Universidad de Concepción

DIRECCIÓN DE DOCENCIA
UNIVERSIDAD DE CONCEPCIÓN

Técnicas Estadísticas

Multivariantes para la Investigación en Administración

Profesor Juan Carlos Manríquez Garay

2022 TÉCNICAS ESTADÍSTICAS MULTIVARIANTES PARA LA INVESTIGACIÓN EN ADMINISTRACIÓN

Registro de Propiedad Intelectual N° 2022-A-787 año 2022

I.S.B.N. 978-956-227-517-0

A mi pequeña hija Antonia, fuente de amor, alegría y motivación,
que día a día me hace disfrutar la vida.

Prólogo

El presente texto está dirigido a aquellos alumnos de pregrado, que ya han tenido un curso formal de estadística y de inferencia estadística y que tengan las ganas de incursionar en las aplicaciones de dichas disciplinas. La motivación de este trabajo nace de la experiencia en la enseñanza y ser profesor guía de la asignatura *Investigación en Administración* y entregar nuevas herramientas a la carrera de Ingeniería Comercial de la Universidad de Concepción. Los contenidos cubren aproximadamente un semestre, teniendo en cuenta que cada capítulo, considera la aplicación de un trabajo práctico por parte del alumno. No se requieren de herramientas matemáticas de gran complejidad para abordar los temas de este texto. Sin embargo, es vital el manejo computacional de algunos software.

Desde la aparición de computadores, capaces de almacenar y procesar una gran cantidad de datos, los modelos multivariantes han sido un insumo imprescindible para la investigación en ciencias sociales, medicina, sicología, administración e ingeniería, entre otras materias. El objetivo de este texto es presentar en forma resumida y práctica las herramientas necesarias que debe tener cualquier profesional que se quiera desempeñar bien en un área de trabajo que se requiera procesar y analizar grandes volúmenes de datos para verificar ciertas hipótesis, propias de cualquier investigación de las áreas disciplinarias ya nombradas.

Es sabido que existen distintos textos que abarcan el material que se expone en este trabajo, algunos con un nivel muy básico y otros con un nivel muy complejo en la parte estadística matemática. También hay muchos de ellos que muestran el desarrollo y los pasos de cada caso que se investiga, pero no son generosos en la entrega de una base de datos, para que el investigador practique con ella. La necesidad de un texto de este tipo surge del convencimiento que falta un material académico simplificado que muestre detalladamente los pasos que debe realizar un estudiante de pre y post grado para realizar una investigación aplicada. La idea fundamental es que el estudiante, con base en los ejemplos que se desprenden en cada capítulo, logre extrapolar el caso seleccionado para aplicarlo en el contexto que él está investigando.

Los ejemplos y aplicaciones, han sido seleccionados del texto Malhotra, algunos de ellos no están resueltos en dicho texto, mientras que otros si lo están. Sin embargo,

a este último grupo se le ha agregado un componente explicativo más detallado, para hacerlo más comprensible y amistoso al estudiante. Así también, en cada uno de los ejemplos que se trabajan como investigación, se provee de la base de datos y los pasos que se deben efectuar para el procesamiento y análisis de éstos. De esta menera, se espera que el manejo computacional no sea una piedra de tope para el estudiante.

Al trabajar con ejemplos prácticos, creo que incentivo al lector a pensar en casos similares que lo podrían llevar a desarrollar un trabajo investigativo del área que más le guste y le acomode.

Índice general

CAPÍTULO I

INTRODUCCIÓN AL ANÁLISIS MULTIVARIANTE

Es usual que en las aulas universitarias los procesos de análisis estadístico parten con la recolección de datos, hacer un análisis estadístico basado en estadística descriptiva, y un clásico modelo de regresión lineal múltiple.

El profesional o investigador construye tablas de frecuencia y gráficas, obtiene medidas de tendencia central y dispersión. Si el objetivo es validar algunas características o medidas de una cierta población, la inferencia estadística es la herramienta que da solución a esta propuesta; sin embargo, la estadística como herramienta tiene un amplio espectro de alternativas para tratar y analizar datos que van más allá de los instrumentos que se han señalado hasta aquí.

Algunos ejemplos típicos que demandan una batería de instrumentos estadísticos más avanzados son: un análisis del mercado donde se exploran segmentos o niveles de satisfacción de clientes, un estudio sociológico de una comunidad en donde existen deficiencias en la calidad de vida de sus habitantes, un censo rural que considera un sin número de variables. Estos y otros ejemplos implican una gran cantidad de variables que requieren métodos estadísticos avanzados para lograr resultados más contundentes y fundamentados.

Se pueden dar un sin número de interrogantes que cotidianamente la sociedad enfrenta, como por ejemplo: ¿cuáles son las variables que determinan el rendimiento académico de una población acotada de estudiantes? ¿quiénes son aquellos consumidores que adquieren departamentos y qué usos pretenden darles? ¿por qué difiere cuantiosamente la esperanza de vida de los habitantes de un país? ¿Qué características poseen aquellos alumnos que con problemas de salud mental en una comunidad universitaria? En fin, son múltiples los problemas y las preguntas de investigación que pueden ser abordadas con técnicas estadísticas más avanzadas.

El propósito de este texto es para ser utilizado como una guía de trabajos de investigación para alumnos de pregrado de la carrera de Ingeniería Comercial. Pero como el desarrollo de las materias que se incluyen cubren un espectro un poco más amplio, es que también puede ser utilizado en otras carreras relacionadas con las ciencias sociales.

El texto consta de siete capítulos, dentro de los cuales el 1 corresponde a la introducción y los siguientes a cada técnica multivariante. El capítulo 2 está referido al *análisis factorial*, y lo dejo como punto de partida con la finalidad de motivar al estudiante, pues es un técnica que permite el descubrimiento de variables latentes; aquellas que no son posibles de detectar con la simple observación o encuesta. El

capítulo 3 corresponde al *análisis de conglomerados*, que puede entenderse como la técnica que debe hacerse previamente al *análisis discriminante* (capítulo 4). En el capítulo 5 se encuentra el análisis de varianza; técnica funcional que se emplea básicamente en la experimentación. Los capítulos 6 y 7 se refieren a dos técnicas estadísticas complementarias: *escalamiento multidimensional* y *análisis conjunto*, respectivamente. La primera trata de encontrar la estructura de un conjunto de medidas de distancia entre objetos o casos, en tanto que el análisis conjunto descubre cuáles son aquellas características o atributos de aquellos objetos o casos que se agrupan y difieren entre otro grupo o conjunto de objetos o casos.

CAPÍTULO II

ANÁLISIS FACTORIAL

2.1. Objetivo

Explicar en qué consiste la técnica de análisis factorial, sus variadas aplicaciones en la solución a problemas de la administración y ciencias sociales en general.

2.2. Antecedentes

El análisis factorial es una técnica estructural o de interdependencia (todas las variables independientes) que tiene como principal propósito resumir la información, para describirla más fácilmente; proceso que lleva a la reducción de variables o dimensiones. Esta técnica, junto con exigir un cierto grado de independencia en la mayoría de las variables, requiere también que tales variables sean métricas o medidas en escala de intervalo o razón.

En el análisis factorial las variables observables y medibles se expresan como combinación lineal de un conjunto reducido de factores comunes (Guisande, 2011), por ejemplo, las diferentes manifestaciones como habilidad numérica, capacidad verbal, memoria, habilidad geométrica, capacidad de asociación, etc., se deben a un único factor: inteligencia.

El análisis de factores surge de una necesidad en psicología y matemáticas. Fue el psicólogo Charles Spearman quién aplicó inicialmente el análisis de factores al intentar resolver el dilema acerca de si la inteligencia respondía a un solo factor general o estaba conformada por un conjunto de habilidades específicas (De la Garza,2013). Sin embargo, a inicios de 1960, se crearon grandes equipos de computación, capaces de procesar una gran cantidad de datos. Esto permitió que el análisis factorial pudiese ser aplicado a otras áreas como la investigación de mercados.

2.3. Áreas de aplicación

Esta técnica se puede aplicar a una diversidad de áreas. Por ejemplo en medicina para agrupar variables que puedan diferenciar síntomas o tratamientos (De la Garza,2013); en economía para agrupar variables y diferenciarlas en la aplicación de una política pública o bien analizar los factores subyacentes que determinan la evolución económica de algunos países; en educación superior para agrupar variables que indiquen las competencias que los alumnos requieren para lograr un desempeño eficiente en las empresas; en marketing para analizar la importancia que los

consumidores dan a una cantidad de variables que consideran relevantes para la compra de un inmueble, o bien para identificar un conjunto de variables asociadas a determinados estilos de vida. Como se podrá deducir, son muchas y variadas las aplicaciones que se le pueden dar al análisis factorial.

2.4. Objetivos específicos del análisis factorial

El análisis factorial tiene por objetivos específicos los siguientes (De la Garsa, 2013):

1. Identificar un conjunto de dimensiones o características que se encuentran latentes (es decir, que no se detectan fácilmente) dentro de un conjunto de variables.
2. Encontrar características que describan a núcleos poblacionales (personas u objetos)
3. Identificar nuevas variables, las cuales podrán utilizarse en análisis posteriores, como el de regresión, el discriminante, etc.
4. Crear datos para las nuevas variables encontradas, a partir de la información original.
5. Obtener los mapas de posicionamiento semántico.

Antes de pasar a otro punto, es importante definir qué significa **factor**. Desde el punto de vista matemático, es una combinación lineal de las variables originales de la investigación, representado por la siguiente ecuación (De la Garsa, 2013):

$$F_i = A_1X_1 + A_2X_2 + \ldots + A_kX_k$$

donde:

F = es el factor o componente i de la observación k.
A = representa la importancia o peso que cada variable tiene con respecto al factor encontrado, dicha importancia es relativa, es decir, es la importancia ponderada que tiene cada variable original con respecto a la característica obtenida, pero se debe recordar que las variables originales fueron estandarizadas.
X_j = es la variable original j - ésima.
i = es el número del factor.
k = es el número de la variable.

Los supuestos que el análisis de factores requiere son (De la Garsa, 2013):

1. Las variables involucradas en el análisis deben provenir de poblaciones con distribuciones normales.
2. Las variables deben tener varianzas constantes (homocedasticidad).
3. Las variables incluidas en el análisis deben ser lineales.
4. Las variables, a pesar de ser independientes, deben presentar cierta relación entre ellas para que el análisis factorial se pueda llevar a cabo y tenga sentido realizarlo.

2.5. El modelo factorial

Considérense las variables observables $X_1, X_2, ..., X_p$ como variables estandarizadas (con media 0 y varianza 1). Se define el modelo factorial de la siguiente forma:

$$\begin{aligned} X_1 &= l_{11}\, F_1 + l_{12}\, F_2 + \ldots + l_{1k}\, F_k + e_1 \\ X_2 &= l_{21}\, F_1 + l_{22}\, F_2 + \ldots + l_{2k}\, F_k + e_2 \\ &\vdots \\ X_p &= l_{p1}\, F_1 + l_{p2}\, F_2 + \ldots + l_{pk}\, F_k + e_p \end{aligned}$$

En que $F_1, F_2, \ldots,\ F_k$ son los factores comunes; $e_1, e_2, \ldots, e_k$ son los factores únicos o factores específicos y l_{jh} es el peso del factor h en la variables X_j, denominado también *carga factorial* o *saturación* de la variable X_j en el factor h. Según la formulación del modelo, cada una de las p variables observables es una combinación lineal de k *factores comunes* a todas las variables ($k < p$) y de un factor único para cada variable. Tanto los factores comunes como los específicos son variables no observables. En forma matricial el modelo se presenta como sigue:

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1k} \\ l_{21} & l_{22} & \cdots & l_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pk} \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_k \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_k \end{pmatrix}$$

o lo que es lo mismo: $\mathbf{X}$=$\mathbf{LF}$+$\mathbf{e}$

2.6. Estadísticas asociadas con el análisis factorial

Las principales estadísticas asociadas con el análisis factorial son las siguientes (Malhotra, 2004):

Cargas de los factores. Correlaciones simples entre las variables y los factores.
Gráfica de acumulación. Gráfica de valores propios y el número de factores en orden de extracción.
Gráfica de las cargas de los factores. Gráfica de las variables originales en la que las coordenadas son las cargas de los factores.
Matriz de correlación. Matriz del triángulo inferior en el que se muestran las correlaciones, r, entre todos los pares posibles incluidos en el análisis. Por lo general se omiten los elementos de la diagonal, que son todos igual a 1.
Matriz factorial. Matriz que contiene las cargas de los factores de todas las variables de todos los factores extraídos.
Medida de la adecuación de la muestra de *Kaiser - Meyer - Olkin (KMO).* Índice con el que se examina si el análisis factorial es el apropiado. Valores elevados (entre 0.5 y 1.0) indican que el análisis factorial es apropiado. Los valores inferiores a 0.5 implican que el análisis factorial no es apropiado.
Porcentaje de varianza. Porcentaje de la varianza total atribuida a cada factor.
Prueba de esfericidad de Bartlett. Ésta es una prueba estadística para examinar la hipótesis de que las variables no se correlacionan en la población. En otras palabras, la matriz de correlación de la población es una matriz de identidad. Si $r = 1$ cada variable se correlaciona perfectamente con ella misma, pero no se correlaciona con las otras ($r = 0$).
Puntuaciones de los factores. Puntuaciones compuestas que se estiman con los factores derivados de cada encuestado.
Residuos. Diferencias entre las correlaciones observadas, como están introducidas en la matriz de correlación, y las correlaciones reproducidas, según se calcula en la matriz factorial.
Valor propio. Varianza total explicada por cada factor.
Variación común. Monto de la varianza que una variable comparte con las demás variables consideradas. También es la proporción de la varianza explicada por los factores comunes.

2.7. Pasos para llevar a cabo el análisis factorial

Los pasos del análisis de factores serán descritos con el siguiente ejemplo:

Se estudió la relación entre el comportamiento hogareño y el comportamiento de compras, se obtuvieron datos sobre los siguientes enunciados de estilo de vida, en una escala de siete puntos (1 = desacuerdo, 7 = de acuerdo). Los resultados del cuestionario se escribieron en la siguiente tabla (base de datos).

Tabla 2.1: Base de datos del estilo de vida (*Malhotra, 2004*).

N°	CASA	PRECIO	REVISTA	ANUNCIO	HOGAR	AHORRO	PUBLICIDAD
1	6	2	7	6	5	3	5
2	5	7	5	6	6	6	4
3	5	3	4	5	6	6	7
4	3	2	2	5	1	3	2
5	4	2	3	2	2	1	3
6	2	6	2	4	3	7	5
7	1	3	3	6	2	5	7
8	3	5	1	4	2	5	6
9	7	3	6	3	5	2	4
10	6	3	3	4	4	6	5
11	6	6	2	6	4	4	7
12	3	2	2	7	6	1	6
13	5	7	6	2	2	6	1
14	6	3	5	5	7	2	1
15	3	2	4	3	2	6	5
16	2	7	5	1	4	5	2
17	3	2	2	7	2	4	6
18	6	4	5	4	7	3	3
19	7	2	6	2	5	2	1
20	5	6	6	3	4	5	3
21	2	3	3	2	1	2	6
22	3	4	2	1	4	3	6
23	2	6	3	2	1	5	3
24	6	5	7	4	5	7	2
25	7	6	5	4	6	5	3

Fuente: Malhotra, 2004

CASA: Prefiero pasar una tarde tranquila en casa que ir a una fiesta.
PRECIO: Siempre verifico los precios, incluso de artículos menores.
REVISTAS: Las revistas son más interesantes que las películas.
ANUNCIOS: No compraría productos anunciados en carteleras espectaculares.
HOGAREÑO: Soy del tipo hogareño.
AHORRO: Ahorro y uso cupones.
PUBLICIDAD: Las compañías gastan mucho dinero en publicidad.

2.7.1. Planteamiento del problema.

En primer lugar, se deben identificar los objetivos del análisis factorial, es decir, si realmente el análisis factorial da respuesta al objetivo de la investigación. En segundo lugar, se deben especificar las variables a incluir en el análisis, sobre la base de investigaciones pasadas, una teoría y el propio juicio del investigador (Malhotra, 2004). Es importante que las variables se midan en una escala de intervalo o razón. Por último, considerar una muestra de tamaño adecuada de variables. Como regla, se sigiere un tamaño de muestra superior a cinco veces más observaciones que variables (Malhotra, 2004).

Asumiendo que el análisis factorial da respuesta a la técnica estadística adecuada para resolver el problema planteado, de debe tomar la decisión del tipo de análisis factorial que se debe aplicar, lo cual se resuelve con la siguiente pregunta:

¿Qué se desea agrupar?. La respuesta a esta interrogante se da con la siguiente tabla:

Tabla 2.2: Tipo de análisis factorial

Si se desea	**Si en la base de datos se tiene en las…**		**Usar análisis factorial**
	columnas…	**filas…**	
Variables	variables	personas	R
	variables	periodos o lugares	P
Personas	variables	personas	Q
	personas	periodos o lugares	S
Periodos o lugares	variables	periodos o lugares	O
	personas	periodos o lugares	T

Fuente: Elaboración propia

De acuerdo con el ejemplo, lo que se quiere es reducir es el número de variables (pues hay siete), lo que se debe hacer es un análisis factorial tipo R.

2.7.2. Estandarización o tipificación de las variables.

Una vez almacenada la información es recomendable estandarizar las variables (pero no es necesario, porque se puede trabajar con la matriz de varianzas y covarianzas); si se desea estandarizar el SPSS lo puede realizar. Más adelante y en este capítulo, se explicará la utilidad que tienen las variables al estar estandarizadas.

2.7.3. Generación de la matriz de correlación.

En el análisis de factores debe existir cierto grado de relación entre los grupos de variables (el valor - p da a conocer cuáles correlaciones son significativas), de no ser así no es necesario usar análisis factorial.

Tabla 2.3: Matriz de correlación del ejemplo comportamiento de compra y hogareño.

		CASA	PRECIO	REVISTA	ANUNCIO	HOGAR	AHORRO	PUBLICIDAD
Correlación	CASA	1,000	-,004	,628	,082	,675	-,100	-,338
	PRECIO	-,004	1,000	,151	-,248	,048	,582	-,251
	REVISTA	,628	,151	1,000	-,182	,480	,090	-,588
	ANUNCIO	,082	-,248	-,182	1,000	,272	,017	,469
	HOGAR	,675	,048	,480	,272	1,000	-,110	-,082
	AHORRO	-,100	,582	,090	,017	-,110	1,000	,014
	PUBLICIDAD	-,338	-,251	-,588	,469	-,082	,014	1,000
Sig. (unilateral)	CASA		,493	,000	,348	,000	,316	,049
	PRECIO	,493		,236	,116	,409	,001	,113
	REVISTA	,000	,236		,192	,008	,334	,001
	ANUNCIO	,348	,116	,192		,094	,469	,009
	HOGAR	,000	,409	,008	,094		,301	,348
	AHORRO	,316	,001	,334	,469	,301		,473
	PUBLICIDAD	,049	,113	,001	,009	,348	,473	

Determinante = 0.062

La tabla 2.3 muestra la correlación que arroja SPSS del ejemplo, relación entre el comportamiento hogareño y el comportamiento de compra. Las correlaciones significativas (sean directas o inversas) se muestran en la parte inferior de esta tabla y se identifican por un valor - p inferior o igual a 0,05.

