simple, análisis clúster, análisis de covariancia, análisis discriminante, regresión logística binaria y regresión logística ordinal y nominal.

Los resultados de la investigación indican que la técnica del análisis de covariancia permite comparar las diferentes secciones de un curso, eliminando el efecto del promedio ponderado acumulativo o de cualquier otra covariable. De igual modo, las otras técnicas multivariadas contribuyen a alcanzar el objetivo.

Palabras clave:
Análisis de correspondencias simple, análisis clúster, análisis de covariancia, análisis discriminante, regresión logística binaria y regresión logística ordinal y nominal. Modelos lineales generalizados y deviance.

* Este artículo no hubiera sido posible sin el apoyo del Instituto de Investigación Científica de la Universidad de Lima (IDIC), que auspició el proyecto "Sistema para el Análisis Estadístico con Técnicas Multivariadas del Rendimiento Académico de los Estudiantes de la Universidad de Lima", realizado por los autores. Del mismo modo, agradecen sinceramente a todos aquellos que de alguna manera los apoyaron en el desarrollo del presente artículo.

1. Introducción

El proceso de toma de decisiones en una institución educativa debe estar basado en métodos cuantitativos, especialmente en aquellos que pertenecen a la ciencia de la estadística. Estos métodos permiten adoptar las mejores decisiones cuando se observan y analizan escenarios que se desarrollan bajo condiciones de incertidumbre, aleatoriedad y variabilidad.

Durante un semestre se registra información de diferentes variables asociadas al rendimiento académico de los alumnos. Algunas de estas variables son: número de desaprobados en las diferentes asignaturas del plan de estudios, cursos con mayor número de desaprobados por semestre, porcentaje de desaprobados por profesor, nivel socioeconómico de los alumnos, condición laboral de los alumnos, categoría del colegio de procedencia, etcétera. En este contexto, el objetivo principal del proyecto es proporcionar un sistema que permita analizar las posibles relaciones entre las notas de las asignaturas obtenidas por los alumnos y algunas variables socioeconómicas de los mismos alumnos con la ayuda de técnicas multivariadas para detectar cursos, alumnos y secciones que presenten comportamientos estadísticos irregulares.

2. Técnicas multivariadas

Las técnicas multivariadas son un conjunto de métodos estadísticos que analizan simultáneamente medidas de diferentes variables de cada individuo u objeto en estudio. Las variables son aleatorias, pueden o no tener distribución normal multivariada y deben estar interrelacionadas de tal forma que sus efectos no pueden ser interpretados separadamente.

Las técnicas multivariadas utilizadas en el presente trabajo son:

- Métodos de interdependencia: análisis de correspondencias y análisis clúster.
- Métodos de dependencia: análisis de covariancia, análisis discriminante y regresión logística.

Los cálculos de las estadísticas asociadas a cada técnica multivariada son laboriosos y complejos en algunos casos, desde que la cantidad de datos que se analizan es generalmente grande con numerosas operaciones matriciales. Para facilitar estos cálculos se usaron los siguientes paquetes estadísticos: Minitab® , SPSS® y R.

2.1 Análisis de Correspondencia Simple

El Análisis de Correspondencia Simple (ACS) es una técnica estadística utilizada para analizar las relaciones de dependencia e independencia de un conjunto de variables categóricas a partir de datos originales. El ACS permite construir un diagrama cartesiano, denominado mapa de percepción de las categorías de varias variables; de forma que la proximidad entre los puntos representados está relacionada con el nivel de asociación entre dichas categorías.

Los principales objetivos del ACS son describir las asociaciones existentes entre las diferentes categorías de las variables analizadas y representar gráficamente, en un espacio bidimensional, las asociaciones existentes entre las diferentes categorías de las variables analizadas. Los tipos de datos utilizados son variables cualitativas nominales o variables cuantitativas continuas categorizadas haciendo uso de intervalos.

La metodología del ACS es aplicada a una tabla de contingencia de dimensiones $r \times c$ (r filas y c columnas) de la manera siguiente:

- Se calcula la tabla de frecuencias relativas, F de las observaciones.
- Se realiza la prueba Chi–Cuadrado para comprobar la existencia de dependencia entre las variables analizadas. La existencia o no de algún tipo de relación entre las variables X e Y se analiza mediante la prueba de hipótesis sobre la independencia de dichas variables. Estas hipótesis son dos: hipótesis nula (Ho) e hipótesis alternativa (Ha). Por consiguiente:

  Ho: Las variables del estudio X e Y son independientes

  Ha: Las variables X e Y son dependientes

- Se calcula la tabla estandarizada Z, de frecuencias relativas de las mismas dimensiones de la tabla original, $r \times c$, dividiendo cada celda de F por la raíz cuadrada de los totales de su fila y columna.
- Se calculan los h (normalmente $h = 2$) vectores propios ligados a valores propios mayores, pero distintos de la unidad, de las matrices ZZ^T y Z^TZ.

En el presente estudio se realizó el Análisis de Correspondencia Simple con la ayuda del software estadístico Minitab. Los reportes de interés emitidos por el mencionado software son *Row profiles;* en relación con las categorías de la variable ubicada en las filas de la tabla de contingencia, nos proporciona el porcentaje relacionado con cada una de las categorías de la variable ubicada en las columnas.

Los column profiles por su parte nos indica, con relación a las categorías de la variable ubicada en las columnas de la tabla de contingencia, el porcentaje relacionado con cada una de las categorías de la variable ubicada en las filas.

También tenemos el *analysis of contingency table,* que nos señala la importancia de cada una de las dimensiones al momento de explicar las dependencias observadas, así como las proporciones de inercia acumulada que ayudan a decidir el número mínimo de dimensiones necesario para explicar dichas dependencias.

Este análisis presenta *las row contributions,* que miden la importancia de cada una de las variables categóricas ubicadas en las filas de la tabla de contingencia en la construcción de los componentes o ejes factoriales construidos por el Análisis de Correspondencias Simple.

Las column contributions miden la importancia de cada una de las variables categóricas ubicadas en las columnas de la tabla de contingencia en la construcción de los componentes o ejes factoriales construidos por el Análisis de Correspondencias Simple.

2.2 Análisis clúster

Con respecto al análisis clúster, que es otra técnica multivariante, este nos permite clasificar objetos, casos o variables en grupos homogéneos llamados conglomerados (clústers), con respecto a algún criterio de selección predeterminado. El análisis clúster es utilizado en este trabajo para identificar y explicar los diferentes grupos que conforman las entidades en estudio, siendo estas agrupaciones realizadas de acuerdo con la similitud existente entre ellas.

El estudio tuvo como finalidad ofrecer una metodología para la realización de agrupaciones y, de esta forma, poder realizar análisis personalizados de acuerdo con las características que presente cada grupo formado.

Al formar los clúster o grupos se busca cumplir que cada elemento pertenezca a uno, y solo uno, de los grupos formados; que los objetos dentro de cada grupo (conglomerado) sean similares entre sí, es decir que exista una alta homogeneidad interna; y que los elementos dentro de cada grupo sean diferentes a los elementos de los otros conglomerados, es decir que exista una alta heterogeneidad externa.

Sobre los tipos de datos, no existe ningún tipo de restricción o condición que estos deban cumplir para proceder a aplicar la técnica, por lo

que casi con cualquier tipo de datos y prácticamente sin preparación previa de estos, se puede realizar un análisis de tipo clúster.

La metodología aplicada en el estudio es la siguiente:

- Planteamiento del problema para ser abordado mediante el análisis clúster; lo más importante es la selección de las variables en las que se basa la agrupación.
- Establecer medidas de semejanza y de distancia entre los elementos u objetos por clasificar en función del tipo de datos analizado. La similaridad o similitud es una medida de semejanza entre los elementos que van a ser agrupados; lo más común es medir la semejanza en términos de la distancia entre los pares de ellos. Así, los elementos con distancias reducidas entre ellos son más parecidos entre sí que aquellos con distancias mayores y se agruparán, por lo tanto, dentro del mismo clúster. Los tres métodos empleados en la medición de la similitud son: las medidas de correlación, las medidas de distancia (variables métricas) y las medidas de asociación (variables categóricas). Como las medidas de distancia son sensibles a la diferencia de escalas o de magnitudes hechas entre variables, a veces, es necesaria la estandarización de datos para evitar que las variables con una gran dispersión tengan un mayor efecto en la similaridad. La forma de estandarización más común es restarle a cada observación la media de la variable y este resultado dividirlo entre su desviación estándar.
- Analizar los métodos de clasificación haciendo especial énfasis en los métodos jerárquicos aglomerativos y en el algoritmo de las k-medias, y determinar el número de grupos. Estos dos procedimientos dependerán de los resultados que se obtengan y de la interpretación derivada de ellos; los dos tipos de procedimientos de agrupación son los jerárquicos y los no jerárquicos.

