*Economía*, XVI, 6 (1991), pp. 51-77

# Análisis discriminante: Estudio del rendimiento estudiantil

*Discriminant analysis: A study of students' performance*

**Elsy Garnica O., Pilar González M., Amelia Díaz de Pascual y Enrique Torres L.***

**Resumen**

Mediante un análisis discriminante, aplicado a un grupo de estudiantes de nuevo ingreso, se establecen algunas características que diferencian a los alumnos de rendimiento satisfactorio en Química 11 de aquellos cuyo rendimiento es bajo. Se midieron doce variables clasificadas en la siguiente forma: tres socioeconómicas, dos referentes a la preparación en bachillerato, tres actitudinales y, por último, la variable grupo que fue la nota en Química. Según las características analizadas se encontró que la frontera que separa al alto y bajo rendimiento es la nota de doce puntos (12 a 20 puntos para alto rendimiento). Tres fueron las variables que resultaron relevantes a la luz de los requisitos del análisis: habilidad numérica, promedio de bachillerato y la motivación hacia el sistema educativo. En base a estas tres variables se obtiene un porcentaje general de buena clasificación (84%) que resulta ser tan alto como el logrado utilizando todas las once variables predictoras iniciales. El porcentaje de buena clasificación, en el análisis definitivo, es similar en ambos grupos.

## 1. Introducción

Los estudiantes universitarios frecuentemente encuentran grandes dificultades en el estudio de la química básica; por ello, el rendimiento en esta asignatura rara vez alcanza niveles aceptablemente altos. Concretamente, en la Facultad de Ciencias de la Universidad de Los Andes (ULA), la materia del primer semestre, Química 11, común para los estudiantes de Biología y Química, presenta un alto porcentaje de aplazados.

* Universidad de Los Andes, Instituto de Investigaciones Económicas y Sociales

Para poder implementar políticas tendientes a mejorar el rendimiento estudiantil en Química, conviene conocer las causas de las dificultades presentadas por los alumnos.

En vías de desarrollar este conocimiento se deben buscar las variables que discriminan el grupo de estudiantes que presentan buen rendimiento de los que no lo tienen.

Los resultados del análisis discriminante pueden ayudar a orientar a los investigadores en esta área del conocimiento en posteriores estudios causales. El conocimiento de las variables discriminantes induce a determinar algunas de las dificultades concretas que encuentran los estudiantes.

## 2. Objetivo

El objetivo del presente trabajo fue buscar variables que discriminaran significativamente al grupo de alumnos que obtuvieron un rendimiento satisfactorio en Química 11 de los que no alcanzaron este nivel.

## 3. Metodología

La investigación sobre rendimiento se efectuó en un grupo de estudiantes de nuevo ingreso en las Licenciaturas en Química y Biología de la Facultad de Ciencias de la ULA, en el Semestre A-84. Se observaron 49 estudiantes, de los cuales 19 optaron por Química y 30, por Biología.

A continuación se especifican las variables y el análisis estadístico multivariante utilizado.

### *3.1. Variables*

Se midieron doce (12) variables que se agruparon en la siguiente forma:

*Variable Grupo*:
NOTAQ (Nota en Química). Es la calificación obtenida en Química 11. Se expresó en una escala de 0 a 20 puntos.

*Variables acerca de la preparación anterior*:
- PRBCH (Promedio en Bachillerato). Es la media aritmética de las notas obtenidas en el ciclo diversificado de bachillerato. Esta variable tiene un campo de variación entre 0 y 20 puntos.
- APRBH (Apreciación del Bachillerato). Se midió en función de las respuestas de los alumnos acerca de su preparación en bachillerato. La escala utilizada fue:
  1. Mala,
  2. Deficiente,
  3. Buena,
  4. Muy Buena y
  5. Excelente

*Variables Socioeconómicas*:
- INSTP (Instrucción del Padre) e INSTM (Instrucción de la Madre). Estas midieron el grado de escolaridad de los progenitores. Se utilizo la escala:
  1. Sin escolaridad,
  2. Educación Primaria,
  3. Educación Secundaria o técnica y
  4. Grado Universitario o equivalente.
- INGFM (ingreso familiar mensual por miembro). Medida como la razón entre el monto, en bolívares, de los recursos económicos familiares mensuales y el número de miembros que integran la familia.

*Variables Aptitudinales*:
- RAZAB (razonamiento abstracto), RAZVB (razonamiento verbal) y HABNU (habilidad numérica). Medidas a través de los respectivos tests de la batería DAT (Bennet – Seashore – Wesman), traducidos y adaptados por el Colegio Americano de Guatemala.

*Variables Actitudinales*:

– MOTES (motivación hacia el estudio) y MOTST (motivación hacia el sistema), mediadas mediante la técnica del diferencial semántico de Osgood (Osgood *et al.*). El valor final de cada individuo se obtuvo sumando sus puntajes en los ítems, el resultado se dividió por el máximo valor posible y luego, se llevó a porcentajes.
– DPERM (Deseo de permanencia). Variable dicotómica que toma el valor de cero (0) cuando el alumno deseaba cambiar de carrera y de uno (1) en casos de que el alumno quería permanecer en ella.

### *3.2. Análisis estadístico*

El análisis factorial discriminante (AFD), cuyo término (discriminación) fue introducido por R. A. Fisher, en 1936, en el primer tratamiento moderno de problemas separatorios, es una técnica multivariante orientada fundamentalmente a lograr dos objetivos básicos:

– Explorar y analizar las posibles diferencias que puedan existir entre g poblaciones excluyentes, previamente definidas por el investigador, en base a las diferencias que puedan presentar en las p variables medidas. Se trata de hallar funciones que dependan de esas p variables originales que separen los g grupos tanto como sea posible. Por ejemplo, se desea clasificar a las familias de una ciudad en tres (g = 3) niveles socioculturales: bajo, medio y alto, en base a cuatro (p = 4) variables: grado de instrucción del padre, grado de instrucción de la madre, número de libros en el hogar y número de suscripciones a publicaciones periódicas.
– A partir del criterio de discriminación obtenido se puede proceder a incluir un nuevo elemento en algunos de los grupos formados. Este es el caso de los individuos que no se les conoce a priori el grupo al cual pertenece, entonces el AFD permite clasificarlos sobre la base de ecuaciones matemáticas, derivadas del análisis de los casos con pertenencia conocida. En el ejemplo anterior, una vez conocidas las funciones discriminantes, se tiene la posibilidad de saber en qué grupo o nivel sociocultural se puede ubicar una familia que no fue seleccionada en el estudio inicial.