La matriz de correlación es un primer análisis que indica cuáles variables posiblemente quedarán agrupadas en el mismo factor y cuáles posiblemente no lo hagan. De esta manera se observa que existe una aceptable correlación entre las variables REVISTA, CASA y HOGAR, por lo que posiblemente las tres variables queden juntas en algún factor. Sinn embargo, se ve también que hay una correlación bastante baja entre HOGAR y PRECIO, por lo que probablemente no queden en el mismo factor; en pocas palabras, estas inspecciones pueden dar cuenta de la posible solución.

2.7.4. Determinante de una matriz de correlación.

El determinante de la matriz de correlación oscila entre 0 y 1 ($0 \leq |R| \leq 1$); en donde R es la matriz de correlación y $|R|$ es el determinante de la matriz de correlación. Si el valor del determinante es cercano a cero indica la conveniencia de realizar análisis factorial para analizar el problema y si es cercano a 1, entonces el análisis factorial no es el adecuado. En este caso se obtuvo un determinante igual a 0,062, lo que se considera cercano a cero.

2.7.5. Prueba de contraste de esfericidad de Bartlett.

La prueba de contraste de esfericidad de Bartlett es una alternativa más precisa para discriminar sobre el uso o no aplicación del análisis factorial para un determinado problema. Como ya se mencionó, es necesario que exista un grado de multicolinealidad entre algunas variables ya que la técnica de análisis de factores identifica las variables que están interrelacionadas, en caso contrario la matriz de correlación sería una matriz identidad y no tendría sentido llevar a cabo el análisis factorial (De la Garza, 2013). Esto se verifica por medio de la prueba de contraste de esfericidad de Bartlett; sus hipótesis serían:

$H_0 : |R| = I$; no se debe utilizar la técnica de análisis factorial para resumir la información de la base de datos; versus la hipótesis alternativa $H_a : |R| \neq I$; no se debe utilizar la técnica de análisis factorial para resumir la información de la base de datos.

Donde:

R : es la matriz de correlación.
|R|: es el determinante de la matriz de correlación.
I : es la matriz identidad.

Volviendo al ejemplo de la relación entre comportamiento hogareño y el comportamiento de compra, SPSS muestra el siguiente resultado:

Tabla 2.4: Prueba KMO y de Bartlett

Medida Kaiser-Meyer-Olkin de adecuación de muestreo		,550
Prueba de esfericidad de Bartlett	Aprox. Chi-cuadrado	57,994
	gl	21
	Sig.	,000

Entonces se realiza la prueba de hipótesis de esfericidad de Bartlett, bajo las hipótesis ya indicadas.

La fórmula para el cálculo de la χ^2 es la siguiente:

$$\chi^2 = -\left[n - 1 - \frac{1}{6}(2m + 5)\right] ln|R|$$

Con los grados de libertad $= 0{,}5(m^2 - m)$.

Donde:

m: es el número de variables.
n: es el número de observaciones.
$|R|$: es el determinante de la matriz de correlación.

De acuerdo con el ejemplo, *m=7*, *n=25* y $|R|$*=0,062*, se calcula un $\chi_c^2 = 57,93$; mientras que por otro lado, con un error tipo I $\alpha = 0,05$ y con *21* grados de libertad, se obtiene de tabla $\chi^2_{0,05,21} = 11,5913$. Como en este caso $\chi_c^2 > \chi^2_{tabla}$, entonces se rechaza la hipótesis nula $H_0 : |R| = I$. Por lo tanto, llevar a cabo el análisis factorial sí tiene sentido.

2.7.6. Análisis de suficiencia general o Kaiser - Meyeer - Olkin.

La medida de suficiencia o adecuación del muestreo general (*MASg*, por sus siglas en inglés) o *KMO* (Kaiser - Meyeer - Olkin) es una medida global que indica si se llevara a cabo el análisis de factores, qué tan fuerte y adecuada sería la posible solución a encontrar (De la Garsa, 2013). Como referencia se puede considerar lo siguiente:

Tabla 2.5: Evaluación para la medida de adecuación.

MASg	Evaluación
Menor a 0,5	Inaceptable
[0,5 a 0,6[	Bajo
[0,6 a 0,7[	Regular
[0,7 a 0,8[	Aceptable
[0,8 a 0,9[	Bueno
De 0,9 en adelante	Excelente

Fuente: Análisis Estadístico Multivariante (De la Garsa, 2004).

Volviendo al ejemplo de la conducta hogareña y comportamiento en el consumo; en la tabla 2.4 se muestra un MASg igual a $0,55$, lo que es considerado bajo. Este es un indicativo de que una variable no debiera estar en el análisis ¿cuál?; se tratará en el siguiente punto.

2.7.7. Análisis de adecuación individual.

Después del análisis global se tiene que analizar cada variable y esto se lleva a cabo en la matriz antiimagen de la matriz de correlación. Aquí se obtiene la *MASi* (por sus siglas en inglés y es la medida de adecuación muestral individual de cada variable), el resto son las correlaciones parciales (De la Garsa, 2003).

Las correlaciones parciales con valores muy grandes indican que los datos de la variable no son adecuados y que ésta no debiera estar en el análisis, es decir, fuera de la diagonal de la matriz antiimagen no deberán colocarse valores grandes (valores absolutos), pero los valores del *MASi*, los de la diagonal principal, deberán ser valores grandes. Si $MASg \geq 0,5$ se ve la matriz antiimagen, pero sin olvidar que si el *MASi* de las variables es menor a *0,5* la variable deberá ser eliminada del estudio.

En la tabla 2.6 se observa que las variables *precio* y *ahorro* tienen una medida de adecuación menor a *0,5*. Sin embargo, el procedimiento sugiere que se debe dejar fuera la variable que muestra una menor *MASi* y volver a realizar el análisis factorial.

Tabla 2.6: Matriz antiimagen del ejemplo comportamiento de compra y hogareño

		CASA	PRECIO	REVISTA	ANUNCIO	HOGAR	AHORRO	PUBLICIDAD
Covarianza anti-imagen	CASA	,404	,044	-,124	-,037	-,202	,014	,071
	PRECIO	,044	,514	,078	,155	-,139	-,340	,126
	REVISTA	-,124	,078	,378	,077	-,126	-,135	,204
	ANUNCIO	-,037	,155	,077	,618	-,156	-,137	-,154
	HOGAR	-,202	-,139	-,126	-,156	,412	,136	-,109
	AHORRO	,014	-,340	-,135	-,137	,136	,536	-,114
	PUBLICIDAD	,071	,126	,204	-,154	-,109	-,114	,463
Correlación anti-imagen	CASA	,716[a]	,097	-,318	-,075	-,495	,029	,165
	PRECIO	,097	,414[a]	,176	,274	-,301	-,648	,259
	REVISTA	-,318	,176	,638[a]	,159	-,319	-,299	,489
	ANUNCIO	-,075	,274	,159	,537[a]	-,308	-,238	-,288
	HOGAR	-,495	-,301	-,319	-,308	,535[a]	,290	-,249
	AHORRO	,029	-,648	-,299	-,238	,290	,345[a]	-,229
	PUBLICIDAD	,165	,259	,489	-,288	-,249	-,229	,586[a]

a:Medidas de adecuación de muestreo

Al dejar fuera la variable ahorro (tiene un $MASi = 0,345$), se vuelve a realizar nuevamente un *MASi*, obteniendo lo siguiente:

Tabla 2.7: Prueba KMO y de Bartlett

Medida Kaiser-Meyer-Olkin de adecuación de muestreo		,664
Prueba de esfericidad de Bartlett	Aprox. Chi-cuadrado	45,712
	gl	15
	Sig.	,000

Como se puede apreciar el *MASg* aumentó de *0,55* a *0,664* con alto nivel de significancia(*0,000*), por lo que llevar a cabo la técnica factorial tiene sentido. A continuación, se presenta nuevamente la matriz antiimagen para revisar los valores del *MASi* (tabla 2.8).

Tabla 2.8: Matriz antiimagen del ejemplo comportamiento de compra y hogareño

		CASA	PRECIO	REVISTA	ANUNCIO	HOGAR	PUBLICIDAD
Covarianza anti-imagen	CASA	,404	,091	-,133	-,036	-,224	,078
	PRECIO	,091	,886	-,015	,123	-,098	,098
	REVISTA	-,133	-,015	,415	,049	-,110	,204
	ANUNCIO	-,036	,123	,049	,655	-,140	-,205
	HOGAR	-,224	-,098	-,110	-,140	,450	-,092
	PUBLICIDAD	,078	,098	,204	-,205	-,092	,489
Correlación anti-imagen	CASA	,688[a]	,152	-,324	-,070	-,526	,176
	PRECIO	,152	,609[a]	-,025	,162	-,155	,148
	REVISTA	-,324	-,025	,728[a]	,095	-,254	,452
	ANUNCIO	-,070	,162	,095	,625[a]	-,257	-,363
	HOGAR	-,526	-,155	-,254	-,257	,621[a]	-,196
	PUBLICIDAD	,176	,148	,452	-,363	-,196	,637[a]

a:Medidas de adecuación de muestreo

En esta matriz antiimagen (tabla 2.8), ninguna variable presenta un *MASi* inferior a *0,5*; en consecuencia para seguir con el análisis factorial, se ha de considerar todas las variables originales a excepción de la variable *ahorro*.

2.8. Determinación del método de análisis factorial

Una vez que se ha determinado que el análisis factorial es una técnica apropiada para analizar los datos, debe elegirse el método apropiado. Los métodos de análisis factorial se distinguen por la manera de derivar los pesos o los coeficientes de las puntuaciones de los factores. Los dos métodos básicos son el análisis de los componentes principales (*ACP*) y el modelo de factor común. El ACP es útil si el objetivo es confirmar una teoría o hipótesis previamente establecida, lo que se llamará *confirmatorio*. El modelo de factor común se enfoca en un análisis de tipo exploratorio, pues se desea estudiar dentro de un conjunto de datos esa estructura latente que hace que las variables muestren interrelación (De la Garsa, 2013).

Para explicar estos modelos se debe recordar que, al estandarizar las variables, la varianza de cada una de ellas es 1. Como ya se mencionó, el análisis factorial busca aspectos comunes entre las variables para agruparlas y la dispersión o variación de los datos indica posibles similitudes entre las variables; el análisis factorial conceptualiza la variación de cada variable como se indica a continuación:

$$Var\ Total = Var\ Común + Var\ Específica + Var\ Aleatoria$$

Donde:

Var Total : variación o dispersión de la variable. Cuando las variables están estandarizadas, el valor de esta variación es 1.
Var Común : variación que hace parecidas a las variables y es lo que las une.

Var Específica : es la que hace diferentes y únicas a las variables, por lo que dicha variación separa a las variables.

Var Aleatoria : representa el error o el azar; se asume su presencia, pero en virtud de que no se puede cuantificar o calcular se le considera como mínimo su efecto.

2.8.1. Componentes pricipales.

Este modelo asume que la variación específica es tan pequeña que la considera como cero y desprecia la variación aleatoria. Por lo tanto, la variación total se considera igual a la variación común.

$$Var\ Total = Var\ Común$$

Esto significa que el máximo valor que puede tomar la variación común es 1. En un análisis de tipo confirmatorio se supone que el factor está formado de una combinación lineal de las variables originales.

$$F_{1j} = U_1X_{1j} + U_2X_{2j} + U_3X_{3j} + \cdots + A_kX_{kj}$$

Donde:

F_{1j} = es el factor 1 de la observación j
U = representa la importancia o peso relativo que cada variable estandarizada tiene con respecto al factor encontrado.
X = es la variable.
j = es el número de la observación.
k = es el número de la variable.

Así toda la información contenida en las variables que se analizan, nada es específico, todo es variación común.

2.8.2. Factor común.

En este modelo se asume que la variación específica es importante y que por lo tanto se deberá calcular y eliminar del modelo de variación para que con el resto se intente agrupar a las variables.

Var Total - Var Específica = Var Común + Var Específica - Var Específica + Var Aleatoria

Var Total - Var Específica = Var Común , *Var Aleatoria = 0*

Esto significa que el máximo valor que puede tomar la variación común es menor que 1, pero si la variación específica es cero, entonces puede llegar a valer 1.

En un análisis de tipo exploratorio se supone que las variables comparten algo que es común del grupo y una parte de ellas es específico o propio, esta última se elimina y solo se trabaja con la parte común. Así se tiene la siguiente ecuación conceptual en la que se asume que están formados los datos de cualquiera de las variables.

$$
\begin{aligned}
X_{1j} &= U_{11j}F_1 + U_{21j}F_2 + U_{31j}F_3 + \cdots + U_{m1j}F_m + e_{1j} \\
X_{2j} &= U_{12j}F_1 + U_{22j}F_2 + U_{32j}F_3 + \cdots + U_{m2j}F_m + e_{2j} \\
&\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
X_{ij} &= U_{1ij}F_1 + U_{2ij}F_2 + U_{3ij}F_3 + \cdots + U_{mij}F_m + e_{ij} \\
&\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
X_{pj} &= U_{1pj}F_1 + U_{2pj}F_2 + U_{3pj}F_3 + \cdots + U_{mpj}F_m + e_{pj}
\end{aligned}
$$

Donde:

F = son los factores comunes.
U = representa la importancia o peso relativo que cada variable estandarizada tiene con respecto al factor encontrado.
X = es la variable i de la observación j.
i = es el número de la variable.
j = es el número de la observación.
m = es el número máximo de factores.
p = es el número máximo de variables.
e = son los factores únicos o específicos.

2.9. Descomposición espectral, singular o única

Una vez que se ha escogido, el modelo, a la matriz de correlación que es la entrada de esta técnica, se le aplica el procedimiento matemático llamado *descomposición espectral* o *singular*.

Definición:

Sea A una matriz cuadrada, entonces la expresión:

$$A - \lambda I$$

Donde:

A = es la matriz característica.
λ = son los valores propios (*eigenvalues*) de la matriz A.
I = es la matriz identidad.

Los pasos para la descomposición espectral son los siguientes:

1) Se calculan los valores propios para cualquier matriz cuadrada A, ésta tendrá un valor de λ que satisface la siguiente expresión:

$$|A - \lambda I| = 0,$$

es decir el determinante de la matriz característica de A es igual a cero, y existirá más de un vector propio (*eigenvector*) dependiendo del rango de la matriz A. Si se hace $A = R$, donde R es la matriz de correlación:

$$|R - \lambda I| = 0,$$

entonces a la matriz de correlación se le extraerán los eigenvalores y el número de éstos depende de las dimensiones de la matriz y de cuántas variables se tienen en el análisis.

2) Se calcula para cada valor propio su vector propio. Si ν representa los vectores propios de A y además satisfacen la siguiente ecuación matricial:

$$(A - \lambda I)\nu = 0$$

Para el caso $A = R$, entonces los eigenvectores a extraer serán los que satisfacen la ecuación siguiente:

$$(R - \lambda I)\nu = 0$$

Volviendo al ejemplo, en el análisis se tenían siete variables, pero una fue eliminada por el criterio de la MAS_i, por lo que la matriz de correlación sería de $6x6$, así se extraerían 6 valores propios, los cuales se presentan en la siguiente tabla:

Tabla 2.9: Variación total explicada que muestra los valores propios

Componente	Autovalores iniciales			Suma de extracción de cargas al cuadrado			Suma de rotación de cargas al cuadrado		
	Total	% Varianza	% Acumulado	Total	% Varianza	% Acumulado	Total	% Varianza	% Acumulado
1	2,484	41,402	41,402	2,484	41,402	41,402	2,318	38,641	38,641
2	1,677	27,947	69,349	1,677	27,947	69,349	1,842	30,708	69,349
3	,862	14,372	83,721						
4	,428	7,128	90,849						
5	,291	4,847	95,696						
6	,258	4,304	1,000						

Método de extracción: Análisis de componentes principales

La tabla 2.9 muestra las 6 componentes o factores (el máximo de componentes siempre será igual al número de variables, lo cual no tiene sentido) donde cada una tiene asociado un valor propio, dicho número indica la cantidad de varianza (información) explicada. Por ejemplo, a la componente número dos, le corresponde el valor propio $\lambda_2 = 1,677$ y éste concentra un $27,9\,\%$ de la información. En la cuarta columna de la tabla se ve el porcentaje de información acumulada por los factores. Así se tiene que, en conjunto los factores 1 y 2 acumulan un $69,349\,\%$ de la información (se pierde solo un $30,651\,\%$ de ella).

En este ejemplo, se han dejado solamente dos factores (componentes 1 y 2) para explicar solo dos estilos de vida, dado el conjunto de variables seleccionadas para el análisis. El criterio que se utilizó para dejar solo dos factores corresponde al *criterio de la raíz latente*, cuyo método se explicará más adelante.

Para un único valor propio hay asociado un único vector propio y cada vector propio se compone de cargas del factor, las que representan el grado de relación que existe entre ellas y el factor (componente).

Como se han dejado dos componentes, la tabla 2.10 muestra solo los vectores propios de los factores 1 y 2. En la columna de la primera componente se ven buenas correlaciones con las variables *casa, revista, hogar* y *publicidad*. Por ejemplo, la variable *revista* muestra una correlación positiva de 0,886 con la componente 1, pero una muy baja correlación con la segunda componente $(-0,043)$. Para calcular el valor propio de cada factor, cada carga en la matriz no rotada (tabla 2.10) se eleva al cuadrado y se suma para cada factor. Los valores propios de cada factor aparecen en la tabla 2.9.

Tabla 2.10: Matriz de componentes

	Componente	
	1	2
CASA	,823	,371
PRECIO	,258	-,474
REVISTA	,886	-,043
ANUNCIO	-,199	,833
HOGAR	,671	,575
PUBLICIDAD	-,682	,536

Dos componentes extraidos

Como ya se mencionó, la tabla 2.10 es una matriz de componentes o factores no rotados. Dicha matriz permite dar una primera aproximación para interpretar el grado de correlación entre cada variable y su respectivo factor. Sin embrago no permite una interpretación plena del grado de correlación entre los elementos ya señalados. Por tal razón, se hace uso preferente de la matriz de factores rotados (tabla 2.11). Obsérvese como las variables *hogar* y *publicidad* en la matriz de componentes (tabla 2.10) tienen valores absolutos superiores a 0,5 para las componentes 1 y 2 ($0,671$ con $-0,682$ en la primera componente y $0,575$ con $0,536$ en la segunda componente, respectivamente). Al efectuar la rotación, se obtienen las correlaciones más precisas de cada variable con cada factor. Ahora las correlaciones de las variables *hogar* y *publicidad* quedan claramente determinadas. Por ejemplo, la variable *hogar* está fuertemente correlacionada con la primera componente $(0,859)$ y la variable *publicidad* está también fuertemente correlacionada con la componente 2.

Tabla 2.11: Componentes rotados

	Componente	
	1	2
CASA	,902	-,042
PRECIO	,015	-,540
REVISTA	,771	-,439
ANUNCIO	,200	,833
HOGAR	,859	,209
PUBLICIDAD	-,365	,787

Dos componentes extraídos

2.9.1. Comunalidad.

La comunalidad expresa la proporción de varianza de una variable explicada por el conjunto de factores seleccionados (Pedret, 2000). Indirectamente, explica el porcentaje de información que se está perdiendo al trabajar con un espacio determinado. Una comunalidad elevada (cercana a 1) implicará una correlación elevada con *al menos uno* de los factores seleccionados, en cambio una comunalidad baja implicará una correlación baja con *todos* los factores seleccionados, es decir, estas variables estarán correlacionadas con otros factores.