Los procedimientos jerárquicos se caracterizan por el desarrollo de una jerarquía o estructura de árbol (dendograma). De este modo, los clústers están formados solamente por la unión de los grupos existentes, así cualquier miembro de un clúster puede trazar su relación en una ruta que no se rompe y que comenzaría con una simple relación. Los métodos jerárquicos pueden ser por aglomeración o por división.

Entre los procedimientos no jerárquicos, el más usado es la Agrupación K – medias. Este algoritmo requiere un único parámetro, K, el número de agrupamientos que debe encontrar. Se aplica inicializando arbitrariamente los centros de los K grupos; luego se asigna cada patrón

al grupo más cercano y se recalculan los centros sobre la base de esta asignación. Algunos patrones pueden cambiar de agrupamiento y, en consecuencia, los centros de estos; si esto ocurre, se trata de repetir lo realizado hasta que no se modifiquen los centros. Cuando no haya modificaciones se considera que se ha encontrado una buena partición y se termina el agrupamiento.

No existen criterios objetivos y ampliamente válidos para la elección del número óptimo de grupos, pero existe una idea importante: A medida que se van formando grupos, estos son menos homogéneos, pero la estructura es más clara; por tanto, se puede fijar un objetivo, que es identificar el punto de equilibrio entre la estructura incompleta y la estructura mezclada o confusa.

2.3 Análisis de covariancia

En el proceso de la gestión educativa con frecuencia se plantean las siguientes interrogantes: ¿es uniforme la calificación que reciben los alumnos en las diferentes secciones de un curso?, ¿enseñan todos los profesores de un curso al mismo nivel?, ¿cómo eliminar la presencia de factores extraños (por ejemplo, en algunas secciones los promedios ponderados son altos en tanto que en otras son bajos) que no permiten analizar únicamente el nivel de enseñanza que se ofrece en las diferentes secciones de un curso?

Para responder las anteriores preguntas se eligió utilizar el análisis de covariancia (Ancova), técnica estadística que es una combinación de análisis de regresión y análisis de variancia.

El Ancova es utilizado en este trabajo para ajustar las notas finales de los alumnos de un curso mediante la covariable promedio ponderado acumulativo (PPA) en un diseño completamente al azar. Este ajuste es realizado con la finalidad de comparar los rendimientos académicos de los alumnos de las diferentes secciones de un curso sin considerar el efecto del PPA. El Ancova aprovecha la relación existente entre las notas obtenidas por los alumnos y sus correspondientes promedios ponderados acumulativos con la finalidad de reducir el error experimental y, de esta manera, hacer del Ancova una herramienta apropiada para el análisis de las diferentes secciones.

Luego, el modelo del Ancova para una covariable en un diseño completamente al azar es:

$$Y_{ij} = \mu + \tau_i + \beta(X_{ij} - \overline{X}_{..}) + e_{ij} \qquad (1)$$

Y_{ij} es la variable respuesta de tipo cuantitativa. i = 1, 2,..., r representa una sección; j = 1, 2,..., n_i. representa al alumno dentro de la sección. μ es el promedio general. β es el coeficiente de regresión de la covariable X y mide el efecto de X en la variable respuesta $Y_{i\,j}$. La covariable cuantitativa $X_{i\,j}$ está relacionada linealmente con $Y_{i\,j.}$ τ_i es el efecto del i-ésimo tratamiento en la variable respuesta Y_{ij}. El efecto τ_i es una variable binaria. e_{ij} es el término aleatorio.

Las suposiciones del modelo y otras propiedades se pueden encontrar en Neter, J. y Wasserman, W. (1974).

Las hipótesis del modelo del Ancova son:

Ho: $\tau_1 = \tau_2 = \ldots = \tau_r = 0$

Ha: Al menos un $\tau_i \neq 0$

Si los tratamientos difieren, la siguiente etapa en el análisis es la comparación por pares de los efectos de los tratamientos $\tau_k - \tau_h$ (la distancia vertical entre las líneas de regresión de tratamientos). Contrastes más generales de los τ'_i s también pueden ser utilizados.

La solución de un problema de Ancova puede realizarse por dos aproximaciones:

- Análisis de variancia Anova – Modelo lineal general.
- Regresión lineal múltiple.

La solución más difundida es la de Anova – Modelo lineal general, que se obtiene de Minitab, SPSS y R. Esta es la solución que se utiliza en el presente trabajo de investigación. Para la solución de la regresión lineal múltiple se recomienda el libro de Neter y Wasserman, quienes desarrollan en forma detallada esta aproximación.

2.4 Análisis discriminante

El análisis discriminante es un método que permite modelar la pertenencia a un grupo de individuos en función de datos de varias variables, además de predecir al grupo más probable de un individuo, conociendo solamente los valores de las variables que la caracterizan. Este método asigna objetos a grupos, estos objetos pueden ser personas, empresas, países, plantas, animales, etcétera.

El modelo de análisis discriminante está definido por una combinación lineal y se le conoce con el nombre de función discriminante.

En la figura 1 está representada la función D, que es una combinación lineal de las variables; sobre la función D se ha representado la proyección de las dos nubes de puntos, las líneas punteadas representan la ubicación proyectada de los puntos medios de cada grupo (centroides). La función lineal general se define como:

$$D = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon_{ij} \qquad (2)$$

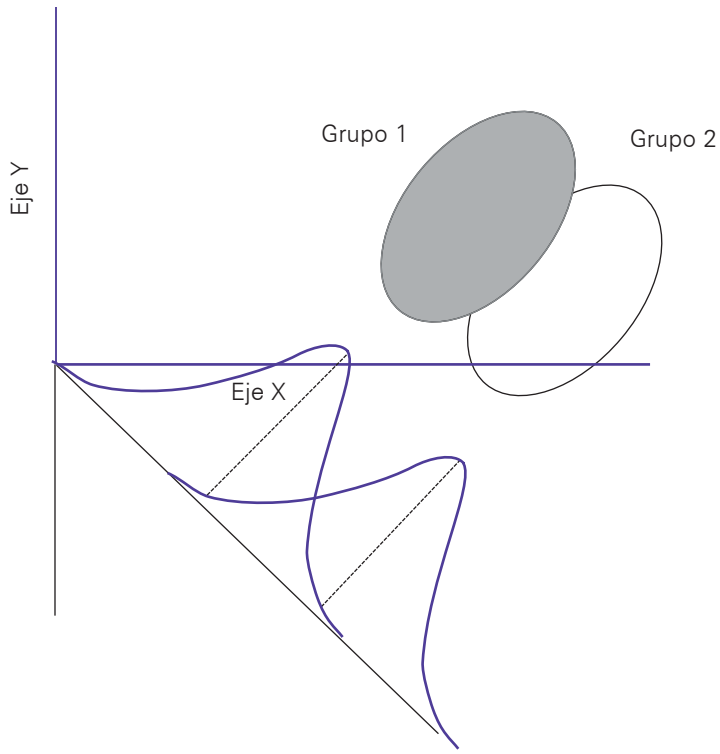

Figura 1. Representación de la función discriminante

Esta técnica parte de una tabla de datos de n individuos a los que se les ha medido k variables cuantitativas independientes que actúan como perfil de características de cada uno de ellos. Desde el punto de vista del análisis de varianza es saber si dos o más grupos son significativamente diferentes, si se prueba que la media de una variable es significativamente diferente en varios grupos, se puede asegurar que esta variable discrimina entre grupos.

Se recomienda considerar 20 sujetos por cada variable independiente, los tamaños de cada grupo pueden ser distintos; cuando ocurre esto, el grupo de menor tamaño debe ser al menos tres veces superior al número de variables predictoras incluidas en el análisis.

Las variables discriminantes deben ser métricas y la dependiente categórica, si la variable dependiente no es categórica hay que transformarla mediante las técnicas de exploración.