Las variables utilizadas en este análisis estadístico son denominadas *variables discriminantes*. Estas deben ser medidas en la escala de intervalo o razón para que las medias y varianzas puedan ser calculadas e interpretadas. Un requerimiento para la utilización del AFD es que el número de casos observados (n) debe exceder por más de dos, al número de variables.

Ninguna variable original puede ser combinación lineal de otras variables discriminantes ya que se tendría una redundancia en la información. Una combinación lineal es la suma de una o más variables que pueden haber sido ponderadas por términos constantes. Del mismo modo, dos variables que están perfectamente correlacionadas no pueden ser usadas al mismo tiempo.

Aunque existe un enfoque empírico en el que no se requiere una forma particular de la distribución de las poblaciones (función lineal *discrimínate* de Fisher), en este estudio se sigue el esquema tradicional del AFD que exige los siguientes supuestos:

a. Cada grupo debe ser considerado como una muestra extraída de una población normal multivalente. Cuando cada variable tiene una distribución normal con valores fijos para todas las otras, se puede considerar que cada grupo es extraído de una población normal multivariante. Con un gráfico apropiado (por ejemplo el P – P plot) puede probarse este supuesto.
b. Las matrices de covarianzas poblacionales de todos los grupos deben ser iguales (esto puede probarse con el test para la homogeneidad de varianzas de Box). Este supuesto permite la simplificación de la fórmula lineal discriminante y la adecuada interpretación de los resultados de los test de significación.

A veces estos supuestos son difíciles de hallar. Algunos autores, entre ellos. Lachennbruch han demostrado que el AFD es una técnica robusta que puede tolerar ciertas desviaciones de estos supuestos (1975).

Si se cumple el primer supuesto (normalidad), las funciones lineales discriminantes minimizan la probabilidad de mala clasificación[1]. Existen diversas reglas de clasificación, entre ellas: la regla de clasificación

de máxima verosimilitud, la regla de clasificación de Bayes, regla de clasificación generalizada de Bayes y la función discriminante cuadrática en poblaciones normales; además, se disponen de diversos métodos de evaluación de las funciones de clasificación: holdout, resubstitución técnica, *jackknife*, entre otros, (Márquez, 1989).

Cuando el supuesto de normalidad es violado, el cálculo de probabilidades no es exacto. En este caso, al tener probabilidades aproximadas, los resultados deben interpretarse con cuidado ya que puede haber una reducción en eficiencia y seguridad, sobre todo en el caso de muestras muy pequeñas.

En los casos frontera, los pequeños errores debido a la violación del supuesto de normalidad podrían causar una clasificación incorrecta. Por ejemplo, si un caso tiene una probabilidad de 0,95 de pertenecer al grupo 1 y 0,05 de pertenecer al grupo 2, no importa la imprecisión debida a la violación de los supuestos, ya que la decisión para asignar el caso al grupo 1 será posiblemente correcta; si el caso tiene una probabilidad de 0,51 de pertenecía del grupo 1 y de 0,49 para el grupo 2, se debe ser cauteloso al tomar una decisión.

Es frecuente conseguir que el requerimiento de igualdad de matrices de covarianzas de los grupos, no se cumpla. Cuando este supuesto es violado se presentan distorsiones en la función discriminante canónica y la ecuación de clasificación. Cuando se determina la existencia de diferencias significativas, aún es posible utilizar, con buenos resultados la función lineal discriminante si las matrices de covarianzas no son muy diferentes[2]. Si se ha comprobado que las matrices de covarianzas de grupos son muy diferentes, se sugiere usar estas matrices para calcular la probabilidad de pertenencia de grupo. Este análisis se denomina *discriminación cuadrática*.

En el AFD se obtiene una combinación de las variables independientes, para cada caso o individuo, denominado *score* (D):

$$D = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p ,$$

que resume la información contenida en todas las variables en un sólo índice. Además, esta combinación lineal se utiliza para asignar nuevos casos o individuos a los grupos ya formados.

Si los puntajes o scores de las funciones discriminantes están distribuidos normalmente para cada uno de los grupos observados y los parámetros de la distribución pueden ser estimados, entonces es posible calcular la *probabilidad condicional* de obtener un score D dado que el individuo en cuestión sea miembro del grupo $G_i$ (para i = 1,.,g), es decir $P(D/G_i)$. También es posible calcular, por otra parte, la probabilidad a priori $P(G_i)$, la cual es un estimador máximo verosímil de la probabilidad de que un individuo j pertenezca a un grupo particular $G_i$, cuando no hay información disponible. Con el conocimiento de las probabilidades condicionales y a priori es posible calificar individuos en uno de los grupos, utilizando la Regla de Bayes:

$$P(G_i \mid D) = \frac{P(D \mid G_i)\, P(G_i)}{\sum_{i=1}^{g} P(D \mid G_i)\, P(G_i)}$$

Donde:

$P(G_i \mid D)$ : Probabilidad de que un individuo pertenezca al grupo i dado que tiene un score D.

$P(D \mid G_i)$ : Probabilidad de que un individuo tenga un score D, dado que pertenece al grupo i.

$P(G_i)$ : Probabilidad de que un individuo pertenezca al grupo i.

El porcentaje de correcta clasificación es un indicador de gran utilidad cuando la investigación se centra en la búsqueda de una descripción razonable del mundo real. Si este porcentaje es alto, no importa la violación de los supuestos. Pero, si el porcentaje de correcta clasificación es bajo, no se puede asegurar si esto se debe a la violación de los supuestos o al uso de variables discriminantes débiles.

El *centroide* de un grupo es un punto cuyas coordenadas son los promedios, en el grupo, de cada una de las variables discriminantes. En

la interpretación de la función discriminante canónica es de gran utilidad analizar la posición relativa de los datos respecto a los centroides.