Tabla 2.12: Comunalidades

	Inicial	Extracción
CASA	1,000	,815
PRECIO	1,000	,292
REVISTA	1,000	,787
ANUNCIO	1,000	,734
HOGAR	1,000	,781
PUBLICIDAD	1,000	,752

Método de extracción: ACP

En el ejemplo del capítulo, la tabla 2.12 de la salida de SPSS muestra dos columnas. En la columna extracción se ve que todas las variables están bien representadas, algunas más que otras, con excepción de la variable *precio*.

Como ya se ha señalado, la variación de cada variable estandarizada es de 1, por lo tanto el valor máximo de información que se puede manejar con cada variable, o sea el valor máximo posible de la comunalidad, es de 1. Mientras más se aproxime al valor de la comunalidad a 1, más información se tiene en los factores de la variable.

2.9.2. Criterios para la determinación del número de factores a ser extraídos como solución inicial.

Los criterios para determinar el número de factores como posible solución inicial son los siguientes:

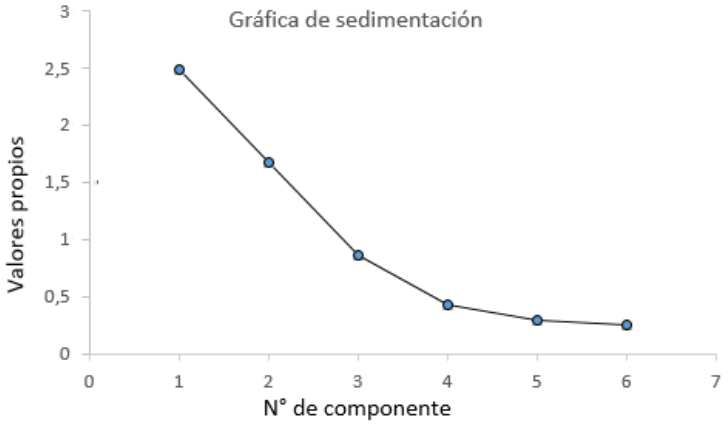
1. *Crierio a priori*. Se utiliza en aquellos casos donde el investigador desea probar alguna teoría o hipótesis y de antemano conoce cuántos factores se deben tener en la solución; el investigador decide el número de factores a pedir con base en su hipótesis. Se recomienda si se usa la técnica de análisis factorial confirmatorio (De la Garsa, 2013). En el ejemplo del capítulo no se estableció un hipótesis sobre el número de factores.

2. *Criterio de la raíz latente.* En éste se pueden tomar en cuenta las consideraciones siguientes: en caso de que los datos no estén estandarizados, la idea es retener a los factores cuya raíz característica exceda la medida de las raíces características. La **raíz característica** es la variación explicada por cada factor. Cuando los datos están estandarizados se considera que un factor debe ser retenido en su solución si su raíz característica es mayor a 1. Debido a que los valores propios o *raíces características* representan las varianzas y las varianzas de las variables estandarizadas son igual a 1, un factor con un valor propio menos a 1 no es importante y se considera que no tiene la cantidad de información significativa captada; lo lógico es mantener a los factores que tengan mayor información (varianza) que cualquiera de las originales (ver tabla 2.9). Se recomienda este criterio si el objetivo del análisis de factores es para un análisis exploratorio. Este criterio es usado por la mayoría de los paquetes computacionales cuando no se les especifica un cierto número de factores en la solución (De la Garsa, 2013).

3. *Criterio del porcentaje de variación explicada acumulada.* Mediante este criterio se considera que "n" factores deben manejarse como solución inicial, si el porcentaje de variación explicada acumulada se encuentra en un rango de entre 60 y 95 %. Es decir, a través de este criterio se está dispuesto a perder cuando mucho 40 % y cuando menos 5 % de información (De la Garsa, 2013). En la tabla 2.9, se puede deducir que se está perdiendo un 30,651 % de información, al considerar solo los dos primeros factores.

4. *Criterio scree test o gráfica de sedimentación.* Fue desarrollado por el psicólogo británico Raymond Bernard Cattell; *scree* es un término usado en geología, es el escombro al final de un precipicio. La idea en el *scree test* es que si los factores son importantes tendrán una varianza grande. Al extraer los factores, el primero contabiliza la mayor variación, el segundo que se extrae tiene menor variación que el primero y así sucesivamente. En virtud de esta manera natural en que se extraen, para la aplicación de este criterio se realiza una gráfica en la que se ubica el número del factor en el eje de las x y los valores de las varianzas o los valores propios en el eje de las y. Al unir estos puntos se obtiene una figura que es similar al perfil de una montaña con una pendiente fuerte hasta llegar a la base, formada por una meseta con una ligera inclinación. En esta meseta es en donde se acumula el escombro tirado desde la cumbre, es decir, donde se sedimenta. La idea es quedarse con los factores previos a la sedimentación (De la Garsa, 2013).

Figura 2.1: Scree plot o gráfica de sedimentación del ejemplo.



A diferencia del criterio de la raíz latente en el que se pedía que la variación explicada fuera mayor a 1 (valor propio mayor a 1) para aceptar a un nuevo factor en la solución; este criterio podría dar como resultado soluciones con 1 ó 2 factores más que la raíz latente.

En la figura 2.1 si se sigue con el criterio antes explicado en este ejemplo quedarían 3 factores, pues ese es el punto en donde las variaciones (valores propios) se empiezan a sedimentar; es a partir del factor 3, en este caso, en donde la pendiente comienza a ser menos pronunciada.

Estos son los cuatro criterios que ayudan a determinar el número de factores a manejar en una solución inicial del análisis factorial (De la Garsa, 2013).

2.9.3. Obtención de la matriz de factores rotada.

La matriz de factores rotada, al igual que la no rotada, se genera por computadora y contiene información referente al grado de explicación de las variables por los factores, es decir, las cargas de factores, pero ahora serán las definitivas las que permitirán determinar la agrupación de las variables en dichos factores, se busca obtener así una solución final lógica (De la Garsa, 2013).

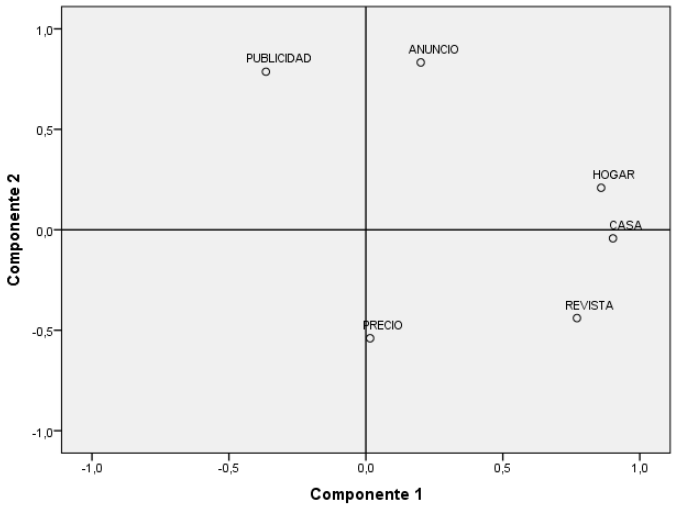
Mediante este proceso se ajustan los ejes coordenados o *ejes factores* con el fin de obtener una solución más sencilla y con mayor significado teórico; esto es, hacer que una de las cargas de la variable sea lo más alta posible para que se identifique con un solo factor, mientras que sus demás cargas sean bajas, de manera que pierda su relación con otros factores (ver tablas 2.10 y 2.11).

Al rotar los factores las cargas cambian en virtud de una nueva posición de los ejes, es decir, cambia el valor de los valores propios para cada factor, pero la variación total explicada y las comunalidades no cambian, a menos que se eliminen o agreguen factores. Entonces, al rotar, no se explica mayor o menor información; la variación total se mantiene, así como las comunalidades.

Gráficamente, lo que se trata de hacer al rotar los factores es que los ejes se acerquen a las variables y queden lo más cerca posible a éstos y que las variables se agrupen bajo una solución característica, lo cual sería la mejor solución.

A continuación, la figura 2.2 muestra la posición que tiene cada variable en sistema de coordenadas factoriales.

Figura 2.2: Gráfico de componente en espacio rotado.



Se puede observar como las variables *anuncio*, *publicidad* y *precio* se adhieren a la componente 1 y las variables *hogar*, *casa* y *revista* tienden a agruparse en torno a la componente 2.

2.9.4. Criterio para la rotación de factores.

Entre los métodos de rotación ortogonal están los siguientes (De la Garsa, 2013):

1. *Varimax.* El término *varimax* proviene de que la varianza se *maximiza*. Mediante este criterio se trata de identificar a un grupo de variables con solo un factor, es decir, simplifica por componente o columna; busca la máxima simplificación al generar tantos unos y ceros como le sea posible en la matriz. Este método de rotación es el más utilizado.

2. *Quartimax*. La rotación de factores tendrá por objetivo identificar a cada variable con al menos uno de los factores, esto es, simplifica por variable o renglón, tratando de que las cargas de la mayoría de los factores sean tan bajas como se pueda y al menos una lo suficientemente alta para considerarse como significativa.

3. *Equamax*. Es una solución intermedia entre las anteriores; la simplificación se hace en forma alternativa, esto es, se simplifica por renglón o por columna indiscriminadamente en la matriz de factores.

Los métodos de rotación oblicua son los siguientes: *Oblimin directo, Promax* y *Quartimin*. No se detallará en que consiste cada uno debido a que presentan algunos problemas en su aplicación (Consultar De la Garsa, 2013, p.362).

2.9.5. Criterio para identificar cargas significativas.

Lo significativo que puede ser una carga factorial, dependerá del tamaño de la muestra y del nivel de significancia manejado en la investigación. Se requiere conocer el tamaño de la muestra, ya que conforme va aumentando se tiene una mayor credibilidad o confianza en la información y por tanto se puede manejar un límite cada vez menor en la carga del factor, para considerarla significativa. Por el contrario, al manejar un nivel de significancia menor se es más estricto con la confiabilidad del factor, y por lo tanto, se manejará un límite cada vez mayor en la carga del factor para considerarla significativa (Consultar De la Garsa, 2013, p.363).

2.9.6. Interpretación de los factores.

La interpretación se facilita si se identifican las variables que tienen cargas grandes en el mismo factor. Así, este factor puede interpretarse en términos de tales variables que tienen mayor carga en él (Malhotra, 2004).

En la matriz factorial rotada (tabla 2.11), el factor 1 tiene coeficientes elevados para las variables *casa* (preferiría pasar una tarde tranquila en casa (...)), *hogar* (soy del tipo hogareño) y *revista* (las revistas son más interesantes (...)). Por tanto el factor 1 puede interpretarse como un estilo de vida hogareño. El factor 2 está muy relacionado con las variables *anuncio* (no compraría productos anunciados en carteleras espectaculares), *publicidad* (las compañías gastan mucho dinero en publicidad) y *precio* (siempre verifico los precios (...)). Así el factor 2 podría interpretarse como un estilo de vida asociado a un comprador racional.

2.10. Cálculo de la puntuación de los factores

Después de la interpretación se calcula la puntuación de los factores, si es necesario. El análisis factorial tiene un valor en sí mismo; ahora bien, si su finalidad es reducir las variables originales a un conjunto menor de variables compuestas (factores) para el siguiente análisis multivariado, es útil calcular puntuaciones de los factores de cada encuestado. Un factor no es más que una combinación lineal de las variables originales (Malhotra, 2004). Las **puntuaciones de los factores** para el factor *i-ésimo* se estiman como sigue:

$$F_{1j} = U_1X_{1j} + U_2X_{2j} + U_3X_{3j} + \cdots + A_kX_{kj}$$

Donde:

F_{1j} = es el factor 1 de la observación j
U = representa la importancia o peso relativo que cada variable estandarizada tiene con respecto al factor encontrado.
X = es la variable estandarizada.
j = es el número de la observación.
k = es el número de la variable.

Los pesos, o coeficientes de las puntuaciones de los factores, con los que se combinan las variables estandarizadas se obtienen de la matriz de los coeficientes de esas puntuaciones. La mayoría de los programas de cómputo aceptan solicitudes de puntuaciones de factores. Sólo en el caso del análisis de los componentes principales es posible calcular puntuaciones exactas. Además, en este análisis las puntuaciones no se correlacionan. En un análisis de factores comunes se obtienen estimaciones de estas puntuaciones y no hay seguridad de que los factores no se correlacionen unos con otros. Las puntuaciones de los factores pueden usarse en vez de las variables originales en el análisis multiviariado siguiente.

CAPÍTULO III

ANÁLISIS DE CONGLOMERADOS

3.1. Objetivo

Entender y aplicar los conceptos de análisis de conglomerados en situaciones del área administrativa.

3.2. Antecedentes

El análsis de conglomerados es una técnica del grupo estructural (también llamadas independientes), el cual tiene por objeto resumir información; en este análisis todas las variables son consideradas independientes, trabaja con variables medidas en escala de intervalo, de razón y nominales convertidas a *dummy*; para realizar este análisis las variables deberán ser todas en escala de intervalo y de razón o todas *dummy*, esto es, no se pueden manejar en un mismo análisis diferentes tipos de escalas (De la Garsa, 2013).

El análisis de conglomerados tiene como propósito formar grupos de objetos o individuos lo más homogéneos entre sí; es decir, al comparar los elementos que conforma un grupo, éstos deberán ser tan parecidos como sea posible, y lo más heterogéneos al compararse un grupo con otro.

El análisis de conglomerados (*cluster analysis*) tiene su inicio en la biología al clasificar diferentes especies en grupos homogéneos.

3.3. Áreas de aplicación

El análisis de conglomerados se puede utilizar en muchas áreas, por ejemplo, en la medicina se pueden agrupar padecimientos o síntomas y encontrar un remedio para un grupo de pacientes. En psiquiatría el diagnóstico correcto para enfermedades mentales como paranoia esquizofrenia. Esto es esencial para definir la terapia a aplicar (De la Garsa, 2013).

El análisis de conglomerados también es ampliamente utilizado en marketing para encontrar segmentos de mercado cuando estos no están claros y dirigir la estrategia comercial acertada o para encontrar segmentos de consumidores para un nuevo producto. En economía se podría utilizar para agrupar países dependiendo de sus recursos económicos o naturales, etc.

En recursos humanos, el análisis de conglomerados se podría utilizar para entender las diferentes motivaciones laborales que tienen los empleados de una empresa, entre otras aplicaciones.

3.4. Estadísticas del análisis por conglomerados

La mayoría de los métodos por conglomerados son heurísticos, basados en algoritmos. Así, el análisis por conglomerados contrasta notablemente con el análisis de varianza, regresión, discriminante y factorial, que se basan en conjuntos extensos de razonamientos estadísticos. Las siguientes son las estadísticas fundamentales que se relacionan con el análisis por conglomerados (Malhotra, 2004).

Centroide del conglomerado. El centroide del conglomerado lo componen los valores de las medias de las variables en todos los casos u objetos de ese conglomerado.
Gentros del conglomerado. Son los puntos iniciales de un conglomerado no jerárquico. Los conglomerados se forman alrededor de estos centros o semillas.
Dendograma. Un dendograma o *gráfica de árbol*, es un medio gráficos para desplegar los resultados de un conglomerado. Las líneas verticales representan conglomerados conjuntados. La posición de la línea en la escala indica las distancias a las que fueron unidos los conglomerados. El dendograma se lee de izquierda a derecha.
Distancia entre los centros de los conglomerados. Esta distancia indica qué tan separados están los pares individuales de los conglomerados. Los conglomerados que están ampliamente separados son distintos y, por tanto, deseables.
Diagrama de carámbanos. Es una representación gráfica de los resultados de un agrupamiento. Se llama así porque parece una fila de carámbanos que cuelgan de las salientes de una casa. Las columnas corresponden a los objetos agrupados y las filas al número de conglomerados. Los diagramas de carámbanos se leen de abajo hacia arriba.
Esquema de aglomeración. Ofrece información sobre los objetos o casos que se combinan en cada etapa del proceso de agrupamiento jerárquico.
Matriz de coeficientes de semejanza y distancia. Es una matriz de un triángulo inferior que contiene las distancias en pares entre objetos o casos.
Pertenencia a un conglomerado. Indica el conglomerado al que pertenece cada objeto o caso.

3.5. Pasos para realizar el análisis de conglomerados

Los pasos para realizar el análisis de conglomerados, se desarrollarán en paralelo con un ejemplo consistente en la actitud que tiene el consumidor frente a la compra.

3.5.1. Planteamiento del problema.

Lo más importante es elegir acuciosamente las variables en las que se basará el conglomerado. La inclusión de apenas una o dos variables irrelevantes puede distorsionar una solución de agrupamiento que de otra manera sería útil. Básicamente las variables a seleccionar deben describir la semejanza entre objetos, en términos de que sean pertinentes para la investigación (Malhotra, 2004). Se sugiere que las variables sean elegidas de acuerdo a investigaciones pasadas, o realizando una investigación exploratoria de la mano con el buen juicio e intuición del investigador.

El siguiente ejemplo consiste en identificar grupos de consumidores que tienen diferente actitud respecto a la compra. Se pidió a los consumidores que expresaran su grado de acuerdo con los siguientes enunciados en una escala de siete puntos (1= en desacuerdo, 7= de acuerdo).

DIVERTIDO: Ir de compras es divertido.
MALO: Ir de compras es malo para el presupuesto.
COMPRO: Cuando voy de compras como fuera.
OFERTAS: Trato de obtener las mejores ofertas cuando compro.
NMINTERESA: No me interesa ir de compras.
PRECIOS: Se ahorra mucho dinero cuando se comparan precios.

Los datos obtenidos en una prueba previa de 20 encuestados se resumen en la base de datos (tabla 3.1)

Tabla 3.1: Actitudes hacia las compras (*Malhotra, 2004*)

N°	DIVERTIDO	MALO	COMPRO	OFERTAS	NMINTERESA	PRECIOS
1	6	4	7	3	2	3
2	2	3	1	4	5	4
3	7	2	6	4	1	3
4	4	6	4	5	3	6
5	1	3	2	2	6	4
6	6	4	6	3	3	4
7	5	3	6	3	3	4
8	7	3	7	4	1	4
9	2	4	3	3	6	3
10	3	5	3	6	4	6
11	1	3	2	3	5	3
12	5	4	5	4	2	4
13	2	2	1	5	4	4
14	4	6	4	6	4	7
15	6	5	4	2	1	4
16	3	5	4	6	4	7
17	4	4	7	2	2	5
18	3	7	2	6	4	3
19	4	6	3	7	2	7
20	2	3	2	4	7	2

Fuente: Malhotra, 2004

3.5.2. Selección de una medida de distancia o semejanza.

Como el objetivo del conglomerado es reunir objetos semejantes, se necesita alguna medida para evaluar que tan similares o diferentes son (Malhotra, 2004).

La **proximidad** expresa el grado de similitud o diferencia que existe entre parejas de individuos, objetos y variables. Las proximidades son de dos tipos: (a) **disimilaridad** o desemejanza, cuando las proximidades se interpretan en términos de los diferente y (b) **similaridad** o semejanza, cuando las proximidades se interpretan por lo parecido (De la Garsa, 2013). El método más común consiste en medir la semejanza en términos de la distancia entre pares de objetos. Los objetos con menores distancias son más parecidos que los que tienen mayores distancias. Hay varias formas de calcular las distancias entre objetos.

La medida más frecuente de semejanza es la distancia euclidiana o su cuadrado. La **distancia euclidiana** es la raíz cuadrada de la suma de las diferencias cuadradas de los valores de cada variable.