2.5 Regresión logística binaria

La regresión logística binaria es un tipo de regresión en que la variable respuesta es una variable binaria (0 = si el resultado es un fracaso y 1 = si el resultado es un fracaso) y las variables predictoras (independientes) son continuas o categóricas. La presentación del modelo de regresión logística es realizada desde la perspectiva de los modelos lineales generalizados.

Dobson (2002) señala lo siguiente, sean N variables aleatorias independientes $Y_1, Y_2, \ldots, Y_N$ correspondientes al número de éxitos en N diferentes subgrupos o estratos. La tabla 1 presenta la tabulación de los datos por subgrupos.

Tabla 1
Frecuencias para N variables aleatorias binomiales

	Subgrupos			
	1	2	....	N
Éxitos	Y_1	Y_2	....	Y_N
Fracasos	$n_1 - Y_1$	$n_2 - Y_2$	....	$n_N - Y_N$
Total	n_1	n_2	....	n_N

Dobson (2002) indica que el objetivo de la regresión logística es explicar la proporción de éxitos, $P_i = Y_i / n_i$, en cada subgrupo en términos de variables continuas, niveles de factores u otro tipo de variable explicativa que caracterice el subgrupo. Como $E(Y_i) = n_i\pi_i$ y por tanto $E(P_i) = \pi_i$, se busca modelar las probabilidades π_i como $g(\pi_i) = X_i^T\beta$.

Donde X_i es un vector de variables explicativas (variables dummy para niveles de un factor y valores para covariables), β es un vector de parámetros y g es una función link. El caso más simple es el modelo lineal $\pi_i = X_i^T\beta$. Este modelo tiene la desventaja de que los valores ajustados

$X_i^T b$ pueden ser menores que cero o mayores que uno, lo que origina una inconsistencia con los valores de π, puesto que π es una probabilidad. Para asegurar que π se encuentre restringido al intervalo [0, 1] se utiliza el modelo logit para obtener:

$$\pi = \frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)} \tag{3}$$

La expresión (3) conduce a la función link (relación)

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_1 + \beta_2 x \tag{4}$$

El término $\log\left(\frac{\pi}{1-\pi}\right)$ es denominado la función logit y se le interpreta como el logaritmo de las apuestas (odds). Dobson (2002) indica que la bondad del ajuste del modelo de regresión logística binaria es

$$D = 2[\ell(b_{max}) - \ell(b)] \tag{5}$$

Donde $\ell(b)$ representa el máximo valor del logaritmo de la función de verosimilitud para el modelo ajustado y $\ell(b_{max})$ el máximo valor del logaritmo de la función de verosimilitud para el modelo saturado. D se distribuye asintóticamente como ji-cuadrado con (N–p) grados de libertad con N= número de observaciones y p=número de parámetros estimados.

Las suposiciones del modelo y otras propiedades se pueden encontrar en Neter, J. y Wasserman, W.

2.6 Regresión logística ordinal y nominal

Cuando la variable respuesta es categórica y tiene más de dos posibles valores, dos posibles aproximaciones pueden aplicarse: uno es la generalización de la regresión logística binaria a más de dos categorías originando la regresión logística ordinal o nominal, y el otro, es el uso de modelos log-lineales para modelar las frecuencias con distribuciones de Poisson. La diferencia entre estas dos aproximaciones es que en la regresión logística ordinal o nominal una de las variables observadas es reconocida como la variable respuesta y las otras como variables predictoras, mientras que en los modelos log-lineales todas las variables son tratadas de la misma manera.

En este trabajo se utiliza la primera aproximación, es decir, la regresión logística ordinal o nominal. Dobson (2002) indica que la regresión logística nominal es utilizada cuando no existe un orden natural entre los valores de la variable respuesta. Un valor (o categoría) de la variable respuesta es elegido arbitrariamente como la *categoría de referencia*. Supóngase que esta es la primera categoría, entonces se pueden definir los logit para las otras categorías de la siguiente manera:

$$\log it(\pi_j) = \log\left(\frac{\pi_j}{\pi_1}\right) = x_j^T \beta_j \quad \text{para } j = 2, ..., J \tag{6}$$

Los estimadores de las probabilidades π_j son:

$$\hat{\pi}_1 = \frac{1}{1 + \sum_{j=2}^{J} \exp(x_j^T b_j)} \qquad \hat{\pi}_j = \frac{\exp(x_j^T b_j)}{1 + \sum_{j=2}^{J} \exp(x_j^T b_j)} \quad j = 2, ..., J. \tag{7}$$

La bondad del ajuste del modelo de regresión logística nominal es medida por (5).

En la regresión logística nominal es más fácil interpretar los efectos de las variables explicatorias en términos de las razones de apuestas en lugar de los parámetros β propiamente dichos. Por ejemplo, sea una variable respuesta con J categorías y una variable explicatoria o predictiva binaria x con dos valores 1 (efecto presente) y 0 (efecto ausente), entonces la razón de apuestas al efecto j (j = 2, 3, …, J) relativo a la categoría de referencia j = 1 es

$$OR_j = \frac{\pi_{jp}}{\pi_{ja}} \Big/ \frac{\pi_{1p}}{\pi_{1a}} \tag{8}$$

donde $\pi_{j\,p}$ y $\pi_{j\,a}$ denotan las probabilidades de la categoría de respuesta j (j = 1, 2, …, J), de acuerdo a si está presente o ausente el efecto de la variable explicatoria x.

Por lo tanto:

$$\log OR_j = \log\left(\frac{\pi_{jp}}{\pi_{1p}}\right) - \log\left(\frac{\pi_{ja}}{\pi_{1a}}\right) = \beta_{1\,j} \tag{9}$$

Intervalos de confianza para OR_j pueden calcularse aplicando la formula

$$\exp[b_{1j} \pm z_{(1-\alpha/2)} \times SE(b_{1j})] \tag{10}$$

3.Sistema de análisis estadístico

En Incose (en línea) aparece la siguiente definición de sistema:

> Un sistema es una construcción o una colección de diversos elementos que juntos producen resultados no obtenibles si se consideran a dichos elementos por separado. Los elementos, o las piezas, pueden incluir personas, hardware, software, instalaciones, políticas, y documentos; es decir, todo lo que se requiere para producir resultados del sistema-nivel. Los resultados incluyen calidad del nivel del sistema, propiedades, características, funciones, comportamiento y rendimiento del nivel de sistema. El valor añadido por el sistema en su totalidad, más allá de lo que contribuye independientemente cada uno de los elementos, es creado principalmente por la relación entre los elementos; es decir, cómo ellos están interconectados. [Traducción de los autores.]

Este es el significado que se aplica en este trabajo de investigación; es decir, plantear un modelo de sistema que estructure y organice los diferentes elementos que existen en una institución de enseñanza superior, desde la perspectiva del análisis estadístico con técnicas multivariadas, para su gestión educativa.

En la figura 2 se presenta el modelo del sistema de análisis estadístico. El sistema tiene como entrada los datos, la información que las unidades administrativas posean de los alumnos y las encuestas que sean necesarias aplicar para obtener la información pertinente al estudio. Los datos corresponden a los alumnos, los profesores y las asignaturas. La parte de procesamiento consiste en identificar el problema que se desea analizar y a continuación elegir la técnica multivariada apropiada. La salida del sistema está compuesta por modelos e indicadores que permitirán a las autoridades tomar las decisiones que correspondan.