La correlación entre una variable y la función discriminante, denominada *coeficiente de estructura total*, es el coseno del ángulo formado por el punto variable y la función. Esta correlación se utiliza para determinar el peso que tiene esa variable en la función discriminante.

Cada una de las funciones discriminantes está asociada a un *valor propio* ($\lambda$). Los valores propios no tienen interpretación directa pero la magnitud de éste indica el poder discriminatorio de la función. Ordenando los valores propios en forma descendente se obtienen las funciones discriminantes organizadas según su poder discriminatorio. En otras palabras, la función con mayor valor propio es el discriminador más poderoso, mientras que la función con el menor valor propio es la más débil. Para comparar los valores propios se acostumbra transformarlos en proporciones relacionados con el suma total de ellos (varianza total). Este valor indica si la función es fuerte o débil relacionada con las demás.

El número de funciones discriminantes del análisis es el mínimo entre p (número de variables) y g (número de grupos menos uno).

Para probar la significancia estadística de las funciones discriminantes se examinan los residuales de la discriminación "a priori". Si la discriminación residual es pequeña, no tiene sentido utilizar las restantes funciones aún si ellas existen matemáticamente. La lambda de Wilks ($\Lambda$) es una medida de la diferencia entre los grupos, respecto a las variables discriminadas.

$$\Lambda = \frac{\text{Suma de cuadrados dentro de grupos}}{\text{Suma de cuadrados total}}$$

Este estadístico toma valores entre cero y uno y su interpretación está en relación inversa a la de la F:

$$F = \frac{\text{Cuadrado medio entre grupos}}{\text{Cuadrado medio dentro de grupos}}$$

Los valores de lambda de Wilks cerca de cero denotan alta discriminación. Esto quiere decir que los centroides están muy separados entre sí. A medida que el lambda de Wilks se acerca a uno el poder discriminatorio de la función se hace más débil. Cuando tiene exactamente el valor de uno, los centroides de los grupos son idénticos (no hay diferencias entre los grupos ).

Como la distribución del lambda de Wilks puede ser aproximada a una distribución $X^2$ (j i cuadrada), el nivel de significación se puede determinar comparando el valor calculado u observado con el valor tabulado de esta distribución.

El AFD no es un análisis causal porque las variables no se definen como dependientes o independientes. Si en determinado estudio se define la variable grupo como independiente de las variables discriminantes se tiene una situación análoga al análisis de regresión múltiple.

## 4. Análisis de los resultados

Se estudiaron varias alternativas de análisis que se anuncian en esta sección. Se indican, en cada una de ellas, las dificultades encontradas con respecto al porcentaje de clasificación errada. Finalmente se explica, en forma detallada, los resultados del AFD definitivo.

### *4.1. Primer análisis: dos grupos y 11 variables predictoras*

Teniendo en cuenta que el sistema de evaluación de la ULA, establece como nota mínima aprobatoria 10 puntos, el primer análisis de este trabajo consistió en definir *dos grupos* de estudiantes: aplazados con notas en Química 11, bajo 10 puntos y *aprobados* con notas de 10 o más. Es decir, la variable grupo quedó codificada como:

– Grupo 1 (aplazados): notas entre 0 y 9 y
– Grupo 2 (aprobados): notas entre 10 y 20

En este primer intento, se encontraron altas correlaciones entre algunas de las 11 variables del análisis. Además, estas variables no resultaron significativas en lo que se refiere a la prueba de análisis de varianza entre grupos (estadístico F). Por último, el coeficiente de correlación canónica fue bajo (49%).

Una vez efectuado el análisis se observó que el 29% de los estudiantes del Grupo 2 (aprobados) estaban mejor ubicados en el Grupo 1 (aplazados). Por otra parte, un 27% de los estudiantes del Grupo 1 (aplazados) estaban más adecuadamente clasificados en el Grupo 2 (aprobados) (véase el Cuadro 1).

El resultado anterior indujo a pensar que la frontera de clasificación que representaban los 10 puntos (como nota mínima aprobatoria) no es totalmente adecuada. Para salvar la dificultad que se produce en la separación rígida de aprobados y aplazados se trabajó sobre la idea de formar una agrupación distinta que tomara en cuenta aprobados con altas notas, aprobados con bajas notas, aplazados con altas notas y aplazados con bajas notas.

**Cuadro 1**. Resultados de clasificación para el primer análisis

| Grupo Actual | Número de casos | Predicción de pertenencia de grupo | |
|---|---|---|---|
| | | 1 | 2 |
| GRUPO 1 | 15 | 11<br>73% | 4<br>27% |
| GRUPO 2 | 34 | 10<br>29% | 24<br>71% |
| Porcentaje de "agrupación" de casos correctamente clasificados: 71% | | | |

### *4.2. Segundo análisis: cuatro grupos y 11 variables predictoras*

En este análisis se definieron cuatro grupos de estudiantes. Grupo 1 (rendimiento bajo) notas menores o iguales que 6. Grupo 2 (rendimiento medio bajo): notas entre 7 y 10. Grupo 3 (rendimiento medio alto): con notas entre 11 y 12. Grupo 4 (rendimiento alto): notas iguales o mayores que 13.

En este segundo análisis se mantuvieron altas correlaciones entre algunas de las 11 variables predictoras.

Por otra parte, las variables *Razab, Razvb, Habnu,Prbch, Aprbh* y *Motst* resultan significativas en la prueba de análisis de varianza entre los cuatro grupos escogidos. Este último es indicador de que, en al menos dos de los grupos, se produjeron diferencias significativas respecto a las variables mencionadas.

Los coeficientes de correlación canónica[3] de cada una de las tres funciones discriminantes fueron del 75, 54 y 45% respectivamente. Cada una de ellas con el 66, 21 y 13% de varianza explicada.