Si las variables se miden en unidades muy diferentes, tendrán un efecto en la solución de conglomerado. En un estudio de las compras en un supermercado, quizá las variables de actitudes se midan en una escala de Likert de nueve puntos; la clientela, en términos de la frecuencia de visitas por mes y la cantidad gastada, y la lealtad a la marca, en términos del porcentaje de gasto en abarrotes destinado al supermercado preferido. En estos casos, antes de conglomerar a los encuestados, debemos uniformar los datos, lo que se hace cambiando las escalas de las variables para que tengan media de cero y desviación estándar de uno. Aunque la homogeneización suprime la influencia de la unidad de medida, también reduce las diferencias entre grupos en las variables que mejor distinguirían grupos o conglomerados. También es deseable eliminar valores extremos atípicos (Malhotra, 2004).

El uso de diferentes medidas de distancia puede llevar a resultados de conglomerados distintos. Por tanto, es aconsejable tomar diversas medidas y comparar los resultados. Luego de elegir una medida de distancia o semejanza se procede a escoger un procedimiento de conglomerado.

En la tabla 3.2, se presentan siete ecuaciones que corresponden a medidas de disimilaridad.

Tabla 3.2: Medida de disimilaridades más comunes

Nombre	Medida de similaridad o métrica
Ecuación 2-1 Distancia euclidiana	$d_{ij} = \left[\sum_{k=1}^{r}(X_{ik} - X_{jk})^2\right]^{1/2}$
Ecuación 2-2 Distancia euclidiana al cuadrado	$d_{ij}^2 = \sum_{k=1}^{r}(X_{ik} - X_{jk})^2$
Ecuación 2-3 Distancia de Chebychev	$C_{ij} = Max\|X_{ik} - X_{jk}\|$
Ecuación 2-4 Distancia de Mahalanobis (solo matricialmente)	$d_{ij} = \left[(X_i - X_j)^T \sum_{k=1}^{r}{}^{-1}(X_i - X_j)^2\right]^{1/2}$ y cuando se compara con el centroide sería $d_{ij} = \left[(X_i - X_k)^T \sum_{k=1}^{r}{}^{-1}(X_i - X_k)^2\right]^{1/2}$
Ecuación 2-5 Distancia de Manhattan o *city block* métrica	$d_{ij} = \sum_{k=1}^{r}\|X_{ik} - X_{jk}\|$
Ecuación 2-6 Distancia de Minkowski métrica	$d_{ij} = \left[\sum_{k=1}^{r}\|X_{ik} - X_{jk}\|^{\lambda}\right]^{1/\lambda}$, $\lambda \geq 1$
Ecuación 2-7 Distancia en un poder métrico absoluto	$d_{ij} = \left[\sum_{k=1}^{r}\|X_{ik} - X_{jk}\|^{p}\right]^{1/r}$

Fuente: De la Garsa, 2013

La d_{ij} es la distancia que se encuentra entre el objeto i y el objeto j, cuando dicha distancia es cero, quiere decir que no hay diferencia entre esos objetos; una distancia muy pequeña significa que los objetos o las personas que se comparan son parecidas y una distancia grande significa que son muy diferentes. El X_{ik} es el valor que tiene el objeto i en la variable k y el X_{jk} es el valor que toma el objeto j en la variable k. Por ejemplo, $X_{48} - X_{58}$ es la evaluación que posee en la variable 8 los objetos o personas 4 y 5, mientras que r, que se utiliza en la fórmula de la distancia euclidiana, representa el número de variables. Al usar estas medidas, se obtendrán las distancias o disimilaridades.

Las medidas de similaridad métricas se dejarán para que el propio estudiante las consulte, como así también las medidas de similitud o semejanza que están diseñadas para el caso de variables binarias (De la Garsa, pp 403 - 407, 2013). Por lo demás, el ejemplo de la actitud hacia la compra, que se trabaja en este capítulo considera el uso de la distancia euclidiana al cuadrado, cuya fórmula fue dada en la tabla 3.2

3.5.3. Selección de un procedimiento de conglomerado.

La clasificación del procedimiento de conglomerado parte con los conglomerados **jerárquicos** y los **no jerárquicos**. Los conglomerados jerárquicos se clasifican **por aglomeración** y **por división**. El conglomerado por aglomeración comienza con cada objeto en un grupo separado. Los conglomerados se forman reuniendo objetos en grupos cada vez mayores. El proceso continúa hasta que todos los objetos son miembros de un único conglomerado. El **conglomerado por división** comienza con todos los objetos reunidos en un solo conglomerado, que se divide hasta que cada objeto está en un grupo separado (Malhotra, 2004).

Los métodos por aglomeración son los que más se usan en la investigación de mercados. Comprenden los métodos de enlace, métodos de sumas de errores cuadrados o varianza y métodos de centroides. Los métodos de enlace abarcan los de enlace único, enlace completo y enlace promedio. El método de **enlace único** se basa en la regla de la mínima distancia o del vecino más cercano. Los primeros dos objetos agrupados son los que tienen la menor distancia entre ellos. Se identifica la siguiente distancia menor y o bien se conglomera un tercer objeto con los primeros dos o se forma un nuevo conglomerado de dos objetos (Malhotra, 2004). El método de enlace simple no funciona bien si los conglomerados están mal definidos. El método del **enlace completo** es semejante al de enlace único, salvo que se basa en la regla de la distancia máxima o el vecino más alejado. En el enlace completo, la distancia entre dos conglomerados se calcula como la distancia entre sus dos puntos más lejanos. El método del **enlace promedio** opera de la misma manera, pero la distancia entre dos conglomerados se define como el promedio de la distancia entre todos los pares de objetos, tomando cada elemento de los pares de un conglomerado.

El **método de Ward**, también llamado *método de la varianza mínima*, busca a los dos grupos o conglomerados cuya unión conlleve al menor incremento de la varianza. Esto significa que en cada paso se debe probar con todas las combinaciones posibles de dos grupos, calcular el valor del índice de la suma de cuadrados y seleccionar aquel con el menor valor. La desventaja es que tiende a formar grupos compactos y del mismo tamaño, utiliza más información sobre el contenido de los grupos que otros métodos, pero es el que ha demostrado mayor eficacia en estudios de simulación.

Por último, dentro de los métodos jerárquicos y específicamente de aglomeración, se tiene el **método del centroide** el cual considera que, al unirse dos elementos y formar un grupo, las características que prevalecerán con respecto a un tercer elemento estarán dadas por el promedio de las que originalmente poseían.

Para comprender en mayor profundidad cada uno de los métodos de conglomerados jerárquicos ya definidos, se recomienda revisar texto de *Análisis Estadístico Multivariante* (De la Garsa, 2013).

Aunque el método de agrupación no jerárquica es bastante razonable, tiene tres desventajas importantes.

La primera es que el procedimiento exige que, en principio, se infiera el número de agrupamientos que van a existir. En segundo lugar es que la elección arbitraria de las simientes (puntos semilla o que se seleccionan inicialmente para el agrupamiento) influye mucho sobre el procedimiento. Por último, con mucha frecuencia, el procedimiento no es factible desde el punto de vista del cálculo, porque hay precisamente demasiadas elecciones posibles, no solo para el número de agrupamientos, sino también para las ubicaciones de las simientes.

3.5.4. Un ejemplo de aplicación.

Para ilustrar el conglomerado jerárquico, se tomará el procedimiento de Ward (varianza mínima). En las tablas 3.3 y 3.4 se dan resultados de conglomerar los datos de la tabla 3.1. En el esquema de aglomeración (tabla 3.3) hay información útil que muestra el número de casos o conglomerados que se combinan en cada etapa. La primera línea representa la etapa 1, con 19 conglomerados. Aquí se combinan los entrevistados 14 y 16, como se observa en la columna titulada “ Clúster combinado ”. La distancia euclidiana cuadrada entre estos dos encuestados se da en la columna “ Coeficientes ”. En la columna “ Primera aparición del clúster de etapa ” se indica la etapa en que se forma un conglomerado. Para ilustrarlo, una entrada de 1 en la etapa 6 indica que el encuestado 14 se agrupó por primera vez en la etapa 1. La última columna, Etapa “ siguiente ”, señala la etapa en que se combina con éste otro caso (encuestado) o conglomerado. Como la cifra de la primera línea de la última columna es 6, vemos que en la etapa 6 el encuestado 10 se combina con 14 y 16 para formar un conglomerado único. Del mismo modo, la segunda línea representa la etapa 2 con 18 conglomerados. En la etapa 2 se agrupan los encuestados 6 y 7.

Tabla 3.3: Esquema de conglomeración.

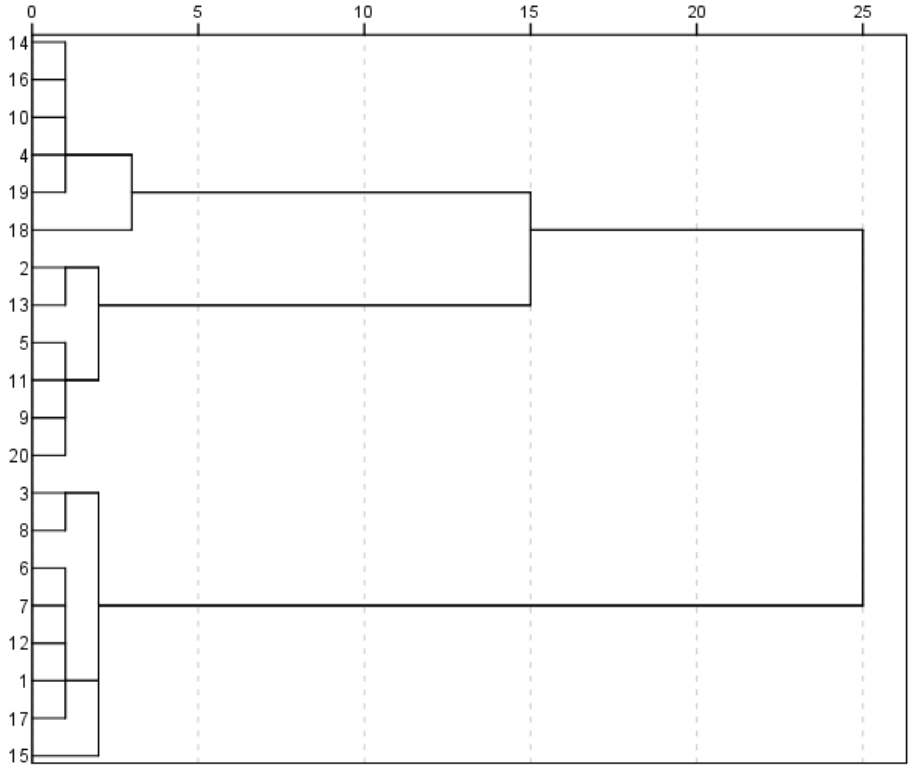
Etapa	Clúster combinado		Coeficientes	Primera aparición del clúster de etapa		Etapa siguiente
	Clúster 1	Clúster 2		Clúster 1	Clúster 2	
1	14	16	1,000	0	0	6
2	6	7	2,000	0	0	7
3	2	13	3,500	0	0	15
4	5	11	5,000	0	0	11
5	3	8	6,500	0	0	16
6	10	14	8,167	0	1	9
7	6	12	10,500	2	0	10
8	9	20	13,000	0	0	11
9	4	10	15,583	0	6	12
10	1	6	18,500	0	7	13
11	5	9	23,000	4	8	15
12	4	19	27,750	9	0	17
13	1	17	33,100	10	0	14
14	1	15	41,333	13	0	16
15	2	5	51,833	3	11	18
16	1	3	64,500	14	5	19
17	4	18	79,667	12	0	18
18	2	4	172,667	15	17	19
19	1	2	328,600	16	18	0

Otro medio gráfico útil para desplegar información del conglomerado es el dendograma (figura 3.1). El dendograma se lee de izquierda a derecha. El eje horizontal del dendograma corresponde a las distancias en la que se agrupa una determinada cantidad de clúster. El eje vertical señala el número del caso (entrevistado o unidad observable). Las líneas verticales segmentadas representan conglomerados reunidos. La posición de una línea en la escala indica las distancias a las que se unieron los conglomerados; por ejemplo, el número 20 ubicado en la horizontal superior significa que la unión de los dos conglomerados con mayor cantidad de entrevistados se logró a una distancia promedio de 20 unidades. Al trazar una vertical imaginaria, ésta va a cortar una cantidad de rectas horizontales que dependerá donde se trace. A modo de ejemplo, si esta vertical se dibuja en el punto medio de los valores 10 y 15 de la escala, se verá que corta a tres líneas horizontales, es decir, a una distancia promedio de 12,5 unidades se obtienen 3 conglomerados. Esta información es útil para decidir sobre el número de conglomerados.

Analizar el árbol de clasificación o dendograma para determinar el número de grupos es un proceso subjetivo. En general, se comienza por buscar "huecos" entre uniones a lo largo del eje horizontal. De derecha a izquierda hay un hueco entre 20 y 25, que divide a los individuos en dos grupos: Un grupo está formado por los individuos (14), (16), (10), (4), (19), (18), (2), (13), (5), (11), (9) y (20), y el otro grupo está formado por los individuos (3), (8), (6), (7), (12), (1), (17) y (15)

Hay otro hueco entre 5 y 10 que sugiere tres clústeres: (14, 16, 10, 4, 19); (2, 13, 5, 11, 9, 20) y (3, 8, 6, 7, 12, 1, 17, 15).

Figura 3.1: Dendograma que utiliza una vinculación de Ward.



También es posible obtener información sobre el conglomerado al que pertenecen los casos si se especifica el número de conglomerados. En la tabla 3.4 se detallan a qué conglometrados pertenecen los casos, dependiendo de que la solución final contenga, dos, tres o cuatro conglomerados. Esta información puede obtenerse para cualquier número de conglomerados y es útil para elegir el número de conglomerados. En el caso estudiado, se consideró un rango de soluciones entre dos y cuatro conglomerados. Supóngase que el analista tiene la sospecha de que existen tres grupos con características diferentes (entre grupos), pero similares al interior de cada grupo; entonces casos 1, 3, 6, 7, 12, 15 y 17 forman el clúster 1; los casos 2, 5, 9, 11, 13 y 20 forman el clúster 2 y el resto de los casos forman el clúster 3.

Tabla 3.4: Clúster de pertenencia.

Caso	4 clústeres	3 clústeres	2 clústeres
1	1	1	1
2	2	2	2
3	1	1	1
4	3	3	2
5	2	2	2
6	1	1	1
7	1	1	1
8	1	1	1
9	2	2	2
10	3	3	2
11	2	2	2
12	1	1	1
13	2	2	2
14	3	3	2
15	1	1	1
16	3	3	2
17	1	1	1
18	4	3	2
19	3	3	2
20	2	2	2

3.5.5. Interpretación y descripción de los conglomerados.

Interpretar y describir los conglomerados implica examinar sus centroides, los cuales representan los valores promedio de los objetos contenidos en el conglomerado en cada una de las variables. Los centroides nos permiten describir cada conglomerado al asignarle un nombre o etiqueta. Si el calendario de conglomeración no imprime esta información, puede obtenerse mediante el análisis discriminante (Malhotra, 2009). La tabla 3.5 proporciona los centroides o valores promedio de cada conglomerado de nuestro ejemplo. El conglomerado 3 tiene valores relativamente altos en las variables $DIVERT$ (ir de compras es divertido) y $COMPRO$ (cuando voy de compras aprovecho para comer fuera). También tiene un valor bajo en $NMINTERE$ (no me interesa ir de compras). De modo que al conglomerado 3 se le puede etiquetar como "compradores divertidos e interesados". Este conglomerado consta de los casos 1, 3, 6, 7, 8, 12, 15 y 18.

Tabla 3.5: Centros de clústeres finales.

	Clúster		
	1	2	3
DIVERT	3,50	1,67	5,75
MALO	5,83	3,00	3,63
COMPRO	3,33	1,83	6,00
OFERTAS	6,00	3,50	3,13
NMINTERE	3,50	5,50	1,88
PRECIOS	6,00	3,33	3,88

El conglomerado 2 es justo el contrario, con valores bajos en $DIVERT$ y $COMPRO$, y valor alto en $NMINTERE$, por lo que este conglomerado puede etiquetarse "compradores apáticos". Los miembros del conglomerado 2 son los casos 2, 5, 9, 11, 13 y 20. El conglomerado 3 tiene valores altos en $MALO$ (las compras desequilibran mi presupuesto), $OFERTAS$ (trato de encontrar las mejores ofertas cuando voy de compras) y $PRECIOS$ (puede ahorrarse mucho dinero si se comparan precios). Por lo que este conglomerado puede etiquetarse como "compradores ahorrativos". El conglomerado 3 abarca los casos 4, 10, 14, 16, 18 y 19.

Tabla 3.6: Clúster de pertenencia.

Número del caso	Clúster	Distancia
1	3	1,414
2	2	1,323
3	3	2,550
4	1	1,404
5	2	1,848
6	3	1,225
7	3	1,500
8	3	2,121
9	2	1,756
10	1	1,143
11	2	1,041
12	3	1,581
13	2	2,598
14	1	1,404
15	3	2,828
16	1	1,624
17	3	2,598
18	1	3,555
19	1	2,154
20	2	2,102

La tabla 3.6 muestra los resultados finales de la pertenencia de cada caso a un conglomerado. Esta tabla también puede obtenerse del programa SPSS.

CAPÍTULO IV

ANÁLISIS DISCRIMINANTE

4.1. Objetivo

En primer lugar, el análisis discriminante permite determinar cuáles son las variables (de entre la serie de variables seleccionadas previamente por el investigador), que mejor explican la pertenencia de un individuo a un grupo determinado (Pedret, 2000).

En segundo lugar el análisis discriminante o de clasificación es producir una regla o un esquema de clasificación que permita a un investigador predecir la población de la que es lo más probable que tenga que venir una observación (Johnson, 2000).

4.2. Antecedentes

El análisis discriminante pertenece al grupo de técnicas funcionales y, por lo tanto, su uso principal es para hacer pronósticos. Esta técnica utiliza "m" variables independientes, todas ellas métricas y una sola variable dependiente no métrica, nominal (categórica) (De la Garsa, 2013).

La historia del análisis discriminante se inicia en 1920 con los trabajos del estadístico inglés Karl Pearson, pero en 1930, con el estadístico R. A. Fisher, que se propone una metodología para obtener la combinación lineal de variables (la ecuación discriminante de Fisher), que hasta la fecha se utiliza como parte del proceso de análisis. Metodólogos de la universidad de Harvard realizaron aplicaciones del análisis discriminante en el área de la educación y en psicología en las décadas de los 50 y 60.

En las primeras décadas del siglo XX se usa el discriminante *descriptivo*. A partir de la década del 60 en adelante se utiliza el análisis discriminante *predictivo*.

4.3. Áreas de aplicación

El problema de discriminación aparece en muchas situaciones en que necesitamos clasificar elementos con información incompleta. Por ejemplo, los sistemas automáticos de concesión de créditos (credit scoring) implantados en muchas

instituciones financieras tienen que utilizar variables medibles hoy (ingresos, antigüedad en el trabajo, patrimonio, etc.) para prever el comportamiento futuro. En otros casos la información podría estar disponible, pero puede requerir destruir el elemento, como en el control de calidad de la resistencia a la tensión de unos componentes (Peña, 2002). Otros ejemplos de aplicación se presentan en áreas como la psicología en donde ayuda a evaluar si una persona sufre algún trastorno según sus patologías o perfil mental. En recursos humanos sirve para realizar un filtro previo a una entrevista; en marketing para decidir sobre qué tipo de producto (bien o servicio) ofrecerle a un cliente o bien estudiar la eficacia de alguna campaña publicitaria. En antropología ayuda para saber la antigüedad, el tipo, la raza o género de un fósil. En política, para realizar estudios anteriores a las elecciones con el fin de determinar la eficacia de las campañas políticas o para buscar por qué las personas tienen mayor preferencia por algún partido. En sociología para determinar la razón del comportamiento humano. En seguros para predecir el tipo de cobertura que se le podría ofrecer a un cliente o bien para pronosticar el tipo de siniestros, entre muchas otras aplicaciones (De la Garsa, 2013).