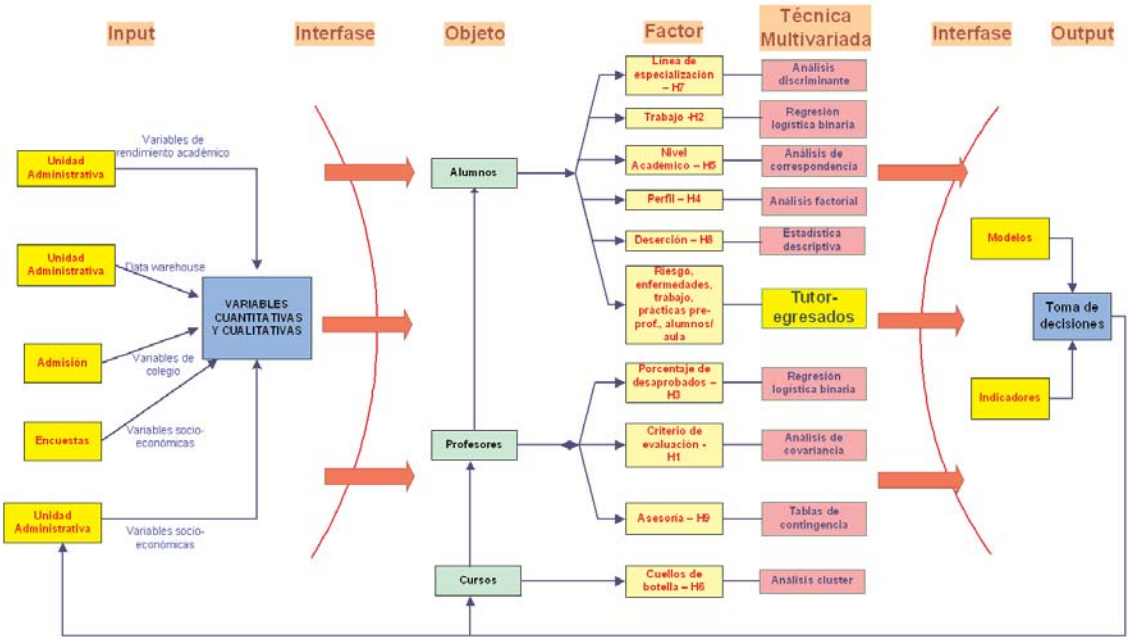

Figura 2. Modelo del Sistema de Análisis Estadístico

4. Resultados

4.1 Análisis de correspondencia simple

El análisis de correspondencia simple fue utilizado para determinar la existencia y el tipo de relación entre las siguientes variables de un periodo o semestre académico:

- Nivel socioeconómico y rendimiento académico de alumnos.
- Categoría del colegio de procedencia y rendimiento académico de los alumnos.
- Tipo de postulante y rendimiento académico de los alumnos.
- Condición laboral y ránking académico de los alumnos.

Las hipótesis de trabajo para nivel socioeconómico y rendimiento académico de alumnos fueron:

Ho: El nivel socioeconómico y el rendimiento académico de los alumnos son independientes.

Ha: El nivel socioeconómico y el rendimiento académico de los alumnos son dependientes.

Luego de procesar los datos con Minitab se obtuvo la figura 3, que se presenta a continuación.

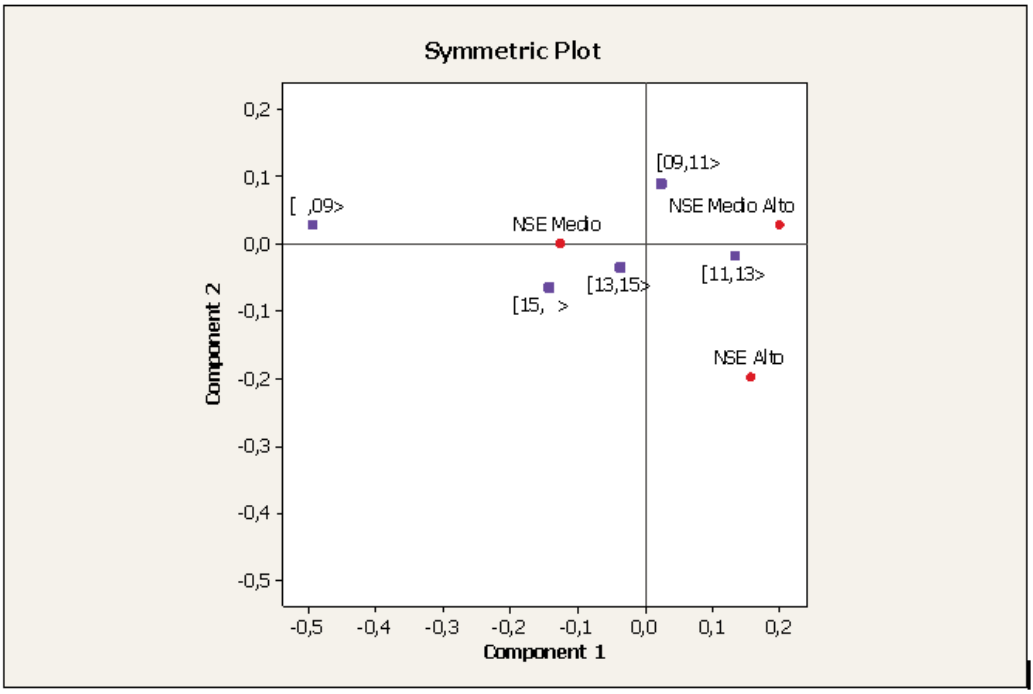

Figura 3. Mapa perceptual – Nivel socioeconómico y rendimiento académico

De la figura 3 se deduce lo siguiente:

- El nivel socioeconómico del alumno tiene relación con su rendimiento académico.
- Los alumnos del nivel socioeconómico medio se encuentran más asociados a obtener un promedio ponderado académico de 13 a más.
- Los alumnos de nivel socioeconómico medio alto se encuentran más asociados a obtener un promedio ponderado de 09 a 13.
- Los alumnos del nivel socioeconómico alto se encuentran más asociados a obtener un promedio ponderado académico de 11 a 13.

4.2 Análisis clúster

Con respecto al análisis clúster, se analizó la similitud existente en las variables para lograr la:

- Clasificación de los cursos, según el número total de alumnos matriculados por primera, segunda y tercera vez.
- Clasificación de los alumnos de los distintos niveles por los promedios obtenidos.

Para la "clasificación de los cursos, según el número total de alumnos matriculados por primera, segunda y tercera vez", se utilizó la siguiente información: relación de cursos de una de las unidades académicas en estudio, considerando código y nombre, número total de alumnos matriculados en los mencionados cursos, número total de alumnos matriculados por primera, segunda y tercera vez. Se consideraron los cursos dictados durante cinco ciclos académicos.

Cabe mencionar que en la unidad académica se dictan cursos de especialidad que pertenecen a la misma unidad académica en estudio y cursos comunes que pertenecen a la unidad académica en estudio y a otra unidad académica afín.

Debido a la naturaleza de los datos, previamente se realizó un análisis exploratorio de los datos y una estandarización de estos. Luego de procesar los datos con Minitab y mediante el uso de la medida de similitud Distancia Euclídea, y el método de agrupación Ward Linkage, se obtuvo una clasificación en seis grupos de acuerdo con las similitudes existentes entre los cursos impartidos en la unidad académica en estudio. La figura 4 presenta el Dendograma correspondiente a la clasificación de los cursos, según el número total de alumnos matriculados por primera, segunda y tercera vez.

- Grupo 01: Conformado por 4 cursos. En este grupo se encontraban los cursos con la mayor cantidad de alumnos matriculados, que presentaban una considerable cantidad de alumnos matriculados por segunda y tercera vez. Cabe destacar que la totalidad de estos cursos eran cursos de especialidad.
- Grupo 02: Conformado por 10 cursos. En este grupo se encontraban los cursos con una elevada cantidad de alumnos matriculados, pero no tanto como los del grupo 01, que presentaban una mayor cantidad de alumnos matriculados por segunda y tercera vez. En términos generales, podría considerarse a estos cursos como "cuellos de bote-

lla"; es decir, cursos que pueden ocasionar un retraso en la culminación del plan de estudios por parte de los alumnos de la unidad académica. Esto se apreció en la totalidad de cursos comunes.

- Grupo 03: Conformado por 8 cursos. En este grupo se encontraban los cursos con una elevada cantidad de alumnos matriculados, pero no tanto como los de los grupos 01 y 02, que presentaban una elevada cantidad de alumnos matriculados por segunda y tercera vez. Después de los cursos "cuellos de botella", estos cursos también representan una ligera dificultad para la culminación del plan de estudios por parte de los alumnos de la unidad académica. Cabe destacar que la mayoría de estos cursos son propios de la unidad académica en estudio.
- Grupo 04: Conformado por 12 cursos. En este grupo se encontraban los cursos con una cantidad estándar de alumnos matriculados, que presentaban una mediana cantidad de alumnos matriculados por segunda y tercera vez. Estos cursos también representan una ligera dificultad para la culminación del plan de estudios por parte de los alumnos de la unidad académica, pero menos considerable que las de los grupos 02 y 03. También aquí cabe destacar que la mayoría de estos cursos son cursos propios, con excepción de 3 cursos.
- Grupo 05: Conformado por 16 cursos. En este grupo se encontraban los cursos con una cantidad estándar de alumnos matriculados, que presentaban una baja cantidad de alumnos matriculados por segunda y tercera vez. Estos cursos no representan una dificultad para la culminación del plan de estudios por parte de los alumnos de la unidad académica. La gran mayoría de estos cursos son cursos propios de la unidad académica, con excepción de dos cursos.
- Grupo 06: Conformado por 26 cursos. En este grupo se encuentran los cursos con una cantidad poco significativa de alumnos matriculados, que presentan una baja o nula cantidad de alumnos matriculados por segunda y tercera vez. Estos cursos no representan una dificultad para la culminación del plan de estudios por parte de los alumnos de la unidad académica.