**Cuadro 2**. Resultados de clasificación del segundo análisis

| Grupo Actual | Número de casos | Predicción de pertenencia de grupo | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| GRUPO 1 | 8 | 6<br>75% | 2<br>25% | 0<br>0% | 0<br>0% |
| GRUPO 2 | 14 | 4<br>29% | 7<br>50% | 2<br>14% | 1<br>7% |
| GRUPO 3 | 15 | 1<br>7% | 4<br>26% | 9<br>60% | 1<br>7% |
| GRUPO 4 | 12 | 0<br>0% | 0<br>0% | 2<br>17% | 10<br>83% |
| Porcentaje de "agrupación" de casos correctamente clasificados: 65% | | | | | |

Los casos erróneamente clasificados alcanzaron un porcentaje total del 35% mucho mayor que el obtenido en el primer análisis (los casos correctamente clasificados sólo alcanzaron el 65%). Debe destacarse que los dos grupos intermedios (grupo 2 y grupo 3) presentaron bajos porcentajes de buena clasificación (50 y 60%, respectivamente) y que los dos grupos extremos (grupo 1 y grupo 4) tuvieron un porcentaje aceptable de buena clasificación (75 y 83%, respectivamente, véase el Cuadro 2).

Como la clasificación no mejoró el análisis preliminar, se estudia una nueva alternativa en la cual se definieron los grupos de bajos, regular y alto rendimiento.

### *4.3. Tercer análisis: tres grupos y 11 variables predictoras*

En este tercer intento se definieron tres grupos. *Grupo 1 (bajo rendimiento) notas del o al 8 Grupo 2 (regular rendimiento)*: notas de 9 al 11 *Grupo 3 (alto rendimiento)*: notas mayores o iguales a 12. Esta agrupación se realizó con el fin de recoger, en una grupo intermedio, los casos no identificados con los grupos extremos.

En este análisis se mantuvieron altas correlaciones entre alguna de las 11 variables analizados y se observaron algunos variables significativas en el análisis de varianza entre las tres grupos (*Razab, Razvb, Habnu, Prbch, Aprbh*, e *Instm*). Por otra parte, las correlaciones canónicas entre los variables grupo de las dos funciones lineales discriminantes fueron, respectivamente, 73 y 43%. La primera de estas funciones explicaba un 84% de la varianza y la segunda, el 16%.

Aunque el porcentaje total de correcta clasificaciones aumentos (de 65 a 67%) se siguió observando porcentajes bajos en los grupos 1 y 2 (67 y 54% respectivamente) mientras que en el grupo 3 aumentó notablemente (92%) (véase el Cuadro 3). Este resultado condujo, en forma natural, a una cuarta alternativa de estudio en la cual el grupo 3 (alto rendimiento -notas mayores o iguales a 12 puntos) se aisló por sí sólo y los dos grupos restantes se fusionaron para integrar un sólo conjunto. Así, se procedió a tomar la nota de 12 punto como el límite entre los alumnos de bajo y alto rendimiento.

### *4.4. Cuarto análisis: dos grupos y 11 variables predictoras.*

En este cuarto análisis se formaron sólo dos grupos separados por la nota frontera 12 puntos. La nueva estructura de los individuos observados quedó conformada en *Grupo 1 (bajo rendimiento)*: notas entre 0 y 11 puntos, y *Grupo 2 (alto rendimiento)*: notas iguales o superiores a 12 puntos.

**Cuadro 3**. Resultados de clasificación para el tercer análisis

| Grupo Actual | Número de casos | Predicción de pertenencia de grupo | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| GRUPO 1 | 15 | 10<br>67% | 3<br>20% | 2<br>13% |
| GRUPO 2 | 22 | 7<br>32% | 12<br>54% | 3<br>14% |
| GRUPO 3 | 12 | 1<br>8% | 0<br>0% | 11<br>92% |
| Porcentaje de "agrupación" de casos correctamente clasificados: 67% | | | | |

El porcentaje de clasificación correcta total fue de un 90% (véase el Cuadro 4). Pero aunque este indicador aumentó significativamente con respecto a los análisis anteriores, fue preciso observar los resultados de las pruebas estadísticas univariantes (F) y las correlaciones simples entre cada par de variables discriminantes.

Con finalidad de elegir las variables más importantes del estudio, se analizaron a través del estadístico F, las diferencias significativas de cada una de las variables originales, entre los dos grupos propuestos de estudiantes (véase el Cuadro 5).

**Cuadro 4**. Resultados de clasificación para el cuarto análisis

| Grupo Actual | Número de casos | Predicción de pertenencia de grupo | |
|---|---|---|---|
| | | 1 | 2 |
| GRUPO 1 | 31 | 28<br>90% | 3<br>10% |
| GRUPO 2 | 18 | 2<br>11% | 16<br>89% |
| Porcentaje de "agrupación" de casos correctamente clasificados: 90% | | | |

**Cuadro 5**. Lambda de Wilks y Razon f Univariante con 1 y 47 grados de libertad

| Variable | Lambda de Wilks | F | Significación |
|---|---|---|---|
| RAZAB | 0,77251 | 13,8400 | 0,0005 |
| RAZVB | 0,91015 | 4,6400 | 0,0364 |
| HABNU | 0,68650 | 21,4600 | 0,0000 |
| PRBCH | 0,89950 | 5,2510 | 0,0265 |
| APRBH | 0,97067 | 1,9240 | 0,1719 |
| MOTES | 0,98795 | 0,5734 | 0,4527 |
| MOTST | 0,90986 | 4,6560 | 0,0361 |
| DPERM | 0,99214 | 0,3724 | 0,5447 |
| INSTP | 0,99301 | 0,3310 | 0,5678 |
| INSTM | 0,99880 | 0,0157 | 0,8131 |
| INGFM | 0,99370 | 0,2978 | 0,5878 |

Las variables que causaron diferencias significativas, el nivel del 5% (véase el Cuadro 1) son: *Razab* (razonamiento abstracto), *Razvb* (razonamiento verbal), *Habnu* (habilidad numérica), *Prbch* (promedio de bachillerato), *Motst* (motivación hacia el sistema).

Al estudiar las correlaciones simples combinadas[4] de estas cinco variables particulares (véase Cuadro 6), se notó la existencia de relaciones significativas ($\alpha$ = 0,05) entre *Razab* y *Razvb* (0,46); *Razab* y *Habnu* (0,58); *Razvb* y *Habnu* (0,36); y *Habnu* y *Motst* (0,39).