4.4. Modelo de análisis discriminante

El discriminante genera una ecuación que permitirá realizar pronósticos y que determinará la existencia de grupos en una población de interés. Dicha ecuación está representada por la función discriminante lineal de Fisher, la cual tiene la siguiente expresión:

$$Z = K_1X_1 + K_2X_2 + \ldots + K_mX_m$$

donde

$Z =$: representa el puntaje discriminante o puntaje discriminatorio (número que servirá para efectuar la discriminación).
K_i: es el factor de ponderación o importancia que tiene la variable X_i para discriminar.
X_i: es la variable independiente *i-ésima*; $i = 1, 2, ..., m$.

Los coeficientes o pesos K_i se estiman de modo se estiman de modo que los grupos difieran cuanto sea posible en los valores de la función discriminante (Malhotra, 2004).

4.5. Estadísticas del análisis discriminante

Las principales estadísticas del análisis discriminante son (Malhotra, 2004):

Correlación canónica. La correlación canónica mide el grado de asociación entre las puntuaciones de discriminación y los grupos. Es una medida de asociación entre la única función discriminante y el conjunto de variables ficticias que definen la pertenencia al grupo.
Centroide. el centroide es la media de las calificaciones discriminantes de un grupo particular. Existen tantos centroides como grupos, porque hay uno para cada grupo. Los centroides del grupo son las medias de ese grupo en todas las funciones.
Matriz de clasificación. llamada a veces también *matriz de confusión* o *de predicción*, contiene el número de casos cuya clasificación fue correcta e incorrecta. Los casos bien clasificados aparecen en la diagonal porque los grupos reales y los pronosticados son los mismos. Los elementos fuera de la diagonal representan casos cuya clasificación fue incorrecta. La suma de los elementos de la diagonal, dividida entre el número total de casos, representa la proporción de aciertos.
Coeficientes de la función discriminante. Los coeficientes de la función discriminante (sin estandarizar) son los multiplicadores de las variables cuando éstas se encuentran en las unidades originales de medición.
Puntuaciones de discriminación. Los coeficientes sin estandarizar se multiplican por los valores de las variables. Los productos se suman al término constante para obtener las puntuaciones de discriminación.
Valor propio. Para cada función discriminante, el valor propio es la proporción de la suma de cuadrados entre grupos y dentro de los grupos. Valores propios grandes significan funciones superiores.
Valores F y su significancia. Los valores F se calculan en una ANOVA de un factor en el que la variable de agrupamiento es la variable independiente categórica. A su vez, cada variable de pronóstico opera como variable métrica dependiente en ese análisis.
Medias de los grupos y desviaciones estándar de los grupos. Estos valores se calculan para cada variable de pronóstico de cada grupo.
Matriz de correlación común del grupo. La matriz de correlación común del grupo se calcula promediando las matrices de covarianza separadas de todos los grupos.
Coeficientes estandarizados de la función discriminante. Los coeficientes estandarizados de la función discriminante son los coeficientes de la función discriminante y se toman como multiplicadores cuando las variables fueron estandarizadas en una media de cero y una varianza de uno.
Correlaciones estructurales. También se llaman *cargas discriminatorias*. Las correlaciones estructurales representan las correlaciones simples entre las variables de pronóstico y la función discriminante.

Matriz de correlación total. Si los casos se tratan como si procedieran de una sola muestra y se calculan las correlaciones, se obtiene una matriz de correlación total.
λ de Wilks. También se le llama *estadística U*. La λ de Wilks de cada variable de pronóstico es la proporción de la suma de cuadrados dentro del grupo a la suma de cuadrados total. Su valor varía entre 0 y 1. Valores de λ altos (cercanos a 1) indican que las medias del grupo no parecen diferentes. Valores de λ bajos (cercanos a 0) indican que las medias del grupo parecen diferentes.

4.6. Requerimientos para llevar a cabo un análisis discriminante

Los grupos deben formarse a priori, en forma natural o por una técnica anterior (De la Garsa,2013).

- A los grupos los define una serie de características o atributos, pues no están formados al azar; éstas pueden ser conocidas o desconocidas.
- Las zonas que delimitan a los grupos pueden estar bien definidas o pueden existir zonas de confusión, en este caso no se sabe a qué grupo pertenecen ciertos individuos u objetos.
- Ninguna variable discriminante puede ser combinación lineal de otras variables discriminantes (no debe existir multicolinealidad entre ellas).
- El número máximo de funciones discriminantes que se pueden calcular podrá ser igual al número de grupos menos 1.
- Las matrices de varianzas - covarianzas de cada grupo deben ser aproximadamente iguales, es decir, constantes.
- Supuesto de normalidad en las variables independientes.
- Deberá existir variabilidad o dispersión dentro de cada grupo.
- Se asume linealidad, por lo que no debe darse transformación de variables, ni exponenciales.
- Al llevar a cabo una clasificación se supone a priori que el resultado obtenido es tan confiable como lo indica el porcentaje de clasificaciones correctas obtenidas en la etapa de validación del modelo.

4.7. Etapas para realizar un análisis discriminante

Como ya se ha hecho en los capítulos anteriores, las etapas para realizar un análisis discriminante, se desarrollaran y complementaran en forma paralela con un problema en particular:

4.7.1. Formulación del problema o detección de áreas de oportunidad.

Para la elaboración de un A. D. es necesario establecer los objetivos, los cuales pueden ser (De la Garsa, 2013):

- Probar la segmentación que se hizo a priori o corroborar la existencia de grupos que una persona, de manera natural, supone que hay.
- ¿Qué variables, en realidad, están ayudando a la segmentación?
- Pronosticar en qué grupos se encuentran los nuevos individuos.

El siguiente problema servirá de ejemplo para aplicar las etapas del análisis discriminante:

En una prueba previa, se obtuvieron datos de 45 encuestados sobre Nike. Los datos se anotan en la tabla 4.1, en la que se indica *uso*, *sexo*, *conciencia*, *opinión*, *preferencia*, *intención* y *lealtad* a Nike de una muestra de usuarios de la marca. El uso se codificó como 1, 2 y 3 para representar usuarios frecuentes, moderados y esporádicos, respectivamente. El sexo se codificó con 1 para mujeres y 2 para hombres. La conciencia, opinión, preferencia, intención y lealtad se midieron con una escala de siete puntos tipo Likert (1 = muy poco favorable, 7 = muy favorable). Observe que hay cinco encuestados que tienen respuestas faltantes que se denotan con 9.

4.7.2. Selección de las variables independientes y dependientes.

Deben existir por lo menos dos grupos en la variable dependiente. Es muy importante que se haga una lista exhaustiva de todos los grupos posibles de la variable dependiente. Los grupos que forman a dicha variable deben ser mutuamente excluyentes, y si se desea un estudio muy relevante, deberán ser colectivamente exhaustivos.

Tabla 4.1: Base de datos para reconocer tipo de usuario

N°	Uso	Sexo	Conciencia	Opinión	Preferencia	Intención	Lealtad
1	3	2	7	6	5	5	6
2	1	1	2	2	4	6	5
3	1	1	3	3	6	7	6
4	3	2	6	5	5	3	2
5	3	2	5	4	7	4	3
6	2	2	4	3	5	2	3
7	2	1	5	4	4	3	2
8	1	1	2	1	3	4	5
9	2	2	4	4	3	6	5
10	1	1	3	1	2	4	5
11	3	2	6	7	6	4	5
12	3	2	6	5	6	4	4
13	1	1	4	3	3	1	1
14	3	2	6	4	5	3	2
15	1	2	4	3	4	5	6
16	1	2	3	4	2	4	2
17	3	1	7	6	4	5	3
18	2	1	6	5	4	3	2
19	1	1	1	1	3	4	5
20	3	1	5	7	4	1	2
21	3	2	6	6	7	7	5
22	2	2	2	3	1	4	2
23	1	1	1	1	3	2	2
24	3	1	6	7	6	7	6
25	1	2	3	2	2	1	1
26	2	2	5	3	4	4	5
27	3	2	7	6	6	5	7
28	2	1	6	4	2	5	6
29	1	1	9	2	3	1	3
30	2	2	5	9	4	6	5
31	1	2	1	2	9	3	2
32	1	2	4	6	5	9	3
33	2	1	3	4	3	2	9
34	2	1	4	6	5	7	6
35	3	1	5	7	7	3	3
36	3	1	6	5	7	3	4
37	3	2	6	7	5	3	4
38	3	2	5	6	4	3	2
39	3	2	7	7	6	3	4
40	1	1	4	3	4	6	5
41	1	1	2	3	4	5	6
42	1	1	1	3	2	3	4
43	1	1	2	4	3	6	7
44	1	1	3	3	4	6	5
45	1	1	1	1	4	5	3

Fuente: Malhotra, 2004

Si los grupos fueron formados a *priori* se recomienda que en el análisis se incluyan las variables que pueden discriminar o explicar los segmentos. Si los grupos se integraron de forma natural, se deberá tener una idea de qué atributos o características segmentan o discriminan. Por último, si el investigador puede manejar la información, se recomienda reunir a las variables y a los elementos más representativos de cada grupo (De la Garsa, 2013).

En el ejemplo expuesto, se realizarán por separado dos tipos de análisis discriminante (simple y múltiple), lo que lleva a definir en un primer caso dos variables dependientes y en un segundo caso tres variables dependientes.

4.7.3. Consideraciones sobre el tamaño de la muestra.

Un ejemplo: Si la población está dividida en un 55 % de individuos con cierta característica A, y un 45 % posee una característica contraria B, entonces al tomar una muestra de 20 observaciones, se requiere que 11 (55 %) posea la característica A.

Con respecto a la validación de la muestra, se debe tener en cuenta que ésta deberá separarse en 2 y para la etapa de derivación se utilizarán 60 ó 70 % de los datos, y el restante se dejará para la etapa de validación del modelo, ya que no sería valido el grado de error encontrado al pronosticar con los mismos datos con que se derivó la ecuación.

4.7.4. Tipos de análisis discriminante.

Son dos los tipos de análisis discriminante: simple y múltiple. El primero corresponde cuando la variable dependiente está formada por solo dos grupos; en el ejemplo expuesto en este capítulo, un grupo serían las mujeres y el otro los hombres, si el objetivo es discriminar con base al sexo. Para el caso del análisis discriminante múltiple, primero se debe tener en cuenta la cantidad de grupos que se desea formar (tres o más grupos). Para el ejemplo expuesto, se realizará discriminante múltiple con la variable "Uso", pues considera las categorías de usuarios: frecuentes, moderados y esporádicos.

Se puede destacar además que el número de variables independientes no tiene nada que ver con el hecho de que se considere simple o múltiple el modelo discriminante.

Como ya se dijo, la ecuación discriminante se presenta mediante la siguiente expresión:

$$Z = K_1X_1 + K_2X_2 + \cdots + K_mX_m$$

En la derivación del modelo se pretende encontrar los valores de los pesos que tienen las variables para discriminar, es decir, los valores de las constantes (K). Se debe recordar que se tiene como criterio maximizar la relación entre las diferencias de los grupos con respecto a la variabilidad en los datos.

Dado un conjunto de n observaciones, podemos compactar en forma matricial:

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1k} \\ X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pk} \end{pmatrix} \begin{pmatrix} K_1 \\ K_2 \\ \vdots \\ K_k \end{pmatrix}$$

Donde:

Z_i: puntaje discriminante en la observación i (número que servirá para efectuar la discriminación.
K_j: factor de ponderación o de importancia que se tiene para la variable j que sirve para discriminar.
X_{ij}: variable j en la observación i.

En notación matricial más compacta, $\mathbf{Z} = \mathbf{XK}$, donde:

Z: vector de n observaciones a clasificar.
K: vector de todos los factores de ponderación.
X: matriz de todas las variables discriminadoras.

El criterio de la función discriminante mencionada es:

$$Maximizar = \frac{Variabilidad\ entre\ grupos}{Variabilidad\ dentro\ de\ los\ grupos}$$

El resultado de esta división se interpreta como el valor propio o *eigenvalor* (λ), el cual se interpreta como la cantidad de variación explicada por la ecuación. El objetivo es la maximización del valor propio, por lo tanto, el manejo de la mayor información o variación posible a través del modelo.

Si la varianza entre los grupos es grande, pero la varianza dentro de cada grupo es pequeña, entonces se dice que la función discriminante separa bien los grupos, esto es, internamente los individuos son muy homogéneos y a la vez muy diferentes entre sí (entre grupos).

Para obtener los coeficientes del vector **K** de la ecuación discriminante, dado que el objetivo implica una maximización de una relación, se deberá usar un proceso de derivación, solamente que los términos involucrados son expresiones matriciales.

Para un mejor entendimiento de la mecánica a seguir en el cálculo de las constantes, se hará uso del ejemplo de este capítulo.

En la etapa de derivación, que busca la obtención y comprobación del modelo utilizando para discriminar, lo primero que se hace es partir la muestra, esto es, tomar 60 ó 70 % de los datos. Considérese 70 %. Por tanto, $0,7 * 22,5 = 15,75 \approx 16$; es decir se escogerán 16 de cada grupo para la muestra de diseño y el resto (13 observaciones) queda para la muestra de validación. La tabla 4.2 contempla esta muestra aleatoria.

4.7.5. Prueba de igualdad de medias.

Con los datos de la tabla 4.2, se realizará la prueba de variables para asegurar que se seleccionaron las mejores para discriminar. Procedimiento que se realiza con una prueba de hipótesis.

H_0: La variable X_i no es buena para discriminar.
H_a: La variable X_i es buena para discriminar.

g : Es el número de grupos en el análisis discriminante, $g - 1$ son los grados de libertad del numerador.
n : Es el número de datos de las variables, $n - g$ son los grados de libertad del denominador.

Para realizar la prueba de hipótesis, la F_c se compara con la F de tabla, la cual tiene la nomenclatura $F_{\alpha,\, g-1,\, n-g}$.

Lo anterior está hecho para realizarse en forma global, pero cuando se hacen las pruebas para cada variable es necesario sacar la información de cada una de las matrices, para formar la lambda de Wilks de cada una de ellas. Esto es:

$$\Lambda_i = \frac{Variación\ no\ explicada\ de\ X_i}{Variación\ total\ de\ X_i} = \frac{|W_{X_i}|}{|T_{X_i}|}$$

Tabla 4.2: Base de datos para la etapa de diseño.

N°	Uso	Sexo	Conciencia	Opinión	Preferencia	Intención	Lealtad
2	1	1	2	2	4	6	5
3	1	1	3	3	6	7	6
7	2	1	5	4	4	3	2
8	1	1	2	1	3	4	5
10	1	1	3	1	2	4	5
13	1	1	4	3	3	1	1
17	1	1	3	1	2	3	1
18	1	1	2	2	3	1	1
19	1	1	1	1	3	4	5
20	3	1	5	7	4	1	2
23	1	1	1	1	3	2	2
24	3	1	6	7	6	7	6
28	2	1	6	4	2	5	6
29	1	1	9	2	3	1	3
33	2	1	3	4	3	2	9
34	2	1	4	6	5	7	6
1	3	2	7	6	5	5	6
4	3	2	6	5	5	3	2
5	3	2	5	4	7	4	3
6	2	2	4	3	5	2	3
9	2	2	4	4	3	6	5
11	3	2	6	7	6	4	5
12	3	2	6	5	6	4	4
14	3	2	6	4	5	3	2
15	3	2	6	2	4	4	3
16	3	2	6	2	2	3	1
21	3	2	6	6	7	7	5
22	2	2	2	3	1	4	2
25	1	2	3	2	2	1	1
26	2	2	5	3	4	4	5
27	3	2	7	6	6	5	7
30	2	2	5	9	4	6	5

Fuente: Malhotra, 2004

Esta prueba se contrasta con la estadística de Fisher, de tal manera que:

$$F_c = \left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{n-g}{g-1}\right)$$

O lo que es lo mismo:

$$F_c = \frac{(1-\Lambda)/(g-1)}{\Lambda/(n-g)}$$

en donde:

Λ : Lambda de Wilks, y $\Lambda = \frac{|W|}{|T|}$
W y T : Son matrices.

4.7.6. Prueba del modelo.

Más adelante se verá que el programa SPSS permite encontrar estos resultados de manera automática. Por el momento se asumirá que todas las variables explicativas son buenas para discriminar.

Ahora, el siguiente paso es obtener los valores propios propios λ (o eigen values), lo cual implica resolver el siguiente determinante:

$$|W^{-1}A - \lambda I| = 0$$

en donde:

W^{-1}: indica la inversa de la matriz W que es la matriz de variación no explicada.
A : es la matriz de variación explicada.
λ : son los valores propios.
I : es la matriz identidad.

El valor propio es sinónimo de cantidad de variación o información manejada, entonces hay que saber si ésta es relevante o significativa. Se utiliza principalmente para evaluar la importancia de cada una de las funciones discriminantes. Aunque un autovalor tiene un mínimo de cero, no tiene un máximo, lo cual lo hace difícil de interpretar por sí solo, pero el valor se utiliza para probar el modelo, como se explicará más adelante.

Existen dos métodos para probar si el modelo es bueno para discriminar o no, uno de ellos es el que maneja SPSS que a continuación se explicará y la otra prueba es a través del estadístico de Mahalanobis.

Para realizar esta prueba se recurrirá al estadístico V de Bartlett, que es una función de la Lambda de Wilks y tiene una distribución asintótica ji cuadrada, la fórmula es la siguiente:

$$V = \left\{ n - 1 - \frac{k+g}{2} \right\} ln(1+\lambda)$$

donde:

n : es el número de observaciones.
k : es en número de variables independientes o discriminadoras.

Las hipótesis para la prueba del modelo son:

H_0: El modelo **no** es bueno para discriminar.
H_a : El modelo **si** es bueno para discriminar.

El estadístico V de Bartlett sigue una distribución ji cuadrada, con los grados de libertad que se indican:

$$\chi^2_{(k-j)(g-j-1)}$$

donde:

k : número de variables independientes o discriminadoras.
g : número de grupos en la variable dependiente.
j : número correspondiente a cada una de las funciones discriminantes generadas.

Es importante recalcar que el número de ecuaciones discriminantes que se generan tienen que ver con el **mínimo** de:

- el número de grupos menos 1.
- el número de variables.

Como las hipótesis de prueba son:

H_0: el modelo no es bueno para discriminar.
H_a: el modelo si es bueno para discriminar.

y para rechazar H_0 es necesario que $\chi^2_c > \chi^2_{(k-j)(g-j-1)}$ como es lo que ocurre en este caso, entonces concluimos que con probabilidad $0,95$ el modelo si es bueno para discriminar.

4.7.7. Indicadores del modelo.

Los indicadores que reporta SPSS con respecto al modelo son los siguientes:

Valores propios

$\lambda = 1,096$

El valor propio es sinónimo de la cantidad de variación o información manejada, por lo que entre mayor sea su valor es mejor; si se tienen varios valores propios se escogerá el mayor.

Lambda de Wilks

$\Lambda = 0,477$

Los valores de Λ varían entre 0 y 1. Los próximos a cero indican mucha discriminación, es decir, los grupos están muy separados, mientras que los cercanos a 1 representan escasa discriminación o poca diferencia entre grupos. En el ejemplo $\Lambda = 0,477$ muestra que los grupos están separados; pero esa separación no es tan evidente y por tanto, a pesar de que hay discriminación, ésta no es tan alta.