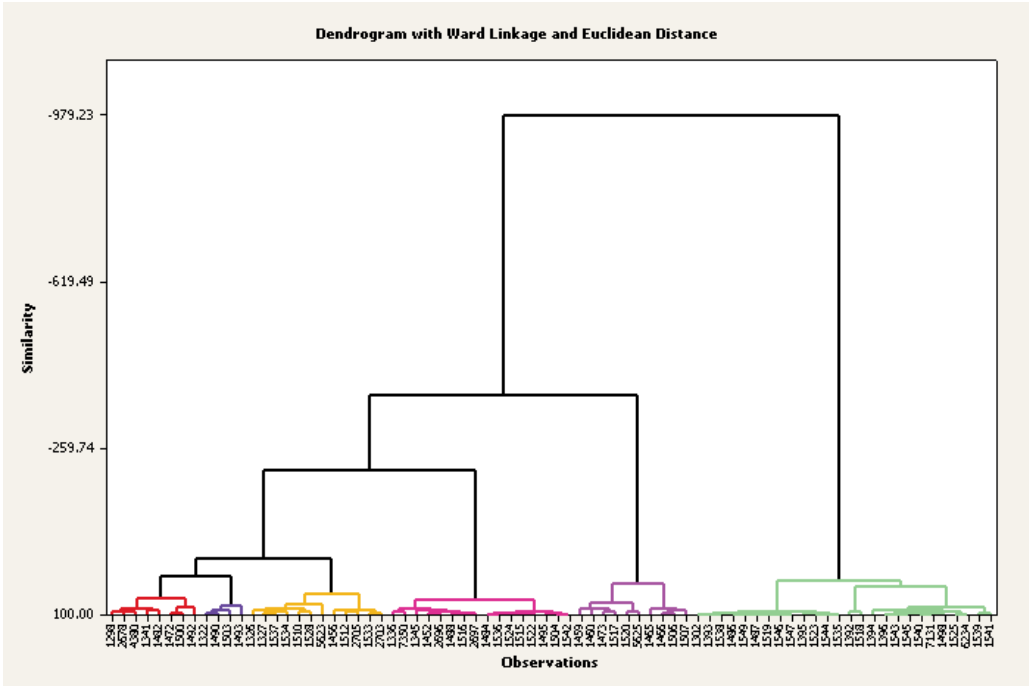

Figura 4. Dendograma – Clasificación de los cursos, según el número totaL de alumnos matriculados por primera, segunda y tercera vez.

4.3 Análisis de covariancia

A continuación se presentan los resultados de la aplicación del análisis de covariancia a las notas de los alumnos de la asignatura A de un semestre académico. La covariable utilizada fue el PPA (promedio ponderado acumulativo).

El curso A se ofreció en 13 secciones con 10 profesores. Algunos profesores dictaron 1 o 2 secciones. Las hipótesis de trabajo fueron las siguientes:

- Hipótesis nula: Los profesores del curso A tienen el mismo nivel de enseñanza y, por tanto, el mismo criterio para evaluar a sus alumnos.
- Hipótesis alternante: Al menos dos profesores son diferentes.

La información utilizada para realizar esta prueba fue: sección del curso, nota del examen final, promedio ponderado acumulativo. El análisis se realizó en dos etapas: la primera consistió en la prueba de normalidad de los datos y la segunda en el análisis de covariancia propiamente dicho.

Primera etapa:

Prueba de normalidad

- Solo se consideraron las notas en el intervalo [4, 17] para satisfacer la suposición de normalidad.
- Se aplicó la prueba de normalidad de Bonett-Seier de la curtosis de Geary porque es una prueba específica para la distribución normal. El reporte del software R, usando la prueba de Bonett-Seier, permitió concluir que los datos tienen distribución normal con p-value = 0.03673

Segunda etapa:

El reporte de Minitab es el siguiente:

```
General Linear Model: NOTAFINAL versus SECCIONCURSO
Factor         Type   Levels  Values
SECCIONCURSO fixed     13      301; 302; 303; 304; 305; 306; 307; 308; 309;
                               310; 311; 313; 316
Analysis of Variance for NOTAFINAL, using Adjusted SS for Tests
Source                DF     Seq SS     Adj SS   Adj MS      F         P
PPA antes              1     1882,15     784,99  784,99   187,11    0,000
SECCIONCURSO          12      182,02     182,02   15,17     3,62    0,000
Error                550     2307,38    2307,38    4,20
Total                563     4371,55
S = 2,04823   R-Sq = 47,22%   R-Sq(adj) = 45,97%
Term             Coef      SE Coef     T       P
Constant       -1,6769    0,9911   -1,69   0,091
PPA antes       1,00423   0,07341  13,68   0,00
```

En las figuras 5 y 6 se presentan las gráficas de los promedios de las 13 secciones del curso A, antes y después de realizar el ajuste con la técnica del análisis de covariancia.

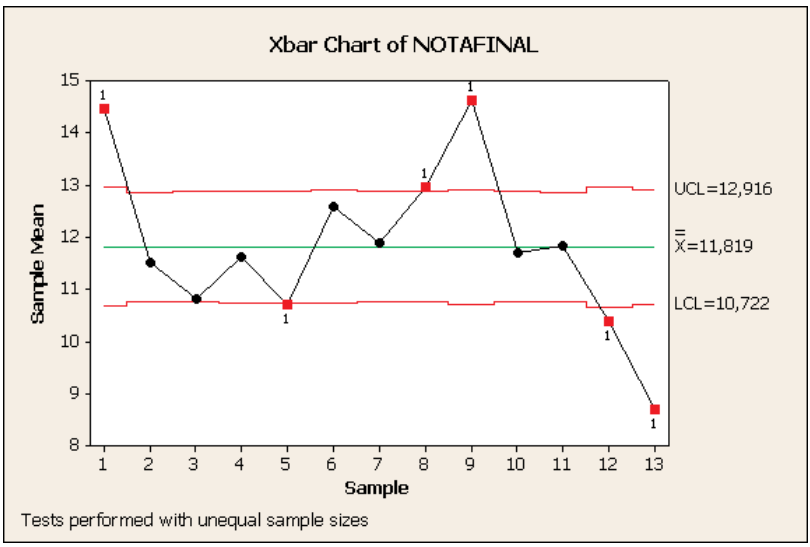

Figura 5. Gráfica de los promedios - Notas de las 13 secciones del curso A.

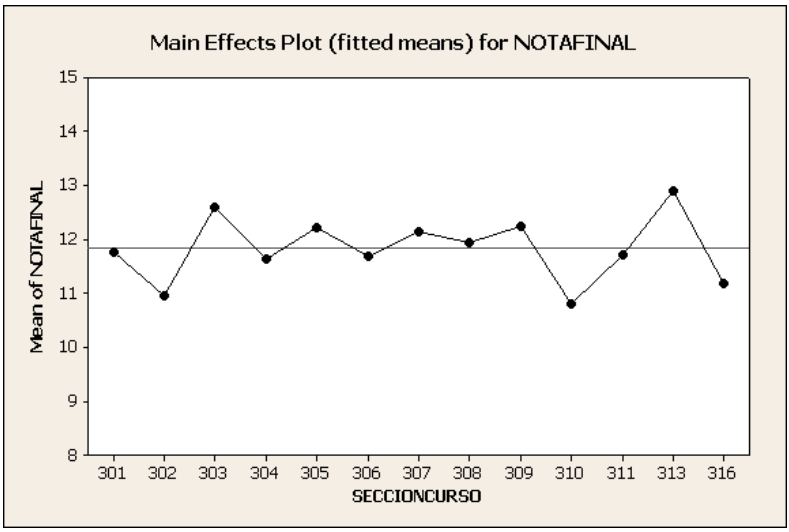

Figura 6. Gráfica de los promedios – Notas ajustadas de las 13 secciones del curso A.