Aunque *Habnu* y *Motst* presentaron una correlación al borde de la significación se tomó la decisión de dejarlas en el análisis. La justificación que se hace a este punto es que ambas características, motivación hacia el sistema (*Motst*) y habilidad numérica (*Habnu*), miden aspectos diferentes del individuo. La primera de ellas es una variable actitudinal y la segunda aptitud. De las cinco variables significativas solamente se descartaron *Razab* y *Razvb*.

**Cuadro 6**. Matriz de correlación combinada entre grupos

| | RAZAB | RAZVB | HABNU | PRBCH | APRBH | MOTES | MOTST | DEPRM | INSTP | INSTM | INGFM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| RAZAB | 1,000 | | | | | | | | | | |
| RAZVB | 0,460 | 1,000 | | | | | | | | | |
| HABNU | 0,584 | 0,357 | 1,000 | | | | | | | | |
| PRBCH | 0,069 | 0,256 | 0,002 | 1,000 | | | | | | | |
| APRBH | 0,294 | 0,327 | 0,178 | 0,510 | 1,000 | | | | | | |
| MOTES | 0,222 | 0,117 | 0,328 | 0,078 | 0,146 | 1,000 | | | | | |
| MOTST | 0,314 | 0,237 | 0,388 | -0,002 | 0,194 | 0,706 | 1,000 | | | | |
| DPERM | -0,024 | 0,172 | 0,026 | 0,106 | -0,027 | 0,366 | 0,203 | 1,000 | | | |
| INSTP | 0,269 | 0,192 | 0,371 | -0,029 | 0,109 | 0,210 | 0,169 | -0,164 | 1,000 | | |
| INSTM | 0,216 | 0,027 | 0,384 | -0,025 | 0,031 | 0,066 | 0,030 | -0,186 | 0,664 | 1,000 | |
| INGFM | -0,001 | 0,105 | -0,036 | 0,060 | 0,214 | 0,226 | 0,149 | 0,054 | 0,383 | 0,173 | 1,000 |

Con el nuevo conjunto formado por las tres variables discriminantes, *Habnu, Prbch* y *Motst*, se procedió a realizar el análisis estadístico discriminante definitivo. Es conveniente trabajar con menos cantidad de variables observadas (el análisis se redujo de 11 variables a 3) si se producen resultados satisfactorios.

### *4.5. Quinto análisis: dos grupos y tres variables predictoras.*

En este quinto y último análisis se formaron dos grupos tomando, al igual que en el análisis previo, la nota de 12 puntos como frontera entre los estudiantes de bajo y alto rendimiento. La agrupación definitiva es la siguiente: *Grupo 1 (bajo rendimiento)*: notas entre 0 y 11 puntos, y *Grupo 2 (alto rendimiento)*: notas entre y 12 y 20 puntos.

En el análisis discriminante definitivo se tomaron en cuenta sólo tres variables: *Habnu, Prbch* y *Motst*. Los resultados obtenidos se describen a continuación.

*4.5.1. Promedios para cada grupo*

En el cuadro 7 puede observarse que el grupo 2 (alto rendimiento) presentó mayores medidas en dos de las variables: *Habnu* y *Prbch*. Este es un resultado esperado por la relación existente entre el rendimiento estudiantil, la habilidad numérica y el promedio de bachillerato.

**Cuadro 7**. Medidas de grupo

| Nota Q | HABNU | PRBCH | MOTST |
|---|---|---|---|
| 1 | 4,23 | 12,45 | 73,94 |
| 2 | 12,24 | 13,42 | 64,39 |
| Total | 7,17 | 12,81 | 70,43 |

La poca diferencia observada en *Prbch* (los promedios de bachillerato) entre los dos grupos estudiados fue debido a que esta característica presentó poca variabilidad, púes sólo se promediaron las notas aprobatorias. El rango muestral que presento esta variable fue de 11,1 a 16,5 puntos.

En cuanto a variable *Motst*, debe recordarse que este tiene un rango de 0 a 100 y los valores medios observados en ambos grupos fueron relativamente altos (74y 64 para los grupos 1 y 2, respectivamente). El test para medir la motivación hacia el sistema fue aplicado a estudiantes de nuevo ingreso, al inicio del semestre. Como ellos, supuestamente, no habían experimentado las dificultades propias del aprendizaje, presentaban altas expectativas hacia el sistema de educación superior.

Analizando el cuadro 7, pareciera que los promedios observados de la variable *Motst* son contradictorios, en el sentido de que los alumnos con alto rendimiento presentaron un promedio observado más bajo que el de los alumnos con bajo rendimiento. Una posible explicación de este comportamiento puede ser el hecho de que los estudiantes con alto rendimiento sean más exigentes para con ellos mismo y para con el sistema, produciendo este resultado aparentemente inesperado.

### *4.5.2. Matriz de correlaciones*

Cuando existen altas correlaciones entre las P variables discriminantes, el investigador debe cuidarse de interpretaciones erróneas de los coeficientes de las funciones discriminantes, porque las variables relacionadas están compartiendo el peso en la función.

Por lo general, cuando hay correlaciones altamente significativas los signos de los coeficientes correspondientes a esas variables son inversos (Norusis, 1986, p. B-16).

**Cuadro 8**. Matriz de correlaciones

| | HABNU | PRBCH | MOTST |
|---|---|---|---|
| HABNU | 1,00 | | |
| PRBCH | 0,00 | 1,00 | |
| MOTST | 0,39 | 0,00 | 1,00 |

En la presente investigación se han eliminado variables altamente correlacionadas ya que la presencia de éstas puede generar ciertas limitaciones en el análisis. Se decidió dejar tanto la variable *Motst* como la variable *Habnu*, que tiene una correlación al borde de la significación, porque ambas características son relevantes.

### *4.5.3. Lamda de Wilks y Razón F Univariante*

Mientras menor sea el lambda de Wilks, mayor es el valor correspondiente de F y más altas son las posibilidades de que las medidas de los grupos sean significativamente diferentes. En en cuadro 9 puede observarse que la variable *Habnu* posee el menor valor de lambda de Wilks y, por supuesto, el mayor valor F demostrando con ello ser la característica que produce las diferencias más significativas entre los dos grupos bajo estudio.