Clorrelación canónica

Correlación canónica $= \sqrt{1 - \Lambda} = \sqrt{1 - 0,477} = 0,723$

La correlación canónica varía entre 0 y 1. Un valor cercano a 0 indica que no se puede explicar la existencia de los 2 grupos por medios de las variables discriminantes escogidas; un valor cercano a 1 indica una fuerte relación entre los grupos de pertenencia y los puntajes encontrados por la función discriminante.

Como en este caso el valor es cercano a 1, entonces existe una relación aceptable entre el grupo de pertenencia y los valores de la función discriminante (puntajes discriminantes).

Para encontrar los coeficientes de la ecuación que nos permita discriminar haremos uso del programa SPSS, siguiendo los siguientes comandos:

1. Elija ANALIZE de la barra menú de SPSS.
2. Haga clic en CLASSIFY y luego en DISCRIMINAT.
3. Cambie " Sexo " al recuadro GROUPING VARIABLE.
4. Haga clic en DEFINE RANGE. Introduzca un 1 en MINIMUN y 2 en MAXIMUN. Haga clic en CONTINUE.
5. Lleve las variables explicativas al recuadro INDEPENDENTS.
6. Elija ENTER INDEPENDENTS TOGETHER (opción predeterminada)
7. Haga clic en STATISTICS. En el recuadro DESCRIPTIVES de la ventana emergente, elija MEANS y UNIVARIATE ANOVAS. En el recuadro MATRICES elija WITHIN - GROUP CORRELATIONS y en el recuadro COEFICIENTES DE LA FUNCIÓN elija FISHER y NO ESTANDARIZADOS. Haga clic en CONTINUE.

8. Haga clic en CLASIFFY ... en la ventana emergente en la opción PRIOR PROBABILITIES, elija ALL GROUPS EQUAL (opción predeterminada). En el recuadro DISPLAY elija SUMMARY TABLE y LEAVE - ONE - OUT CLASSIFICATION. En el recuadro USE COVARIANCE MATRIX elija WITHING - GROUPS. Haga clic en CONTINUE.

9. Seleccione OK.

A continuación se muestra parte de la salida de SPSS al procesar la base de datos de diseño del ejemplo.

Tabla 4.3: Estadísticos de grupo.

SEXO		Media	Desviación estándar	N válido (por lista)	
				No ponderados	Ponderados
1	CONCIENCIA	4,19	2,536	16	16,000
	OPINIÓN	2,81	1,974	16	16,000
	PREFERENCIA	3,31	1,250	16	16,000
	INTENCIÓN	3,44	2,159	16	16,000
	LEALTAD	3,69	1,991	16	16,000
	USO	1,44	,727	16	16,000
2	CONCIENCIA	5,38	1,258	16	16,000
	OPINIÓN	4,50	1,932	16	16,000
	PREFERENCIA	4,75	1,693	16	16,000
	INTENCIÓN	4,25	1,291	16	16,000
	LEALTAD	3,88	1,668	16	16,000
	USO	2,63	,500	16	16,000
Total	CONCIENCIA	4,78	2,059	32	32,000
	OPINIÓN	3,66	2,104	32	32,000
	PREFERENCIA	4,03	1,636	32	32,000
	INTENCIÓN	3,84	1,798	32	32,000
	LEALTAD	3,78	1,809	32	32,000
	USO	2,03	,861	32	32,000

En la tabla 4.3 se presentan los resultados de correr en los datos de la tabla 4.2 un análisis discriminante de dos grupos mediante el uso SPSS. El examen de las medias y desviaciones estándar del grupo brinda una idea intuitiva de los resultados. Parece que los dos grupos están más separados en términos de la " opinión " y el " uso " que de otras variables. La " lealtad " es una variable que no separa los grupos. La diferencia entre los dos grupos respecto de la " conciencia" es pequeña, sin embargo la desviación estándar de esta variable es más grande en el grupo 2 (varones); lo que nos habla de una mayor dispersión en la respuesta de esta variable, por parte del grupo 2.

Un análisis que confirma lo anterior se obtiene, observando los resultados de la tabla 4.4 siguiente.

Tabla 4.4: Pruebas de igualdad de las medias de los grupos.

	Lambda de Wilks	F	df1	df2	Sig.
CONCIENCIA	,914	2,816	1	30	,104
OPINIÓN	,834	5,972	1	30	,021
PREFERENCIA	,801	7,465	1	30	,010
INTENCIÓN	,947	1,669	1	30	,206
LEALTAD	,997	,083	1	30	,775
USO	,509	28,957	1	30	,000

La prueba de igualdad nos permite determinar cuál de las variables discrimina más. En este caso se ve que las variables " opinión ", " preferencia " y " uso " son significativas en la discriminación (valor - p <0,05 las tres); sin embargo la variable " uso ", por tener un mayor valor F, discrimina más que las otras dos.

Un diagnóstico adicional acerca de la capacidad discriminadora que tienen las variables elejidas, lo entrega la siguiente matriz de correlaciones.

Tabla 4.5: Matriz intra - grupos combinados

		CONCIENCIA	OPINIÓN	PREFERENCIA	INTENCIÓN	LEALTAD	USO
Correlación	CONCIENCIA	1,000	,371	,197	-,129	,133	,452
	OPINIÓN	,371	1,000	,435	,320	,354	,610
	PREFERENCIA	,197	,435	1,000	,287	,318	,370
	INTENCIÓN	-,129	,320	,287	1,000	,772	,133
	LEALTAD	,133	,354	,318	,772	1,000	,100
	USO	,452	,610	,370	,133	,100	1,000

Esta matriz intra - grupos entrega información respecto de las correlaciones entre las variables. Se observa que variables como " opinión " y " uso " tienen una correlación por sobre 0,5; algo similar ocurre con el par de variables " lealtad " e " intensión " (correlación 0,722), lo que no es bueno, pues para llevar a cabo el análisis discriminante es contar con variables que no estén correlacionadas. Sin ambargo, en la gran mayoría de los casos de esta matriz esto no ocurre.

Tabla 4.6: Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	1,096[a]	100,0	100,0	,723

a. Se ha empleado solo una función discriminate canónica en el análisis

Los autovalores explican la cantidad de información de cada función discriminante (en este caso como se tiene solo una función discriminante, ésta explica un 100 % de la discriminación). El valor de la correlación canónica 0,723 (cercano a 1) señala la existencia de los dos grupos, lo cual indica también que las variables son buenas discriminadoras (permiten la separación de grupos)

Tabla 4.7: Lambda de Wilks

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	,477	19,979	6	,003

a. Se empleó solo una función discriminate canónica en el análisis

Un valor cercano a cero en el lambda de Wilks muestra que los grupos se encuentran poco mezclados. Si bien es cierto, en la tabla 4.7 esto no se cumple, es también claro que el modelo es significativo en la discriminación (valor -p = 0,003 <0,05), ello implica que la función es buena para discriminar.

A continuación se presentan tres tablas relacionadas con los coeficientes de cada variable.

Tabla 4.8: Coeficientes estandarizados de la función discriminante canónica.

Variables	Función
USO	1,041
CONCIENCIA	-0,036
OPINIÓN	-0,301
PREFERENCIA	0,221
INTENCIÓN	0,309
LEALTAD	-0,251

Tabla 4.9: Matriz estructura

Variables	Función
USO	0,939
PREFERENCIA	0,477
OPINIÓN	0,426
CONCIENCIA	0,293
INTENCIÓN	0,225
LEALTAD	0,050

Correlaciones intra – grupo combinadas entre las variables discriminantes y las funciones discriminantes canónicas tipificadas.

Tabla 4.10: Coeficientes de la función discriminante canónica.

Variables	Función
USO	1,668
CONCIENCIA	-0,018
OPINIÓN	-0,154
PREFERENCIA	0,148
INTENCIÓN	0,173
LEALTAD	-0,136
(Constante)	-3,486

Coeficientes no tipificados.

Con los coeficientes estandarizados de la tabla 4.8 se conoce con certeza y en porcentaje el aporte que tiene cada variable en la discriminación. Así, si se suman

los valores absolutos de cada uno de ellos, resulta 2,159 y al dividir 1,041 por 2,159 se obtiene 0,482. Es decir, la variable " uso " aporta un 48,2 % en la discriminación.

La matriz estructura (Tabla 4.9) presenta las variables ordenadas por su grado de correlación con la (de mayor a menor) con la función discriminante. Este orden puede ser distinto en el que aparecen en otras tablas y del orden en que han sido incluidas en el análisis.

Esto confirma los análisis anteriores al verificar como la variable " uso " es la que en mayor grado está correlacionada con la variable respuesta " sexo " y en gran parte ello puede deberse a cómo dicha variable explicativa ha capitalizado la información del resto de las variables explicativas.

Finalmente con la tabla 4.10 se obtiene la función discriminante. Esta función permite realizar pruebas de error, evaluando observaciones (no tipificadas) de la muestra de validación y comparar el resultado estimado con el observado.

$$Z = K_0 + K_1X_1 + \ldots + K_mX_m$$

$$Z = -3,486 + 1,668\,uso - 0,018\,conc - 0,154\,opin + 0,148\,prefer + 0,173\,inten - 0,136\,lealtad$$

Tabla 4.11: Funciones en los centroides de los grupos

GRUPO	Función
1	-1,014
2	1,014

Las funciones discriminantes canónicas sin estandarizar se han evaluado en las medias de los grupos.

La tabla 4.11 concentra la información referida a los centroides. Los centroides son valores que indican el punto capaz de equilibrar las distancias que hay desde cada individuo al centroide. En este caso cada valor de centroide asocia cada grupo de sexo. En el grupo 1 (mujeres) el valor del centroide es -1,014 y en el grupo 2 (hombres) el valor del centroide es 1,014. En consecuencia, si en la función discriminante se evalúan los valores de las variables para una cierta unidad observable, se obtendrá un puntaje que clasifica a dicha unidad observable en uno de estos grupos, tal como se muestra en la tabla 4.12. Por ejemplo la observación 32 es un hombre y es pronosticada en el grupo de las mujeres, pues obtiene un puntaje negativo y cercano al centroide del grupo 1 (mujeres). Esta incorrecta clasificación se repite en las observaciones 32, 35 y 36. Sin embargo, para el resto de las observaciones de la muestra de validación, los resultados indican correctas clasificaciones. Debe tenerse en cuenta que ningún modelo es perfecto.

Tabla 4.12: Resultados de clasificación con la muestra de validación.

Observación N°	Uso	Conciencia	Opinión	Preferencia	Intención	Lealtad	Grupo real	Puntaje o valor Z	Grupo pronosticado	Clasificación
31	1	1	2	9	3	2	2	-0,565	1	Incorrecta
32	1	4	6	5	9	3	2	-0,925	1	Incorrecta
35	3	5	7	7	3	3	1	1,497	2	Incorrecta
36	3	6	5	7	3	4	1	1,651	2	Incorrecta
37	3	6	7	5	3	4	2	1,047	2	Correcta
38	3	5	6	4	3	2	2	1,343	2	Correcta
39	3	7	7	6	3	4	2	1,177	2	Correcta
40	1	4	3	4	6	5	1	-1,402	1	Correcta
41	1	2	3	4	5	6	1	-1,675	1	Correcta
42	1	1	3	2	3	4	1	-2,027	1	Correcta
43	1	2	4	3	6	7	1	-1,94	1	Correcta
44	1	3	3	4	6	5	1	-1,384	1	Correcta
45	1	1	1	4	5	3	1	-0,941	1	Correcta

1: Mujer 2: Hombre

$$Sexo = -3{,}486 + 1{,}668\, uso - 0{,}018\, conc - 0{,}154\, opin + 0{,}148\, prefer + 0{,}173\, inten - 0{,}136\, lealtad$$

Es interesante también conocer cómo responde el modelo discriminante con los datos de la muestra de diseño. Para este propósito, SPSS entrega un tabla de resultados, que es la siguiente:

Tabla 4.13: Resultados de clasificación.

			Pertenencia a grupos pronosticada		
		SEXO	1	2	Total
Original	Recuento	1	14	2	16
		2	5	11	16
	%	1	87,5	12,5	100,0
		2	31,3	68,8	100,0
Validación cruzada[b]	Recuento	1	13	3	16
		2	5	11	16
	%	1	81,3	18,8	100,0
		2	31,3	68,8	100,0

a. 78,1% de casos agrupados originales clasificados correctamente.

b. La validación cruzada se ha realizado sólo para aquellos casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas de todos los casos distintos a dicho caso.

c. 75,0% de casos agrupados validados de forma cruzada clasificados correctamente.

La tabla 4.13, sobre el ejemplo Nike, también muestra los resultados de la clasificación basados en la muestra de análisis. La proporción de aciertos, o porcentaje de casos correctamente clasificados, es $(14 + 11)/32 = 0{,}781$ ó 78,1 por ciento. Podría pensarse que esta proporción de aciertos se infló de manera artificial, porque los datos usados para el cálculo también se emplearon para la validación. La validación cruzada con exclusión clasifica correctamente sólo $(13 + 11)/32 = 0{,}75$ ó 75 por ciento de los

casos. Realizar un análisis de clasificación de los datos de un conjunto independiente de validación da por resultado una matriz de clasificación con una proporción de aciertos de $9/13 = 0{,}692$ u 69,2 por ciento (véase la tabla 4.12). Dados dos grupos de igual tamaño, podría esperarse por azar una proporción de aciertos de $1/2 = 0{,}50$ o 50 por ciento. Por lo tanto, la mejoría sobre el azar es de más de 25 por ciento por lo que se juzga satisfactoria la validez del análisis discriminante.

4.8 Análisis discriminante múltiple

Con los mismos datos del ejemplo Nike se hará el análisis discriminante múltiple. Para tal propósito, ahora se utilizará la columna de la variable " Uso ", para poder agrupar tres grupos o tipos de usuarios: los frecuentes, moderados y esporádicos (1, 2 y 3 respectivamente). La pregunta de interés es si los tipos de usuarios frecuentes, moderados y esporádicos (USO) se diferencian en términos de su sexo (SEXO), la conciencia por la imagen que tiene de la marca (CONCIENCIA), la opinión de la zapatilla Nike (OPINIÓN), el nivel de preferencia (PREFERENCIA), la intensión de comprar la zapatilla Nike (INTENSIÓN) y la lealtad por la marca Nike (LEALTAD).

Para encontrar los coeficientes de la ecuación que nos permita discriminar haremos uso del programa SPSS, siguiendo los siguientes comandos:

1. Elija ANALIZE de la barra menú de SPSS.
2. Haga clic en CLASSIFY y luego en DISCRIMINAT.
3. Cambie " Uso " al recuadro GROUPING VARIABLE.
4. Haga clic en DEFINE RANGE. Introduzca un 1 en MINIMUN y 3 en MAXIMUN. Haga clic en CONTINUE.
5. Lleve las variables explicativas al recuadro INDEPENDENTS.
6. Elija ENTER INDEPENDENTS TOGETHER (opción predeterminada)
7. Haga clic en STATISTICS. En el recuadro DESCRIPTIVES de la ventana emergente, elija MEANS y UNIVARIATE ANOVAS. En el recuadro MATRICES elija WITHIN - GROUP CORRELATIONS y en el recuadro COEFICIENTES DE LA FUNCIÓN elija FISHER y NO ESTANDARIZADOS. Haga clic en CONTINUE.
8. Haga clic en CLASIFFY ... en la ventana emergente en la opción PRIOR PROBABILITIES, elija ALL GROUPS EQUAL (opción predeterminada). En el recuadro DISPLAY elija SUMMARY TABLE y LEAVE - ONE - OUT CLASSIFICATION. En el recuadro USE COVARIANCE MATRIX elija WITHING - GROUPS. Haga clic en CONTINUE.
9. Seleccione OK.

4.8.1. Estimación de los coeficientes de la función discriminante.

En primer lugar, si se revisa la tabla de estadísticas de los grupos, se deduce a priori, que las variables más aportantes a la discriminación son la " conciencia " y la " opinión ", ya que muestran mucha deiferencia en los promedios de los puntajes de cada uno de los tres grupos. Por el contrario, la " lealtad " parece ser la que menos estaría aportando en la discriminación del tipo de usuario, dado que no exhibe grandes diferencias en los promedios de cada tipo de usuario.

Tabla 4.14: Estadísticas de los grupos.

USO		Media	Desviación estándar	N válido (por lista)	
				No ponderados	Ponderados
1	CONCIENCIA	3,55	2,841	11	11,000
	OPINIÓN	1,73	,786	11	11,000
	PREFERENCIA	3,18	1,079	11	11,000
	INTENCIÓN	3,09	2,119	11	11,000
	LEALTAD	3,36	1,912	11	11,000
	SEXO	1,00	,000	11	11,000
2	CONCIENCIA	4,67	1,225	9	9,000
	OPINIÓN	4,11	1,900	9	9,000
	PREFERENCIA	3,44	1,590	9	9,000
	INTENCIÓN	4,33	1,323	9	9,000
	LEALTAD	4,22	1,563	9	9,000
	SEXO	1,67	,500	9	9,000
3	CONCIENCIA	6,00	,603	12	12,000
	OPINIÓN	5,08	1,782	12	12,000
	PREFERENCIA	5,25	1,422	12	12,000
	INTENCIÓN	4,17	1,697	12	12,000
	LEALTAD	3,83	1,946	12	12,000
	SEXO	1,83	,389	12	12,000
Total	CONCIENCIA	4,78	2,059	32	32,000
	OPINIÓN	3,66	2,104	32	32,000
	PREFERENCIA	4,03	1,636	32	32,000
	INTENCIÓN	3,84	1,798	32	32,000
	LEALTAD	3,78	1,809	32	32,000
	SEXO	1,50	,508	32	32,000

La matriz de correlaciones agrupadas intragrupales (tabla 4.15) indica cierta correlación entre la " intención " y la " lealtad ". Sin embargo, por tratrase de tan solo dos variables dentro del conjunto de variables explicativas, el problema de multicolinealidad no es de gravedad. La significancia asignada a las razones F

univariadas indica que cuando los predictivos se consideran de manera individual, sólo la " intención " y la " lealtad " no son significativas para diferenciar entre los dos grupos. Todas las demás variables explicativas son altamente significativas (tabla 4.16).

Tabla 4.15: Matriz dentro de grupos combinados.

		CONCIENCIA	OPINIÓN	PREFERENCIA	INTENCIÓN	LEALTAD	SEXO
Correlación	CONCIENCIA	1,000	,158	,017	-,220	,108	-,106
	OPINIÓN	,158	1,000	,339	,261	,351	-,218
	PREFERENCIA	,017	,339	1,000	,328	,356	,201
	INTENCIÓN	-,220	,261	,328	1,000	,758	,018
	LEALTAD	,108	,351	,356	,758	1,000	-,088
	SEXO	-,106	-,218	,201	,018	-,088	1,000

Tabla 4.16: Prueba de igualdad de medias de grupos.

	Lambda de Wilks	F	df1	df2	Sig.
CONCIENCIA	,736	5,208	2	29	,012
OPINIÓN	,510	13,929	2	29	,000
PREFERENCIA	,652	7,734	2	29	,002
INTENCIÓN	,904	1,544	2	29	,231
LEALTAD	,964	,549	2	29	,583
SEXO	,458	17,136	2	29	,000

En el análisis discriminante múltiple, si existen G grupos, pueden calcularse $G-1$ funciones discriminantes, si el número de predictivos es mayor que esta cantidad. En general con G grupos y k predictivos, es posible calcular las funciones discriminantes más pequeñas de $G-1$ o k. La primera función tiene la razón más elevada de la suma de cuadrados entre e intragrupos. La segunda función, no correlacionada con la primera, tiene la segunda razón más alta y así sucesivamente. Sin embargo, no todas las funciones pueden ser estadísticamente significativas (Malhotra, 2004).