El uso de la variable PPA es altamente significativa. Hay diferencias altamente significativas entre las diferentes secciones del curso A. El gráfico confirma la conclusión de que hay fuertes diferencias entre las secciones del curso A. La aplicación de la técnica multivariada del análisis de covariancia es sencilla y puede hacerse con R y Minitab.

4.4 Análisis discriminante

Esta técnica multivariada permite identificar los cursos más influyentes en cada área académica.

- Como variables independientes se han considerado las notas de los cursos A, B, C, D, E, F, G, H, I, J, K, L, M, O de los alumnos de tres promociones y del tercer al séptimo ciclo.
- Para crear la variable dependiente se utilizó la técnica del análisis clúster en la que, de acuerdo con la similitud existente entre las variables estudiadas, se han obtenido tres grupos diferenciados, como se observa en la figura 7; el grupo 01 corresponde a los alumnos que tienen las peores notas y son los que han sido desaprobados en dos o más cursos; el grupo 02 son mejores que el grupo anterior y el grupo 03 son los mejores alumnos.

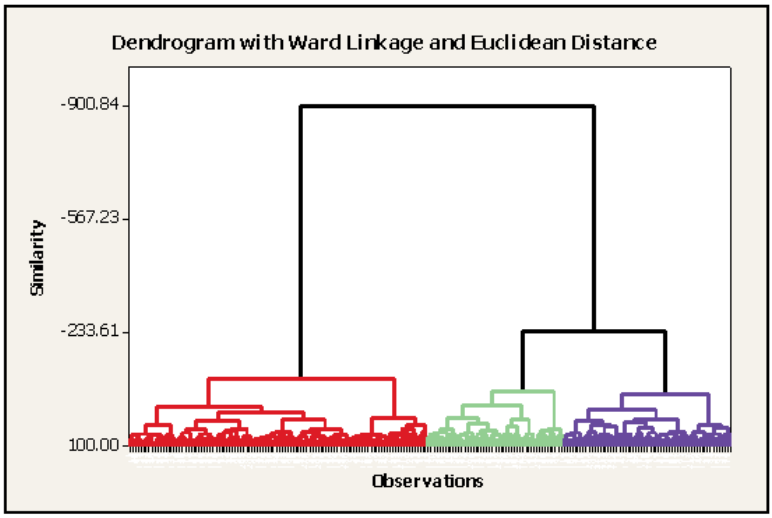

Figura 7. Clasificación de los alumnos según notas

- Se realizó la prueba de normalidad a las variables independientes, como se muestra en la figura 8.

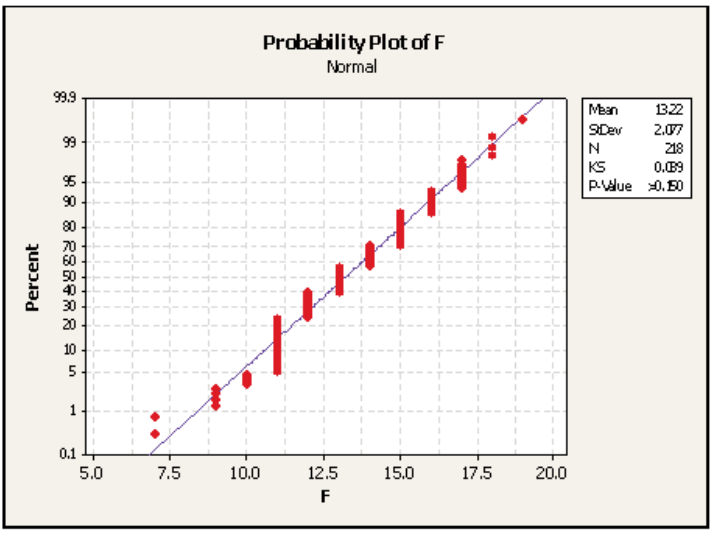

Figura 8. Prueba de normalidad para las notas del curso F

El valor de p-value mayor que 0.15 indica que las notas del curso F pueden ser modeladas por una distribución normal.

- Se aplicó el análisis discriminante utilizando SPSS y se obtuvieron tres grupos:
  - Grupo 1: Alumnos con menor rendimiento, los promedios de las notas de los cursos es de: 10 hasta 13.4.
  - Grupo 2: Alumnos con un mejor rendimiento que el grupo anterior, con promedios de 11.7 a 13.4.
  - Grupo 3: Alumnos con buen rendimiento, con promedios de 13.2 a 16.0.

  De los 218 registros, el 27% corresponde al grupo 1; 58,7% al 2 y 14,7% al 3.

Para formar la función discriminante fueron necesarios siete pasos, tal como se aprecia en la tabla 2; en el primer paso se ha incluido el curso B, en el segundo, el J, y así sucesivamente. Los cursos A, E, F, G, I, L y O han sido excluidos del análisis debido a que no son significativos. El valor mínimo para que ingrese una variable en el análisis es de 3.84 y el valor máximo para retirarla ha sido de 2.71. Los tres grupos conformados por el rendimiento académico están menos traslapados, según el

estadístico Wilks, ya que va disminuyendo en cada paso. En esta aplicación se realizó la prueba de la diferencia de promedios entre los tres grupos, encontrándose que hay diferencia significativa.

Step	Entered	Wilks' Lambda							
		Statistic	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	B	.508	1	2	215.0	104.142	2	215	.000
2	J	.370	2	2	215.0	69.006	4	428	.000
3	H	.311	3	2	215.0	56.246	6	426	.000
4	I	.281	4	2	215.0	47.011	8	424	.000
5	D	.264	5	2	215.0	39.859	10	422	.000
6	C	.250	6	2	215.0	34.977	12	420	.000
7	M	.237	7	2	215.0	31.466	14	418	.000

At each step, the variable that minimizes the overall Wilk´s Lambda is entered.

a. Maximum number of steps is 28.
b. Minimum partial F to enter is 3.84.
c. Maximum partial F to remove is 2.71.
d. F level, tolerance, or VIN insufficient for futher computation.

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.237	.305.167	14	.000
27	.865	30.833	6	.000

Tabla 2. Pasos para formar la función discriminante

En la prueba de Wilks se contrasta la significación de las dos funciones obtenidas. En la primera línea se contrasta la hipótesis nula de que el modelo completo (ambas funciones discriminantes tomadas juntas) no permiten distinguir las medias de los grupos; entonces, se concluye que en el modelo completo sí permite distinguir las medias de los grupos (Sig. = 0). En la segunda línea se contrasta que las medias de los grupos son iguales en la segunda función discriminante. De igual manera, se concluye que la segunda función discriminante permite observar diferencias en al menos dos grupos.

Las tablas 3 y 4 indican lo siguiente: las variables que discriminan más en la primera función son los cursos B, J y H.

Esto permite interpretar que los alumnos con notas altas en estos cursos son clasificados como los mejores alumnos, y aquellos con notas bajas son considerados en el primer grupo, es decir, los alumnos con dificultad en el aprendizaje.

Discr.	Function	
	1	2
1.00	–1.994	.437
2.00	.092	–.329
3.00	3.247	.523

Unstandardized canonical discriminant functions evaluated at group means

Tabla 3. Function at Group centroids function.

Curso	Function	
	1	2
B	.623	.075
C	.218	–.452
D	.160	.523
H	.312	.195
I	.212	–.715
J	.463	.098
M	.159	.569

Tabla 4. Standardized Canonical Discriminant Coefficients.

En la segunda función la mayor ponderación es la de los cursos I y C; es decir que alumnos con mayores notas en estos cursos son clasificados en el segundo grupo.

La clasificación correcta de los casos, en general, ha sido del orden del 89,4%; en el grupo 1 del 96,6%; en el grupo 2 con 84,4% y en el grupo 3 con 96,9%, como se observa en la tabla 5. Se concluye que hay una buena clasificación de los grupos de alumnos.