**Cuadro 9**. Lambda de Wilks y Razon F

| Variable | Lambda de Wilks | F | Significación |
|---|---|---|---|
| HABNU | 0,69 | 21,46 | 0,0000 |
| PRBCH | 0,90 | 5,25 | 0,0265 |
| MOTST | 0,91 | 4,66 | 0,0361 |

### *4.5.4. Matrices de covarianzas de grupos*

En el cuadro 10 pueden observarse grandes diferencias entre las covarianzas de ambos grupos. Por ejemplo, la covarianza entre la variable *Habnu* y la variable *Motst* en el primer grupo es de 1,79, mientras que en el segundo, es de 930,55.

**Cuadro 10**. Matriz de covarianzas por grupo

| | Grupo 1 bajo rendimiento | | | Grupo 2 alto rendimiento | | |
|---|---|---|---|---|---|---|
| | HABNU | PRBCH | MOTST | HABNU | PRBCH | MOTST |
| HABNU | 720,08 | | | 8140,37 | | |
| PRBCH | 3,25 | 1,45 | | -5,34 | 3,06 | |
| MOTST | 1,79 | -2,69 | 180,13 | 930,55 | 4,61 | 298,37 |

Al hacer la décima de hipótesis acerca de la igualdad de matrices de covarianzas utilizando el estadístico M de Box (Norusis, 1986, p. B-33), se concluye que las matrices de covarianzas de ambos grupos son significativamente diferentes (Cuadro 11). Cabe destacar que el tamaño de la muestra en este estudio es pequeño, hecho que puede incidir en el resultado obtenido.

**Cuadro 11**. Resultados de la prueba de igualdad de matrices de covarianzas

| M de Box | F Aroximada | Significación |
|---|---|---|
| 38,227 | 5,86 | 0,0000 |

La función discriminante lineal no puede ser utilizada con fines productivos ya que no se cumple el supuesto de igualdad de matrices covarianzas. La función sólo puede ser utilizada para describir las diferencias entre los dos grupos estudiados.

### *4.5.5. Determinación del número de funciones discriminantes.*

El número de funciones discriminantes debe ser el mínimo entre p y g – 1, En esta investigación, donde se definen dos grupos (g = 2) y existen cuatro variables predictoras (p = 4), se obtiene una sola función discriminante.

Al analizar los coeficientes estandarizados de la función discriminante expuestos en el cuadro 12, se observó que la variable de mayor peso fue *Habnu* con un coeficiente de 0,95838.5

Como hay una relación directa (determinada por el signo positivo del coeficiente) entre la función discriminante y la variable *Habnu*, puede establecerse que cuanto mayor es la habilidad numérica, más alto debe ser el dato D obtenido, caso contrario sucede con la variable *Motst* cuyo peso es –0,69240 (el segundo en importancia); como está relación es inversa se supone que a mayor valor en el test de *Motst*, menor dato D. La justificación de estos resultados se explicó en páginas anteriores cuando analizaron los promedios de las variables discriminantes, por grupos.

**Cuadro 12**. Coeficiente de la función canónica discriminante estandarizada

| | Función 1 |
|---|---|
| HABNU | 0,95838 |
| PRBCH | 0,33824 |
| MOTST | -0,69240 |

A este punto no se debe obviar el conocimiento de correlaciones existentes entre las variables *Motst* y *Habnu* (Cuadro 8), como la contribución de estas dos variables es compartida, se debe tener sumo

cuidado al interpretar los coeficientes de las función discriminante canónica, puesto que los signos y magnitudes de los coeficientes de la función pueden verse seriamente afectados.

### *4.5.6. Correlación combinada dentro de grupos entre las variables discriminantes y la función canónica*

Además de los coeficientes estandarizados señalados en la sección anterior, otras características que se pueden utilizar para evaluar la contribución de cada variable a la función discrimínate, son las correlaciones entre los valores de las variables discriminantes y los valores de la función. Estas correlaciones se especifican en el cuadro 13.

**Cuadro 13**. Correlaciones (combinadas dentro de grupos) entre las variables discriminantes y la función discriminante canónica

| | Función 1 |
|---|---|
| HABNU | 0,69052 |
| PRBCH | 0,34155 |
| MOTST | -0,32162 |

La correlación más alta se observó en la variable *Habnu* y le siguieron, en orden decreciente, las variables *Prbch* y *Motst*.

Las variables *Habnu*, y *Prbch*, presentaron una correlación positiva con la función, quiere decir que los valores altos se asignaron, en promedio, a los estudiantes con alto rendimiento. Por otro lado la variable *Motst* presentó una correlación negativa, indicando así que, en general, los valores altos de esta variable fueron atribuidos a los alumnos de bajo rendimiento.

### *4.5.7. Correlación Canónica*

El análisis produjo una correlación del 70% aproximadamente, considerándose significativa (Cuadro 14), con este resultado puede aseverarse que la función discriminante cumplió con el objetivo de separar eficientemente, los elementos observados en los dos grupos creados.

El Lambda Multivariante de Wilks, arroja un valor de 0,5108, el cual está relativamente alejado de 1. Una prueba relacionada a la lambda de Wilks es el test de la $X^2$:

$$X^2 = - \left[ n - \frac{p + g}{2} - 1 \right] \log_e k$$

con (p – k) (g - k – 1) grados de libertad.
Donde,
n : es el número total de casos observados (n = 49),
p : es el número de variables predictoras (p = 3)
g : es el número de grupos propuestos (g = 2) y
k : es igual a cero antes de la derivación de cualquier función discriminante (Klecka, 1980, p. 40).

**Cuadro 14**. Características de la función canónica discriminante

| Función | Valor propio | Porcentaje de varianza | Correlación canónica | Después de la función (k) | Lambda de Wilks | X2 | g.l | Signif. |
|---|---|---|---|---|---|---|---|---|
| 1 | 0,9577 | 100 | 0,6994 | 0 | 0,5108 | 30,566 | 3 | 0,0000 |

En este estudio específico, el valor de $X^2$ calculado resultó significativo, lo que implicaba que debía procederse a calcular la primera y única función de este análisis.