Dado que en el ejemplo hay tres grupos, como máximo pueden extraerse dos funciones. El valor propio asociado con la primera función es 3,249 y esta función da cuenta del 89,9 % de la varianza explicada. Como el valor propio es grande, es probable que la primera función sea superior. El valor propio de la segunda función es pequeño (0.367) y sólo da cuenta del 10,1 % por ciento de la varianza explicada (tabla 4.17).

Tabla 4.17: Autovalores.

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	3,249[a]	89,9	89,9	,874
2	,367[a]	10,1	100,0	,518

a. Se utilizaron las primeras 2 funciones discriminantes canónicas en el análisis.

4.8.2. Determinación de la significancia de la función discriminante.

En la tabla 4.18 el valor de λ de Wilks es 0.172, que se transforma en una chi cuadrada de 46,612, con 12 grados de libertad, lo cual es significativo más allá del nivel de 0.05. Por tanto, ambas funciones en conjunto hacen una discriminación significativa entre los tres grupos. Sin embargo, cuando se elimina la primera función, la λ de Wilks asociada con la segunda función es 0,732, lo que no es significativa al nivel de 0.05. Por consiguiente, la segunda función no hace una contribución significativa a las diferencias entre los grupos.

Tabla 4.18: Lambda de Wilks.

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1 a 2	,172	46,612	12	,000
2	,732	8,277	5	,142

4.8.3. Interpretación de los resultados.

La interpretación de los resultados se facilita al examinar los coeficientes estandarizados de la función discriminante, la estructura de correlaciones y ciertas gráficas. Los coeficientes estandarizados indican un coeficiente grande para el sexo y opinión en la función 1; mientras que la función 2 tiene coeficientes relativamente altos para preferencia y lealtad. Cuando se examina la matriz estructural se llega a una conclusión similar (véanse las tablas 4.19 y 4.20). Para ayudar a la interpretación de las funciones, se agruparon las variables con coeficientes altos para una función particular. Esas agrupaciones se muestran con asteriscos. Por lo tanto, sexo, opinión y conciencia tienen asteriscos en la función 1, porque estas variables tienen coeficientes mayores en la función 1 que en la función 2. Esas variables se asocian sobre todo con la función 1. Por otro lado, la preferencia, la lealtad y la intención se asocian en forma predominante con la función 2, como lo señalan los asteriscos.

Tabla 4.19: Coeficientes estandarizados de la función discriminante canónica.

Variables	Función 1	Función 2
SEXO	0,763	-0,444
CONCIENCIA	0,431	0,339
OPINIÓN	0,706	-0,296
PREFERENCIA	-0,044	1,066
INTENCIÓN	0,419	0,088
LEALTAD	-0,449	-0,634

Tabla 4.20: Matriz estructuras

Variables	Función 1	Función 2
SEXO	0,601*	-0,143
OPINIÓN	0,544*	0,016
CONCIENCIA	0,320*	0,270
PREFERENCIA	0,333	0,685*
LEALTAD	0,080	-0,216*
INTENCIÓN	0,168	-0,202*

Correlaciones intra – grupo combinadas entre las variables discriminantes y las funciones discriminantes canónicas tipificadas.
* Mayor correlación absoluta entre cada variable y cualquier función discriminante.

En la tabla 4.21 se tienen los coeficientes de las funciones discriminantes. Con esta información se puede construir cada una de las transformaciones lineales (funciones 1 y 2) para poder discriminar.

Tabla 4.21: Coeficientes de la función discriminante canónica.

Variables	Función 1	Función 2
SEXO	2,146	-1,248
CONCIENCIA	0,236	0,186
OPINIÓN	0,454	-0,191
PREFERENCIA	-0,032	0,780
INTENCIÓN	0,237	0,050
LEALTAD	-0,244	-0,345
(Constante)	-5,863	-0,348

Coeficientes no estandarizados

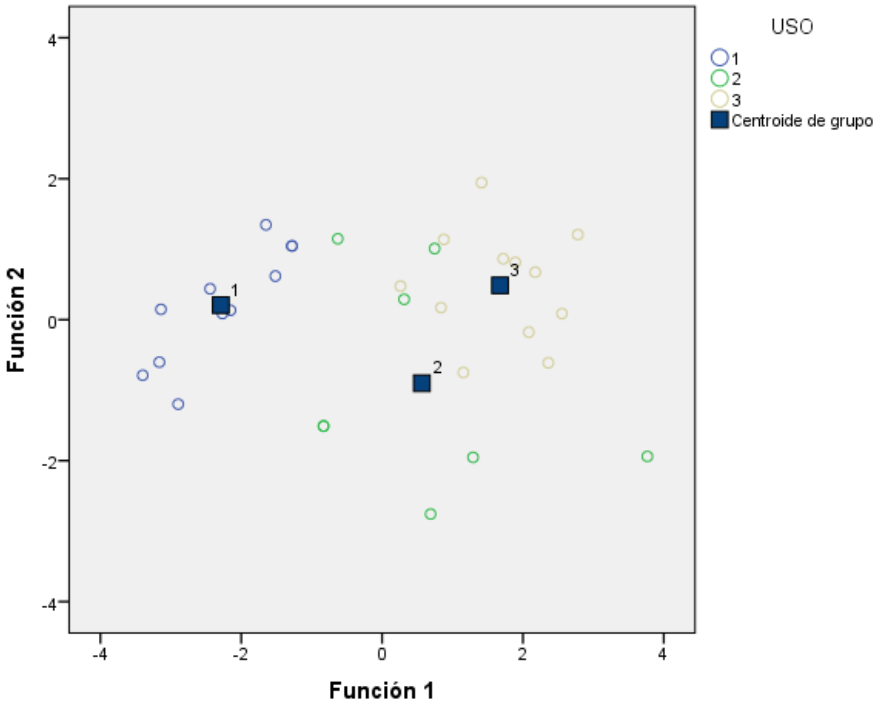
Tabla 4.22: Funciones en centroides de grupo.

	Función	
USO	1	2
1	-2,290	0,207
2	0,565	-0,902
3	1,675	0,487

Las funciones discriminantes canónicas sin estandarizar se han Evaluado en medias de grupos.

Por su parte, la tabla 4.22 muestra los centroides de cada uno de los tres grupos y esta información se puede complementar con el gráfico de dispersión que se muestra en la figura 4.1

Figura 4.1: Funciones discriminantes canónicas.



La figura 4.1 es un diagrama de dispersión de todos los grupos en la función 1 y en la función 2. Se observa que el grupo 3 tiene el valor más alto en la función 1 y el grupo 1 el menor. Dado que la función 1 se asocia sobre todo con el sexo, la opinión y la conciencia, se esperaría que los tres grupos estén ordenados en esas tres variables. Es probable que los varones conscientes de la marca y que tienen una opinión muy favorable de Nike sean los usuarios con mayor frecuencia de las zapatillas Nike y que, a la inversa, la menor frecuencia de uso por dicha zapatilla provenga de las mujeres con baja lealtad, opinión muy poco favorable como así también una disminuida conciencia con la marca Nike. Esta interpretación se fortalece al examinar las medias de los grupos en sexo, opinión y conciencia.

La figura 4.1 indica también que la función 2 tiende a separar el grupo 3 (valor más alto) del grupo 2 (valor más bajo). Esta función se asocia sobre todo con preferencia, lealtad e intención.

Finalmente, se presenta la tabla de resultados de clasificación (tabla 4.23), en donde para las observaciones de la muestra de diseño se tiene que de los 11 individuos pertenecientes al grupo 1 (uso frecuente), la función discriminante los estima correctamente. De nueve individuos pertenecientes al grupo 2 (uso moderado) seis fueron clasificados correctamente y de doce individuos pertenecientes al grupo 3 (uso esporádico) diez fueron clasificados en tal grupo. Estos resultados se resumen en que un 84,4 % de los casos agrupados originales fueron clasificados correctamente.

El mismo análisis se realiza para la validación cruzada; obteniendo solo un 71,9 % de los casos agrupados validados de forma cruzada clasificados correctamente.

Tabla 4.23: Resultados de clasificación.

			Pertenencia a grupos pronosticada			
		USO	1	2	3	Total
Original	Recuento	1	11	0	0	11
		2	1	6	2	9
		3	0	2	10	12
	%	1	100,0	,0	,0	100,0
		2	11,1	66,7	22,2	100,0
		3	,0	16,7	83,3	100,0
Validación cruzada[b]	Recuento	1	11	0	0	11
		2	3	3	3	9
		3	1	2	9	12
	%	1	100,0	,0	,0	100,0
		2	33,3	33,3	33,3	100,0
		3	8,3	16,7	75,0	100,0

a. 84,4% de casos agrupados originales clasificados correctamente.

b. La validación cruzada se ha realizado sólo para aquellos casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas de todos los casos distintos a dicho caso.

c. 71,9% de casos agrupados validados de forma cruzada clasificados correctamente.

En resumen, el análisis discriminante es una técnica bastante útil cuando la variable dependiente o de criterio es categórica y las variables independientes son de una escala de intervalos. Cuando la variable de criterio tiene dos categorías, el análisis discriminante se conoce como simple. Cuando la variable criterio tiene tres o más categorías, se llama análisis discriminante múltiple (Malhotra, 2004).

En el análisis discriminante múltiple, si hay G grupos y k predictivos, es posible calcular la menor de $G - 1$ o k funciones discriminantes (Malhotra, 2004).

CAPÍTULO V

ANÁLISIS DE VARIANZA

5.1. Objetivo

Comprender y aplicar el análisis de varianza univariado y multivariado en problemáticas relacionadas con las ciencias de la administración.

5.2. Antecedentes

Ronald Fisher fue innovador del uso de los métodos estadísticos en el diseño de experimentos (Montgomery, 1991). Durante algunos años estuvo a cargo de la estadística y análisis de datos en una estación agrícola en Londres y fue ahí donde inició su aplicación. Fue él quien desarrolló y usó por primera vez el análisis de varianza como herramienta primaria para el análisis estadístico en el diseño de experimentos. Muchas de las primeras aplicaciones industriales se dieron en el área de la agricultura y ciencias biológicas; sin embargo, sus primeras aplicaciones industriales se hicieron en la década de 1930 en las industrias textil y de lana británica (De la Garsa, 2013)

5.3. Áreas de aplicación

El análisis de varianza tiene muchas aplicaciones y en diferentes disciplinas; su uso es muy importante para la mejora de procesos productivos, para el desarrollo de nuevos procesos, el desarrollo de nuevos productos y la mejora de otros ya existentes (De la Garsa, 2013).

El análisis de varianza descompone la variabilidad del resultado de un experimento en componentes independientes. Por ejemplo, se puede considerar la productividad de tres máquinas compradas a un mismo proveedor que, aunque provienen de la misma empresa, pueden producir cantidades distintas de piezas. Esa variabilidad puede producirse por muchos factores controlables (ajuste de máquina, presión ejercida, etc.), donde cada uno puede presentar diferentes niveles (ajuste de primer nivel, de segundo o cantidades distintas de presión), o bien por factores no controlables (clima, cansancio del operador, etc.).

También es útil para conocer qué campaña publicitaria tiene más impacto en las ventas de algún producto, qué tipo de aprendizaje repercute en las calificaciones de los alumnos de ciertas carreras o cuál dieta ayuda a bajar más de peso. Se pueden

modelar los fenómenos que surjan en los procesos productivos, procedimientos de manufactura, métodos de aprendizaje y sobre todo cuando a los individuos, bienes o servicios se les aplican diferentes tratamientos como en medicina y psicología, entre otros (De la Garsa, 2014).

5.4. Definición

El **análisis de varianza** es una técnica funcional que se emplea básicamente para la experimentación; esta técnica utiliza una o más variables independientes, todas no métricas, y trata de explicar el comportamiento de una o más variables dependientes métricas (De la Garsa, 2013).

En este tipo de técnicas o métodos se descompone el comportamiento de las variables dependientes en tres o más variaciones, las cuales indican qué tanto se logró explicar de dicho comportamiento. Es por ello que el análisis de varianza es el estudio de la variabilidad del resultado de un experimento y se puede dividir en dos partes: una que se origina por los factores que influyen directamente en el resultado del experimento y otra que es producida por el resto de los factores que también influyen en el resultado del experimento con variabilidad no controlable, que se conoce como *error muestral* (De la Garsa, 2013).

5.5. Análisis de varianza unidireccional

Los investigadores de las ciencias administrativas se interesan en examinar las diferencias de las medias en varias categorías de una sola variable dependiente o factor único. Por ejemplo:

- ¿los empleados de los diferentes departamentos difieren por el grado de motivación en el trabajo?
- ¿Los segmentos se diferencian por su volumen de consumo del producto?
- ¿Cuál es el efecto de la familiaridad de los consumidores con las tiendas (medida como alta, mediana y débil) en la preferencia por nuestra tienda?
- ¿Varían las evaluaciones de las marcas hechas por los grupos expuestos a diversos mensajes publicitarios?
- ¿Cómo varía el interés participativo de los integrantes de una junta de vecinos según las diversas propuestas planteadas por el municipio?

Las respuestas a interrogantes como éstas y otras similares, se encuentra en un análisis de varianza de un factor.

5.6. Estadísticas del análisis de varianza unidireccional

η^2. La fuerza de los efectos de X (variable o factor independiente) sobre Y (variable dependiente) se mide por medio de η^2 y su valor va de 0 a 1.

Estadístico F. La hipótesis nula que plantea que las medias de las categorías son iguales en la población se pone a prueba usando un estadístico F, que se basa en la proporción del cuadrado medio con respecto a X y el cuadrado medio relacionado con el error.

Cuadrado medio. El cuadrado medio es la suma de cuadrados dividida entre los grados de libertad adecuados.

SC_{entre}. También simbolizada por SC_x, es la variación de Y relacionada con la variación en las medias de las categorías de X. Ésta representa la variación entre las categorías de X, o la porción de la suma de cuadrados en Y relacionada con X.

SC_{dentro}. También conocida como SC_{error}, es la variación en Y debida a la variación dentro de cada una de las categorías de X. Esta variación no está explicada por X.

SC_y. Variación total de Y.

5.7. Pasos para realizar análisis de varianza unidireccional

El procedimiento para realizar análisis de variancia de un factor. Consiste en identificar las variables dependiente e independiente, descomponer la variación total, medir los efectos, probar la significancia e interpretar los resultados. A continuación, describimos estos pasos de manera detallada y se ejemplificará con una aplicación.

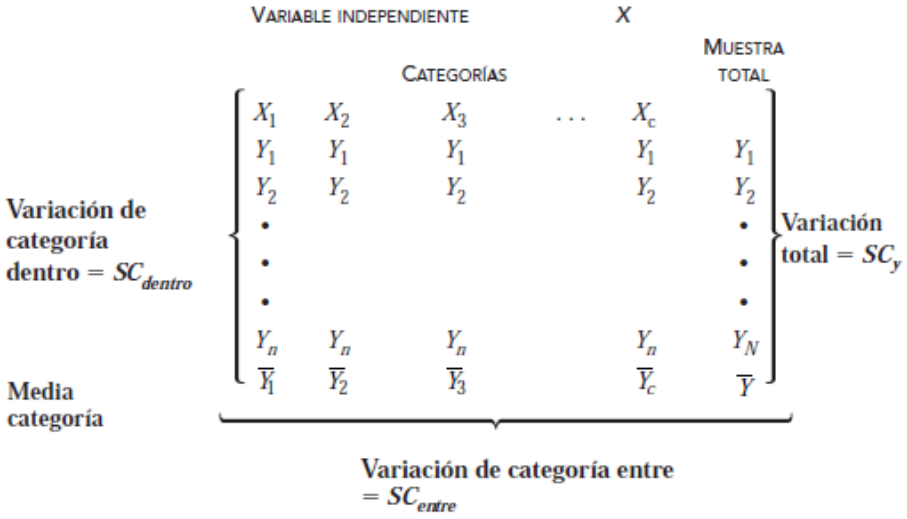
5.7.1. Identificación de las variables dependiente e independiente.

La variable dependiente se simboliza con Y y la variable independiente con X. X es una variable categórica con c categorías. Existen **n** observaciones de Y para cada categoría de X, como se muestra en la figura 5.1. Como se observa, el tamaño de la muestra en cada categoría de X es n y el tamaño total de la muestra $N = n * c$. Aunque por razones de simplicidad se supondrá que los tamaños de las muestras en las categorías de X (los tamaños de los grupos) son iguales, lo que necesariamente no es un requisito.

5.7.2. Descomposición de la variación total.

Al examinar las diferencias entre medias, el análisis de varianza de un factor requiere de la descomposición de la variación total observada en la variable dependiente. Esta variación se mide usando la suma de cuadrados corregida para la media (SC). El análisis de varianza recibe su nombre porque examina la variabilidad o variación en la muestra (variable dependiente) y, con base en la variación, determina si hay alguna razón para creer que las medias poblacionales son diferentes (Malhotra, 2004).

Figura 5.1: Descomposición de la variación total: ANOVA de un factor.



La variación total en Y, simbolizada por SC_y, se resuelve en dos componentes:

$$SC_y = SC_x + SC_{error}$$

Donde SC_y es la variación total en Y, $= SC_x$ es la variación en Y relacionada con la variación en las medias de las categorías de X; representa la variación entre las categorías de X. En otras palabras, SC_x es la porción de la suma de cuadrados en Y relacionada con la variable independiente o factor X y por último SC_{error} no está explicada por X y, por lo tanto, se le conoce como la suma de los cuadrados del error.

El cálculo de cada una de las expresiones anteriores, se presentan a continuación:

$$SC_y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$SC_x = \sum_{j=1}^{c} n(Y_j - \bar{Y})^2$$

$$SC_{error} = \sum_{j=1}^{c} \sum_{i=1}^{n} (Y_{ij} - \bar{Y_j})^2$$

donde

Y_i = observación individual i
$\bar{Y_j}$ = media de la categoría j
$\bar{Y}$ = media de toda la muestra o gran media
Y_{ij} = i - ésima observación de la j- ésima categoría

En el análisis de varianza hay varios grupos (por ejemplo, usuarios frecuentes, intermedios, esporádicos y no usuarios). Si la hipótesis nula es verdadera y todos los grupos tienen la misma media en la población, podemos estimar cuánto deben variar las medias muestrales únicamente por las variaciones de muestreo (aleatorias). Si la variación observada en las medias muestrales es mayor a lo esperado por la variación de muestreo, es razonable concluir que esta variación adicional se relaciona con las diferencias entre las medias grupales de la población (Malhotra, 2004).