		Discr.	Predicted Group Membership			Total
			1.00	2.00	3.00	
Original	Count	1.00	56	2	0	58
		2.00	14	108	6	128
		3.00	0	1	31	32
	%	1.00	96.6	3.4	.0	100.0
		2.00	10.9	84.4	4.7	100.0
		3.00	.0	3.1	96.9	100.0

a. 89.4% of original grouped cases correctly classified.

Tabla 5. Porcentaje de casos clasificados por grupos.

La validación del modelo fue realizada como se observa en la tabla 6.

			Discr.	Predicted Group Membership			Total
				1.00	2.00	3.00	
Cases selected	Original	Count	1.00	32	2	0	34
			2.00	7	54	1	62
			3.00	0	0	17	17
		%	1.00	94.1	5.9	0	100.0
			2.00	11.3	87.1	1.6	100.0
			3.00	.0	0	100.0	100.0
	Cross-validated[a]	Count	1.00	30	4	0	34
			2.00	7	51	4	62
			3.00	0	0	17	17
		%	1.00	88.2	11.8	0	100.0
			2.00	11.3	82.3	6.5	100.0
			3.00	.0	0	100.0	100.0
Cases not selected	Original	Count	1.00	0	0	0	0
			2.00	0	0	0	0
			3.00	0	0	0	0
		%	1.00	0	0	0	100.0
			2.00	0	0	0	100.0
			3.00	0	0	0	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than case.
b. 91.2% of selected original grouped case correctly classified.
c. 0% of unselected original grouped cases correctly classified.
d. 86.7% of selected cross-validated grouped cases correctly.

Tabla 6. Validación del modelo

La tabla 6 indica que la clasificación correcta es de 91,2% y la tasa de aciertos en la validación es de 86,7%; es decir que se espera que la función discriminante obtenida clasifique correctamente al 86,7% de los futuros casos nuevos que se pretenden clasificar.

Finalmente, se concluye que las notas altas de los alumnos en los cursos J, H, I, D, C y M permiten clasificarlos como los mejores alumnos, en tanto que aquellos con notas bajas serán clasificados en el grupo 1, es decir, alumnos con problemas que retrasan la culminación de sus estudios.

4.5 Regresión logística binaria

La técnica de regresión logística binaria se presenta a continuación con dos aplicaciones. En ambos casos, se debe tener presente que la variable dependiente asume valores 0 y 1.

Aplicación 1

Esta primera aplicación busca determinar la relación entre los cursos que reciben los alumnos de la unidad académica en estudio y su capacidad para conseguir trabajo. En términos académicos, todos los cursos de la especialidad serían igualmente importantes, al menos los que se ofrecen en los últimos ciclos. Los datos utilizados corresponden a la promoción de un semestre académico formada por 39 alumnos.

- Hipótesis nula: Los cursos que ofrece la unidad académica son igualmente importantes para los recién egresados que desean conseguir trabajo.
- Hipótesis alternante: Los cursos que ofrece la unidad académica no son igualmente importantes para los recién egresados que desean conseguir trabajo.

La información utilizada para realizar esta aplicación fue:

1. Notas de todos los cursos según el plan de estudios de la promoción estudiada.
2. Situación laboral del egresado (trabaja = 1, no trabaja = 0).

Los resultados de Minitab indicaron lo siguiente:

					Odds	95% CI	
Predictor	Coef	SE Coef	Z	P	Ratio	Lower	Upper
Constant	0,906706	6,41973	0,14	0,888			
Curso 1	-0,694653	0,375103	-1,85	0,064	0,50	0,24	1,04
Curso 2	1,63172	0,835064	1,95	0,051	5,11	1,00	26,27
Curso 3	-0,705843	0,397185	-1,78	0,076	0,49	0,23	1,08

Log-Likelihood = -12,089
Test that all slopes are zero: G = 15,401, DF = 3, P-Value = 0,002

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	30,4086	32	0,547
Deviance	24,1789	32	0,838
Hosmer-Lemeshow	12,3166	8	0,138

Ejemplo de predicción:

Si un alumno tiene las siguientes notas: 16, 13 y 14 en los cursos 1,2 y 3, entonces la probabilidad de que tenga trabajo es:

$$\hat{\pi} = \frac{\exp(0.906706 - 0.694653 \times 16 + 1.63172 \times 13 - 0.705843 \times 14)}{1 + \exp(0.906706 - 0.694653 \times 16 + 1.63172 \times 13 - 0.705843 \times 14)} = 0.75451068 \quad (5)$$

El modelo de regresión logística binaria con las notas de los cursos 1, 2 y 3 es apropiado para modelar la variable binaria trabajar o no trabajar. (Deviance=24,1789 con p-value=0,838). Los resultados obtenidos son válidos únicamente para la promoción 2005-2. Puede suceder que otra promoción de egresados tenga otros resultados.

Aplicación 2

La segunda aplicación consiste en determinar el modelo que mejor se ajuste a la proporción de desaprobados en los exámenes finales de las diferentes secciones de un mismo curso en tres ciclos académicos consecutivos.

- Hipótesis nula: La proporción de desaprobados es igual en todas las secciones de un mismo curso.
- Hipótesis alternante: La proporción de desaprobados no es igual en todas las secciones de un mismo curso.

Tres modelos fueron utilizados para probar las hipótesis mencionadas.

- Modelo 1: logit $\pi_{j\,k} = \alpha_{\,j} + \beta_{\,k}$, α_j = las proporciones de desaprobados son distintas de semestre a semestre. β_k = los profesores tienen diferentes proporciones de desaprobados.
- Modelo 2: logit $\pi_{j\,k} = \alpha_{\,j}$, $\alpha_{\,j}$ = las proporciones de desaprobados son distintas de semestre a semestre.
- Modelo 3: logit $\pi_{j\,k} = \beta_{\,k}$, $\beta_{\,k}$ = los profesores tienen diferentes proporciones de desaprobados

La información utilizada para esta aplicación se presenta en la tabla 7. La variable $Y_{\,i\,k}$ representa el número de desaprobados en el i-ésimo ciclo académico con el k-ésimo profesor. La variable $n_{\,i\,k}$ representa el número de alumnos inscritos en el i-ésimo ciclo académico con el k-ésimo profesor.

Ciclo		Profesor					
		A	B	C	D	E	F
Ciclo Acad. 1	Y_{1k}	19	22	26	14	18	21
	n_{1k}	30	30	29	30	29	28
Ciclo Acad. 2	Y_{2k}	12	13	12	16	15	22
	n_{2k}	29	30	30	30	29	30
Ciclo Acad. 3	Y_{3k}	9	12	10	13	16	9
	n_{3k}	31	33	33	32	31	27

Tabla 7. Datos de la aplicación 2 de la regresión logística binaria.

La gráfica de la proporción de desaprobados se presenta en la figura 9.

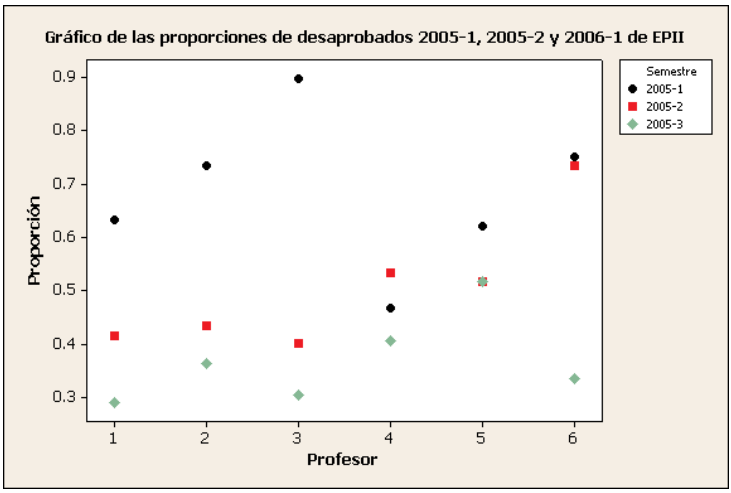

Figura 9. Gráfica de la proporción de desaprobados (tabla 7).

Los resultados que se obtuvieron luego de aplicar el software R se presentan en la tabla 8.