**Cuadro 15**. Centroides de grupos

| Grupo | Función 1 |
|---|---|
| 1 | -0,73035 |
| 2 | 1,25782 |

El resultado de la prueba Lambda de Wilks, o de su homólogo $X^2$, indicó la existencia de diferencia significativa entre los dos centroides de grupos (véase el Cuadro 15).

*4.5.8. Aplicación de la función canónica discriminante.*

Una vez conocidos los coeficientes de la función discriminante se procedió a calcular los valores de ésta para algunos individuos observados. La función discrimínate obtenida en el presente análisis fue la siguiente:

$$D = 0{,}96\ HABNU + 0{,}34PRBCH - 0{,}69\ MOTST$$

Como ejemplo ilustrativo, se procedió a evaluar la función anterior para algunos estudiantes observados cuyos datos se ofrecen en el cuadro 16.

**Cuadro 16**. Valores observados de las variables analizadas en algunos individuos del estudio

| Individuo | Grupo | HABNU | PRBCH | MOTST |
|---|---|---|---|---|
| 10 | 1 | 5 | 11,6 | 88 |
| 24 | 2* | 43 | 12,8 | 92 |
| 34 | 1* | 73 | 14,3 | 58 |
| 37 | 1* | 58 | 13,0 | 53 |
| 41 | 2 | 175 | 12,3 | 60 |
| Media General | | 71,71 | 12,81 | 70,43 |
| Desviación Estándar | | 69,68 | 1,49 | 15,49 |

* Individuo que originalmente pertenece a ese grupo pero que, según la función discrimínate, debería estar ubicado en otro

Los valores estandarizados de los individuos antes señalados, respecto a la media general se especifican en el cuadro 17.

**Cuadro 17**. Valores observados estandarizados de las variables analizadas en algunos individuos del estudio

| Individuo | Grupo | HABNU | PRBCH | MOTST | D |
|---|---|---|---|---|---|
| 10 | 1 | -0,96 | -0,81 | 1,13 | -1,98 |
| 24 | 2* | -0,41 | -0,01 | 1,39 | -1,36 |
| 34 | 1* | 0,02 | 1,00 | -0,80 | 0,91 |
| 37 | 1* | -0,20 | 0,13 | -1,13 | 0,63 |
| 41 | 2 | 1,48 | -0,34 | -0,67 | 1,77 |
| Media del Grupo 1 | | -0,42 | -0,24 | 0,23 | -0,64 |
| Media del grupo 2 | | 0,73 | 0,41 | -0,39 | 1,11 |

*Individuo que originalmente pertenece a ese grupo pero que, según la función discrimínate, debería estar ubicado en otro

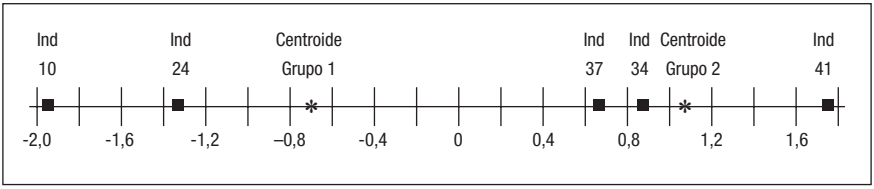


Figura 1. Centroides

### *4.5.9. Resultado de la clasificación*

Con este análisis definitivo se logró un alto porcentaje de buena clasificación (83,6%), que difería poco del obtenido en el cuarto análisis (en el cual se analizaron 11 variables predictoras y dos grupos con notas frontera de 12 puntos), la ventaja de utilizar esta última función discriminante en el lugar de la del cuarto análisis radicó en el menor número de variable predictoras utilizadas.

El porcentaje de buena clasificación para ambos grupos fue aproximadamente igual (83,9 y 83,3% respectivamente), esto permitió establecer que la probabilidad de mala clasificación fuera similar en los dos grupos observados.

La hipótesis nula planteada en la prueba M de Box, es la igualdad de matrices de covarianzas. Tal hipótesis fue rechazada a la luz de los resultados de este análisis. Por consiguiente se debe tener sumo cuidado al utilizar la función discrimínante para clasificar nuevos elementos ya que no se puede esperar que la probabilidad de buena clasificación sea necesaria del 83,6%. Está puede ser mucho menor.

**Cuadro 18**. Resultados de clasificación

| Grupo Actual | Número de casos | Perdición de pertenencia de grupo | |
|---|---|---|---|
| | | 1 | 2 |
| Grupo 1 | 31 | 26<br>83,9% | 5<br>16,1% |
| Grupo 2 | 18 | 3<br>16.7% | 15<br>83,3% |

Porcentaje de “agrupación” de casos correctamente clasificados: 83,6%

## 5. Conclusiones

Para la realización de este estudio originalmente se tomaron once variables, separadas en cuatro grupos:

– Variables aptitudinales
– Variables Actitudinales
– Variables socioeconómicas
– Variables de preparación de bachillerato

Inicialmente los dos grupos, de alto y bajo rendimiento en Química, estaban separados por la nota de diez puntos (nota mínima aprobatoria según el sistema de evaluación vigente en la ULA), el análisis realizado bajo estas condiciones dejaba mucho que desear, por lo tanto se procedió a efectuarlo tomando tres y hasta cuatro grupos con lo cual no se logró mejoramiento en los resultados de clasificación. Finalmente, se volvió a trabajar con la idea inicial de formar sólo dos grupos de estudiantes tomando diferentes notas como punto de separación. Se llegó a la conclusión de que la nota de doce puntos es el límite optimo de división (Grupo 1: de 0 a 11 puntos, Grupo 2: de 12 a 20 puntos).

Las variables que resultaron significativas, de las once variables predictoras originales en cuanto a la separación de los grupos de estudiantes de bajo y alto rendimiento en Química fueron: razonamiento abstracto, razonamiento verbal, habilidad numérica, promedio en bachillerato y motivación hacia el sistema educativo. Las tres primeras son variables aptitudinales, la cuarta mide el grado de preparación en bachillerato y la última es de tipo actitudinal. Debe destacarse que ninguna de las variables socioeconómicas resulto con capacidad discriminatoria.

Con el objeto de simplificar el análisis, sin pérdida de formalidad metodológica, se decidió dejar sólo las variables habilidad numérica, promedio de bachillerato y motivación hacia el sistema educativo para discriminar los dos grupos de estudiantes (bajo y alto rendimiento). Estas variables miden aspectos muy diferenciados del individuo.

Los criterios utilizados para la selección de estas tres variables fueron, fundamentalmente los siguientes: (a) la alta correlación canónica

observada, (b) los niveles de significación de las F y, fundamentalmente, (c) el porcentaje de buena clasificación obtenido. A pesar de que la habilidad numérica y la motivación hacia el sistema tiene una correlación significativa (no muy alta), se decidió dejar ambas variables en el análisis estadístico debido a que ellas miden aspectos diferentes del estudiante.

En el análisis definitivo se logró un porcentaje total de buena clasificación de 83,3%, el cual es consistente en los dos grupos investigados: 83,9% para el grupo 1 (ajo rendimiento) y 83,3% para el grupo 2 (alto rendimiento). Es de hacer notar que utilizado las once (11) variables predictoras se obtiene un 90% de buena clasificación, lo cual no representa una mejora sustancial que justifique el uso de tantas variables por las siguientes razones: 1) algunas no son significativas (prueba F), 2) las altas correlaciones entre algunas de ellas induce a una sobreestimación del porcentaje de buena clasificación.

La correlación canónica observada fue aproximadamente, de 0,70. Este valor representa una relación altamente significativa entre los scores (D) con los valores de la variable grupo.

A pesar de que las variables no tienen una distribución conjunta normal multivariante se puede comprobar que los puntajes de la función discriminante se distribuyen aproximadamente normal, esto quiere decir que las conclusiones sacadas a partir de las pruebas de significación F se pueden considerar confiables. Respecto al test M de Box se concluyó que las matrices de covarianzas son significativamente diferentes. Esto podría limitar los resultados del análisis[7].

## 6. Recomendaciones

En vista de que la nota de diez puntos no resultó ser un límite adecuado, en la separación entre los alumnos de alto y bajo rendimiento, se recomienda efectuar una revisión del sistema de evaluación a fin de medir, con fiabilidad, el rendimiento estudiantil.

Se comprobó que las tres variables que discriminaban entre el bajo y alto rendimiento fueron: habilidad numérica, promedio de bachillerato y motivación hacia el sistema educativo. La primera de ellas

resultó ser la más importante; esta característica podría ser mejorada en el alumno si éste es sometido a un entrenamiento previo para desarrollar la habilidad numérica. Por lo tanto se recomienda la incorporación de este tipo de adiestramiento en los cursos de nivelación universitaria para los estudiantes de las carreras relacionadas a Ciencias.

Considerando que el análisis discriminante es una técnica meramente exploratoria que diferencia grupos respecto a determinadas variables predictoras, se sugiere realizar un análisis causal que permita descubrir otros aspectos que influyen en el aprendizaje y en base a esto, implementar políticas dirigidas al mejoramiento del mismo.

Por último, se recomienda hacer extensivo este análisis a otras áreas del conocimiento universitario con fines comparativos.

## 7. Notas

1 Aunque existen pruebas para determinar la normalidad multívariante, existen tácticas sencillas que el investigador puede utilizar para examinar la normalidad Multivariante tal como el diagrama de puntos Q-Q (Johnson y Wicherm, p 152). Se advierte, sin embargo, que si cada una de las variables está distribuida normalmente no quiere decir que la distribución conjunta sea, necesariamente, una distribución normal Multivariante.

2 Si no se cumple este segundo supuesto, pero la distribución conjunta de las variables discriminantes es normal Multivariante, entonces la regla de clasificación óptima es la función de discriminación cuadrática.

3 La correlación canónica es la correlación r de Pearson entre la función discriminante y la variable grupo.

4 La matriz de correlación combinada dentro de grupos se obtiene promediando las g matrices de covarianza y calculando luego, la matriz de correlación. Algunos autores utilizan la letra w para denominar esta matriz.

5 Anteriormente se mencionó la limitación del análisis de la función discriminante cuando existen correlaciones entre las p variables predictivas. Las variables *Habnu* y *Motst* están correlacionadas y, por ende, ellas comparten sus respectivos pesos en la función.

6 Se pudo comprobar en un análisis adicional, donde se utilizaron sólo las dos variables predictoras *Habnu* y *Prbch*, que disminuye significativamente el porcentaje de buena clasificación así como el valor del coeficiente de correlación canónica. Por otra parte, en este análisis el resultado del test de Box sigue siendo significativo.

7 Al realizar el análisis cuadrático discriminante el porcentaje de correcta clasificación (83,7%) fue similar al del análisis lineal (83,67). Así que en este caso es preferible el análisis discriminante lineal por su facilidad de interpretación de los resultados y porque el porcentaje de correcta clasificación no es muy diferente al cuadrático.

## 8. Referencias

González, Pilar (1982). Análisis estadístico del rendimiento estudiantil en La Universidad de Los Andes (mimeo). Mérida: Departamento de Química, Facultad de Ciencias, ULA.

____. (1986). Análisis del rendimiento estudiantil en la Facultad de Ciencias en la Universidad de Los Andes (mimeo). Mérida: Departamento de Química, Facultad de Ciencias, ULA.

____. (1988). "Indicadores sintéticos del rendimiento estudiantil". *Economía* 2: 69-83.

Jhonson, Richard A. and Dean W. Wichern (1982). *Applied Multivariate Statistical Analysis*. Englewood Cliftfs, New Jersey: Prentice-Hall, Inc.

Klecka, William R. (1980). Discrimmant Analysis, Series: *Quantitative Applications in the Social Sciences, a Sage University Paper*, No. 19.

Márquez M., Víctor A. (1989). Apuntes sobre análisis muItivariante, Vol I. Mérida: Universidad de Los Andes. (Mimeo).

Norusis, Marija J. (1986). SPSS/PC+ for the IBM): PC/XT/AT. Chicago IIlinois: SPSS Inc.

Osgood, Ch. *et al*. (1975). *The Measurement of Meaning*. Urbana: University of IIlinois Press.

Peñaloza, Augusto (1987). "Determinación de variables que influyen en el rendimiento académico de estudiantes en Química Básica", *Investigación Educacional* (14): 29.