Modelo 1	Modelo 2	Modelo 3
a_1 = 1.1708*** a_2 = 0.4154 a_3 = -0.14	a_1 =0.76214** a_2 =0.02247 a_3= -0.53658***	
b_1 = -0.7128* b_2 =-0.4358 b_3 =-0.3577 b_4 =-0.6058 b_5 =-0.2477		b_1= -0.22314 b_2 = 0.02151 b_3 = 0.08701 b_4 = -0.13062 b_5 = 0.20294 b_6 =0.45474*
D_1 = 22.785 con 10 g.l.	D_2 = 29.369 con 15 g.l	D_3 = 59.312 con 12 g.l
	$D_2 - D_1$ = 6.584 con 5 g.l.	$D_3 - D_1$ = 36.527* con 2 g.l.

Tabla 8. Resultados de los tres modelos de la aplicación 2 de la regresión logística binaria

El modelo 1 es el que mejor ajusta los datos pero tiene demasiados parámetros, lo que significa que hay un sobreajuste (Principio de parsimonia). El modelo 2 tiene solo tres parámetros y no difiere significativamente del modelo 1. El modelo 3 es inapropiado porque su desvío es bastante alto. Los porcentajes de desaprobados varían de semestre a semestre con un nivel de significación del 5%. El modelo más apropiado es el 2. (Las proporciones de desaprobados son distintas de semestre a semestre.)

4.6 Regresión logística ordinal y nominal

Para cuantificar las posibilidades que tienen los alumnos de terminar sus estudios en los tiempos previstos, se decidió utilizar los datos de la promoción 2005-2. Los datos fueron los siguientes:

1. Variable dependiente: tiempo de demora en terminar la carrera, codificado de la siguiente manera (1 = 6 años, 2 = 7 años, 3 = 8 o más años).
2. Variable predictora: PPA (se probaron varias variables predictoras pero el promedio ponderado acumulativo PPA demostró ser el mejor).

A continuación se presentan los resultados de Minitab:

```
Ordinal Logistic Regression: Código-Tiempo versus PPA
Link Function: Logit
Response Information
Variable      Value Count
Código-Tiempo  1     15 (Egresados que terminaron en 6 años)
               2     5 (Egresados que terminaron en 7 años)
               3     15 (Egresados que terminaron en 8 o más años)
            Total    35

Logistic Regression Table
                          Odds    95% CI
Predictor     Coef     SE Coef      Z       P     Ratio   Lower  Upper
Const(1)    -22.8389   6.69754   -3.41    0.001
Const(2)    -21.7378   6.55946   -3.31    0.001
PPA          1.72206  0.511616    3.37    0.001    5.60    2.05   15.25
Log-Likelihood = -22.904
Test that all slopes are zero: G = 24.489, DF = 1, P-Value = 0.000
Goodness-of-Fit Tests
Method  Chi-Square   DF      P
Pearson    75.1856   67  0.230
Deviance   45.8075   67  0.978
```

Sobre la base de los coeficientes calculados se procedió a construir la tabla 9 con diferentes valores de PPA y con las probabilidades correspondientes de que el alumno termine sus estudios en 6, 7, 8 o más años.

PPA	Termine en 6 años	Termine en 7 años	Termine en 8 o más años
10	0.003617	0.007184	0.989199
11	0.019914	0.037674	0.942413
12	0.102086	0.152725	0.745188
13	0.388838	0.267934	0.343228
14	0.780718	0.133871	0.085411
15	0.952207	0.031378	0.016414
16	0.991111	0.005916	0.002973
17	0.9984	0.001068	0.000533
18	0.999714	0.000191	0.000095
19	0.999949	0.000034	0.000017

Tabla 9. Valores de PPA y probabilidades de terminar en 6, 7, 8 o más años.

La variable PPA es una variable apropiada para modelar el tiempo de demora en terminar la carrera. Los valores que aparecen como Const (1) y Const (2) son los estimados de los interceptos para los logit de las probabilidades acumuladas para el tiempo de demora de 6 años y 7 años, respectivamente. Debido a que la probabilidad acumulada para la última respuesta es 1, no se presenta el intercepto para 8 o más años. El coeficiente estimado para la covariable PPA es 1.72206 con p-value 0.001. Este p-value indica que hay evidencia significativa para sostener que el PPA afecta el tiempo de demora en terminar los estudios. El signo positivo del coeficiente con una razón de apuestas mayor que uno e igual a 5.6, indica que valores altos de PPA tienden a estar asociados con menores tiempos de demora en terminar los estudios. Específicamente, si el PPA de un alumno aumenta en una unidad, entonces la razón de apuestas de que termine en 6 años versus que termine en 7 años es 5.6 veces; de igual forma que termine en 7 años versus que termine en 8 o más años. El valor del desvío de 45.8075 con p-value de 0.978 indica que el modelo es apropiado. El valor de PPA próximo a 13 es un indicador para determinar si un alumno demora, entre 6 años como máximo y no menos de 7 en terminar su carrera.

4.7 Modelo de herramienta informática para el sistema de análisis estadístico

Con la finalidad de demostrar la viabilidad del sistema de análisis estadístico se construyó una aplicación piloto en Visual Basic que conectase Excel con R. La idea es capturar los datos de Excel y llevarlo a R para su ejecución. La elección de R en esta prueba piloto es porque es un software mucho más potente que Minitab y SPSS. El programa piloto se denomina Rconexion.

5. Conclusiones

- Las técnicas multivariadas son útiles para analizar el rendimiento académico de los alumnos.
- La técnica del análisis de covariancia permite comparar las diferentes secciones de un curso eliminando el efecto del promedio ponderado acumulativo o de cualquier otra variable que se utilice como covariable.
- La regresión logística binaria permite modelar el porcentaje de desaprobados de diferentes ciclos académicos. Luego de probar con

diversos modelos, se eliminaron aquellos que tenían un desvío alto o no cumplían con el principio de parsimonia.

- La regresión logística ordinal permite calcular las probabilidades de que un alumno termine sus estudios en 6, 7, 8 o más años de acuerdo con su promedio ponderado acumulativo.
- El análisis de correspondencia simple permite detectar la influencia de variables cualitativas en el rendimiento académico de los alumnos. Por ejemplo, los alumnos que pertenecen a la condición laboral A (alumnos que no trabajan actualmente), se encuentran más asociados a situarse en el décimo superior y en el tercio superior. Mientras que los alumnos de condición laboral B (con trabajo de medio tiempo), se encuentran más asociados a situarse en el medio superior.
- El análisis clúster es útil para identificar los cursos "cuellos de botella".
- El análisis discriminante permite identificar los cursos importantes dentro de cada grupo de alumnos y cada área académica.
- Debido a que la información disponible está constituida por notas y promedios ponderados, el análisis factorial no pudo ser aplicado porque reduciría innecesariamente el número de variables en estudio. Otro inconveniente para el análisis factorial fue que las variables cualitativas no eran significativas.

6. Bibliografía

Cea D´Ancona, A. *Análisis multivariable*. *Teoría y práctica en la investigación social.* Madrid: Editorial Síntesis, 2002.

Dallas E., Johnson. *Métodos multivariados aplicados al análisis de datos.* Internacional Thomson Editores, 2000.

Dobson, Annette. *An Introduction to Generalized Linear Models.* Florida: Chapman & Hall/CRC, 2002.

Hair, J. F.; Anderson, R. E.; Tatham, R. L. y W.C. Black, *Análisis Multivariante.* 5.ª edición. Traducción de Esme Prentice y Diego Cano. Madrid: Prentice Hall/Iberia, 1999.

Incose. "Definiton of a system" [en línea] <http://www.incose.org/practice/fellows consensus.aspx>. [Consulta: 24 de diciembre del 2006.]

Lévy Mangin, Jean-Pierre y Jesús Mallou Varela. *Análisis multivariable para las ciencias sociales.* Madrid: Pearson/Prentice Hall, 2003.

Luque Martínez, Teodoro. *Técnicas de análisis de datos en investigación de mercados.* Madrid: Ediciones Pirámide, 2000.

Montgomery, D. C. *Diseño y análisis de experimentos*. Traducción de Jaime Delgado Saldívar. México: Grupo Editorial Iberoamérica, 1991.

Neter, J. y W. Wasserman. *Applied Linear Statistical Models*. Illinois: Richard D. Irwin, Inc., 1974.

Peña, D. *Análisis de datos multivariantes*. Madrid: McGraw-Hill, 2002.

Uriel, E. y J. Aldás. *Análisis multivariante aplicado.* Madrid: Thompson, 2005.