En el análisis de varianza se calculan dos medidas de variación: dentro de los grupos ($SC_{dentro} = SC_x$) y entre grupos ($SC_{entre} = SC_{error}$). La variación dentro de los grupos es una medida de cuánto varían dentro de un grupo las observaciones o valores de Y. Esto se utiliza para estimar la varianza dentro de un grupo en la población. Se asume que todos los grupos tienen la misma variación en la población. Sin embargo, debido a que no se sabe si todos los grupos tienen la misma media, no podemos calcular la varianza de todas las observaciones en conjunto. La varianza de cada grupo se debe calcular de manera individual, y luego las varianzas se combinan en una varianza "promedio" o "general". De la misma forma, se puede obtener otro estimado de la varianza de los valores de Y al examinar las variaciones entre las medias (este proceso es inverso a la determinación de la variación en las medias, dadas las varianzas poblacionales). Si la media poblacional es igual en todos los grupos, entonces se puede utilizar la variación de las medias muestrales y el tamaño de los grupos de muestras para estimar la varianza de Y. La sensatez de esta estimación de la varianza de Y depende de si la hipótesis nula es verdadera. Si la hipótesis nula es verdadera y las medias de la población son iguales, la estimación de la varianza basada en la variación entre grupos es correcta. Por otro lado, si los grupos tienen medias diferentes en la población, la estimación de la varianza basada en la variación entre grupos será demasiado grande. Así pues, al comparar los estimados de la varianza de Y con base en la variación entre grupos y dentro de grupos, ponemos a prueba la hipótesis nula. El hecho de descomponer la variación total de esta manera nos permite medir los efectos de X sobre Y (Malhotra, 2004).

5.7.3. Medida de los efectos.

Los efectos de X sobre Y se miden con SC_x. Debido a que SC_x está relacionada con la variación en las medias de las categorías de X, la magnitud relativa de SC_x aumenta conforme se incrementan las diferencias entre las medias de Y en las categorías de X. La magnitud relativa de SC_x también aumenta conforme las variaciones en Y dentro de las categorías de X disminuyen. La fuerza de los efectos de X sobre Y se miden de la siguiente manera:

$$\eta^2 = \frac{SC_x}{SC_y} = \frac{SC_y - SCerror}{SC_y} = 1 - \frac{SCerror}{SC_y}$$

El valor de η^2 varía entre 0 y 1. Se asume un valor de 0 cuando todas las medias de la categoría son iguales, indicando así que X no tiene un efecto sobre Y. El valor de η^2 es 1 cuando no existe variación dentro de cada categoría de X, pero existe cierta variación entre las categorías. De esta manera, η^2 es una medida de la variación en Y que está explicada por la variable independiente X. No sólo podemos medir los efectos de X sobre Y, sino que también podemos hacer una prueba de su significancia (Malhotra, 2004).

5.7.4. Prueba de significación.

En el análisis de varianza de un factor, el interés reside en poner a prueba la hipótesis nula que plantea que las medias de las categorías son iguales en la población. En otras palabras,

$$H_0 : \mu_1 = \mu_2 = \mu_3 \ldots = \mu_c$$

De acuerdo con la hipótesis nula, SC_x y SC_{error} provienen de la misma fuente de variación. En tal caso, el estimado de la varianza poblacional de Y se puede basar en la variación de la categoría entre o en la variación de la categoría dentro. En otras palabras, el estimado de la varianza poblacional de Y,

$SC_y^2 = \frac{SC_x}{c-1}$

$= cuadrado\ medio\ debido\ a\ X$

$= CM_x$

o bien

$SC_y^2 = \frac{SC_{error}}{N-c}$

$= cuadrado\ de\ las\ medias\ debido\ al\ error$

$= CM_{error}$

La hipótesis nula se prueba con el estadístico F, con base en la proporción entre los siguientes dos estimados:

$$F = \frac{SC_x/(c-1)}{SC_{error}/(N-c)} = \frac{CM_x}{CM_{error}}$$

Este estadístico tiene una distribución F, con $(c-1)$ y $(N-c)$ grados de libertad (gl). En la tabla 5 del apéndice estadístico, al final del libro, se incluye una tabla de la distribución F. Como se sabe, la distribución F es una distribución de probabilidad de las proporciones de las varianzas muestrales. Se caracteriza por tener grados de libertad para el numerador y grados de libertad para el denominador.

5.7.5. Interpretación de resultados.

Si la hipótesis nula que plantea medias de categoría iguales no se rechaza, entonces la variable independiente no tiene un efecto significativo sobre la variable dependiente. Por otro lado, si se rechaza la hipótesis nula, entonces el efecto de la variable independiente es significativo. En otras palabras, el valor promedio de la variable dependiente será diferente para distintas categorías de la variable independiente. Una comparación de los valores promedio de las categorías indica la naturaleza del efecto de la variable independiente (Malhotra, 2004).

El análisis de varianza para un factor es importante para comprender los conceptos básicos de esta técnica estadística. Sin embargo, se dejará el ejemplo del análisis de varianza unifactorial para que el alumno por cuenta propia lo investigue, debido a que es mucho más provechoso estudiar el análisis de varianza con **n** factores.

5.8. Análisis de varianza con n factores

En las investigaciones de las ciencias administrativas, es usual interesarse en el efecto simultaneo de dos o más factores. Por ejemplo:

- ¿Cómo varía la intención de comprar una marca con los niveles de precios y los niveles de distribución?
- ¿Cómo se relacionan los niveles de escolaridad (media, técnico, universitaria) y el sector donde se vive (alto, medio, bajo) con la intención de votar en una elección parlamentaria?

- ¿Cuál es el efecto de la familiaridad de los consumidores con las tiendas (medida como alta, mediana y débil) y la imagen de esa tienda (positiva, indiferente y negativa) en la preferencia por ella?

Para determinar este tipo de efectos, se puede emplear un análisis de varianza de n factores. Una de las principales ventajas de esta técnica es que permite al investigador examinar interacciones entre los factores. Las **interacciones** ocurren cuando los efectos de un factor sobre la variable dependiente dependen del nivel (categoría) de los otros factores. El procedimiento para realizar un análisis de varianza de n factores es similar al del análisis de varianza de un factor. Los estadísticos asociados con el análisis de varianza de n factores también se definen de manera similar (MAlhotra, 2004). Considere el caso sencillo de dos factores, X_1 y X_2, con categorías c_1 y c_2. La variación total en este caso se parte de la siguiente manera:

$$SC_{total} = SC debido\ a\ X_1 + SC debido\ a X_2 + SC debido\ a\ X_1\ y\ X_2 + SCE$$

o bien

$$SC_Y = SC_{X_1} + SC_{X_2} + SC_{X_1X_2} + SC_{error}$$

Un mayor efecto de X_1 se reflejará en una mayor diferencia promedio en los niveles de X_1 y en una SC_{X_1} más grande. Lo mismo ocurre con el efecto de X_2. Cuanto más grande sea interacción entre X_1 y X_2, mayor será $SC_{X_1X_2}$. Por otro lado, si X_1 y X_2 son independientes, el valor de $SC_{X_1X_2}$ se acercará a cero (Malhotra, 2004),

La fuerza del efecto conjunto de dos factores, llamado efecto general o η^2 **múltiple**, se mide como sigue:

$$\eta^2_{multiple} = \frac{SC_{X_1} + SC_{X_2} + SC_{X_1X_2}}{SC_Y}$$

La **significancia del efecto general** se prueba con una prueba F, de la siguiente manera:

$$F = \frac{(SC_{X_1} + SC_{X_2} + SC_{X_1X_2})/gl_N}{SC_{error}/gl_D}$$

$$= \frac{SC_{X_1,X_2,X_1X_2}/gl_N}{SC_{error}/gl_D}$$

$$= \frac{CM_{X_1,X_2,X_1X_2}}{CM_{error}}$$

donde

gl_N = grados de libertad del numerador
$gl_N = (c_1 - 1) + (c_2 - 1) + (c_1 - 1) \cdot (c_2 - 1)$
$gl_N = c_1 c_2 - 1$
gl_D = grados de libertad del denominador
$gl_D = N - c_1 c_2$
CM = cuadrado medio

Si el efecto general es significativo, el siguiente paso consiste en examinar la **significancia del efecto de interacción**. Para la hipótesis nula de no interacción, la prueba F apropiada es:

$$F = \frac{SC_{X_1X_2}/gl_N}{SC_{error}/gl_D}$$
$$= \frac{CM_{X_1X_2}}{CM_{error}}$$

donde

$gl_N = (c_1 - 1) \cdot (c_2 - 1)$
$gl_D = N - c_1 c_2$

Si el efecto de interacción resulta significativo, entonces el efecto de X_1 depende del nivel de X_2 y viceversa. Debido a que el efecto de un factor no es uniforme, sino que varía con el nivel del otro factor, generalmente no tiene caso poner a prueba la significancia de los efectos principales. Sin embargo, es importante poner a prueba la significancia de cada efecto principal de cada factor, si el efecto de interacción no es significativo (Malhotra, 2004).

La **significancia del efecto principal** de cada factor se prueba de la siguiente manera para X_1:

$$F = \frac{SC_{X_1}/gl_N}{SC_{error}/gl_D}$$
$$= \frac{CM_{X_1}}{CM_{error}}$$

donde

$gl_N = c_1 - 1$
$gl_D = N - c_1 c_2$

El análisis anterior asume que se trata de un diseño ortogonal o equilibrado (hay el mismo número de casos en cada celda). Si el tamaño de la celda varía, el análisis se vuelve más complejo.

5.9. Ejemplo de aplicación para el análisis de varianza de n factores

Se diseñó un estudio para una tienda departamental que quiere examinar los efectos principales y el efecto conjunto de dos factotres (cupones de descuento y promoción en tienda) sobre las ventas de una tienda departamental. Los resultados de la encuesta aplicada a 30 tiendas se presentan en la tabla 5.1.

Tabla 5.1: Base de datos.

N°de Tienda	CUPONES	PROMO	VENTAS	CLASIFIC
1	1	1	10	9
2	1	1	9	10
3	1	1	10	8
4	1	1	8	4
5	1	1	9	6
6	1	2	8	8
7	1	2	8	4
8	1	2	7	10
9	1	2	9	6
10	1	2	6	9
11	1	3	5	8
12	1	3	7	9
13	1	3	6	6
14	1	3	4	10
15	1	3	5	4
16	2	1	8	10
17	2	1	9	6
18	2	1	7	8
19	2	1	7	4
20	2	1	6	9
21	2	2	4	6
22	2	2	5	8
23	2	2	5	10
24	2	2	6	4
25	2	2	4	9
26	2	3	2	4
27	2	3	3	6
28	2	3	2	10
29	2	3	1	9
30	2	3	2	8

Fuente: Malhotra, 2004

La promoción en tienda incluyó tres niveles: alto (1), medio (2) y bajo (3). Los cupones se manipularon en dos niveles. Se distribuyó un cupón de 20 dólares entre compradores potenciales (simbolizado con 1) o no se distribuyó ninguno (simbolizado con 2 en la tabla 5.1). Se realizó un cruce de la promoción en tienda y de los cupones, lo que resultó en un diseño 3×2 con seis celdas. Se eligieron 30 tiendas al azar, y se asignaron aleatoriamente cinco a cada tratamiento, tal como se muestra en la tabla 16.2. El experimento duró dos meses y luego se midieron las ventas en cada tienda, de manera normalizada para explicar factores extraños (tamaño de la tienda, flujo de personas, etcétera), y luego se convirtieron a una escala de 1 a 10. Además, se realizó una evaluación cualitativa de la afluencia relativa de la clientela en cada tienda, utilizando nuevamente una escala de 1 a 10. En estas escalas, los números más altos indican mayores ventas o mayor afluencia de clientes.

Haciendo usos del programa SPSS, a continuación se desprenden los siguientes resultados del análisis de varianza para dos factores:

Tabla 5.2: Análisis de varianza de dos factores.

Variable dependiente: VENTAS

Origen	Tipo III de suma de cuadrados	gl	Cuadrático promedio	F	Sig.
Modelo corregido	163,505[a]	6	27,251	28,028	,000
Interceptación	103,346	1	103,346	106,294	,000
CLASIFIC	,838	1	,838	,862	,363
DISTRIB	53,333	1	53,333	54,855	,000
PROMO	106,067	2	53,033	54,546	,000
DISTRIB * PROMO	3,267	2	1,633	1,680	,208
Error	22,362	23	,972		
Total	1290,000	30			
Total corregido	185,867	29			

a. R al cuadrado = ,880 (R al cuadrado ajustada = ,848)

Si se consideran las ecuaciones anteriores, se determinan las estadísticas de prueba F para cada caso particular. Así se tiene que para el efecto general del modelo corregido es:

$$F = \frac{27,251}{0,972} = 28,036$$

con 5 y 24 grados de libertad, lo que es significativo al 1 % (0,000 en tabla)

El resto de las pruebas de significancia se resumen en las dos últimas columnas de la tabla 5.2. De esta manera, resulta que los efectos principales de los factores distribución de cupones y promoción en tienda son significativos (valores 0,000 en última columna de la tabla 5.2 para ambos factores). Sin embargo, la estadística de prueba F para el efecto de la interacción entre la promoción en tienda y la distribución de cupones es 1,68 que con 2 y 24 grados de libertad tiene un resultado no significativo al 5 %.

A modo de resumen se infiere que más promoción incrementa las ventas. La distribución de cupones en tienda trae también más ventas. El efecto de cada uno de estos factores es independiente uno del otro.

Con el programa SPSS, los pasos para ejecutar el análisis de varianza de este ejemplo, son los siguientes:

1. Elija ANALIZE de la barra menú de SPSS.
2. Haga clic en MODELO LINEAL GENERAL y luego en UNIVARIANTE.
3. Traslade “VENTAS” en el recuadro VARIABLE DEPENDIENTE.
4. Traslade “DISTRIB” Y “PROMO” al recuadro FACTORES FIJOS.
5. Traslade “CLASIFIC” al recuadro COVARIABLES.
6. Seleccione OK.

CAPÍTULO VI

ESCALAMIENTO MULTIDIMENSIONAL

6.1. Objetivo

Comprender y aplicar la técnica de escalamiento múltidimensional a problemáticas relacionadas con las ciencias de la administración.

6.2. Antecedentes

El escalamiento multidimensional trata de encontrar la estructura de un conjunto de medidas de distancia entre objetos o casos. Esto se logra asignando las observaciones a posiciones específicas en un espacio conceptual (normalmente de dos o tres dimensiones) de modo que las distancias entre los puntos en el espacio concuerden al máximo con las disimilaridades dadas. En muchos casos, las dimensiones de este espacio conceptual son interpretables y se pueden utilizar para comprender mejor los datos. Si las variables se han medido objetivamente, puede utilizar el escalamiento multidimensional como técnica de reducción de datos (Pérez, 2008).

Según Luque (2000), el escalamiento multidimensional (EMD) se origina en psicología como una respuesta a la necesidad de relacionar la intensidad física de ciertos estímulos con su intensidad subjetiva. Torgerson (1958) es considerado como uno de sus principales precursores, contribuyendo decisivamente a la clasificación y utilización de estos métodos. Este autor fue el primero en proponer una generalización del escalamiento. Pronto surgieron nuevos modelos y métodos que paulatina y sistemáticamente fueron cubriendo un amplio abanico de demandas realizadas desde diferentes campos de investigación como la Psicología, la Educación, Sociología, las Ciencias Políticas, la Economía y, por supuesto, el Marketing. Un factor que favoreció su desarrollo fue la evolución experimentada por los equipos informáticos y el software a partir de los años cincuenta (Pérez, 2008).

6.3. Áreas de aplicación

El EMD se aplica en la investigación de mercados para identificar:

1. El número y la naturaleza de las dimensiones que usan los consumidores para percibir diferentes marcas en el mercado.

2. El posicionamiento de las marcas actuales en tales dimensiones.

3. El posicionamiento de la marca ideal de los consumidores en esas dimensiones.

La información que arroja el EMD tiene variadas aplicaciones en ciencias administrativas, como las siguientes:

- *Medición de imagen.* Comparar las percepciones que se tiene de un conjunto de políticas de recursos humanos de quienes avalan la medida y de quienes no la respaldan con la percepción que tiene el propio departamento de recursos humanos, para identificar discrepancias perceptuales.
- *Desarrollo de nuevos productos.* Buscar discrepancias en el mapa perceptual, lo que indicaría posibles oportunidades para posicionar nuevos productos y marcas actuales a título de prueba, para determinar cómo perciben los clientes esos nuevos conceptos. La proporción de preferencias por cada producto es un indicador de su éxito.
- *Evaluar la eficacia de una campaña política.* Los mapas espaciales sirven también para determinar si la campaña política ha logrado convencer a los potenciales votantes.
- *Elaboración de escalas de opinión.* Las técnicas de EMD sirven para establecer las dimensiones apropiadas y la configuración del espacio de opinión.
- *Análisis de precios.* Pueden compararse mapas espaciales desarrollados con y sin información sobre los precios para determinar su impacto.

6.4. Estadísticas propias del escalamiento multidimensional

- *Juicios de semejanza* Los juicios de semejanza son calificaciones en una escala tipo Likert de todos los pares posibles de marcas u otros estímulos en términos de su semejanza.
- *Ordenamientos de preferencia* Los ordenamientos de preferencias son rangos ordenados de las marcas u otros estímulos de los más a los menos preferidos. Por lo general se obtienen de los encuestados.
- *Estrés* Es la falta de ajuste de la medida: los valores más altos de estrés indican un ajuste más pobre.
- *R cuadrada* Es un índice de correlación elevado al cuadrado, que indica la proporción de varianza de los datos escalados en forma óptima, que puede explicarse mediante el procedimiento del EMD. Es una medida de la bondad del ajuste.

- *Mapa espacial* Las relaciones percibidas entre marcas u otros estímulos se representan en relaciones geométricas entre puntos en un espacio multidimensional llamado mapa espacial.
- *Coordenadas* Indican el posicionamiento de la marca o el estímulo en un mapa espacial.
- *Despliegue* Es la representación de marcas y encuestados como puntos en el mismo espacio.

6.5. Pasos para la realización de escalamiento dimensional múltiple

Lo fundamental para llevar a buen puerto el escalamiento multidimensional es el planteamiento claro del problema y la forma apropiada para la obtención de los datos. A continuación, se detalla cada una de estas etapas.

6.5.1. Planteamiento del problema.

Primero que todo, el investigador debe plantear claramente el propósito a que se destinan los resultados del escalamiento multidimensional y elegir los estímulos o marcas que se incluirán en el análisis. Se recomienda entre 8 y 25 marcas o estímulos. En realidad, considerar más de 25 será más pesado para el encuestado (Malhotra, 2004).

La inclusión del número de marcas o estímulos específicos debe basarse en el enunciado del problema de investigación, el marco teórico y el buen juicio del investigador.

6.5.2. Recopilación de datos de entrada.

En el escalamiento multidimensional, los datos de entrada obtenidos de los encuestados pueden ser de *percepciones* o *preferencias* (Malhotra, 2004).

6.5.2.1. Datos de percepción directos.

En los enfoques directos para la recopilación de datos de percepción, se pide a los encuestados que, según su criterio, juzguen qué tan semejantes o diferentes son las diversas marcas o estímulos. Usualmente se les pide que califiquen la semejanza de todos los posibles pares de marcas o estímulos en una escala de Likert. Estos datos se llaman juicios de semejanza.

6.5.2.2. Datos de percepción derivados.

Para la obtención de datos de percepción son procedimientos que se basan en los atributos, y que piden a los encuestados que usen escalas Likert o de diferencial semántico, para calificar los estímulos en los atributos identificados.

6.5.2.3. Métodos directos vs. derivados.

Los enfoques directos tienen la ventaja de que el investigador no tiene que identificar un conjunto de atributos sobresalientes. Los encuestados usan su criterio para hacer los juicios de semejanza, como lo harían en circunstancia normales. La desventaja es que los criterios son influidos por las marcas o estímulos evaluados. Si las distintas marcas de automóviles evaluados están en el mismo rango de precios, entonces el precio no surgirá como un factor importante (Malhotra. 2004).

La ventaja del enfoque basado en los atributos es que resulta sencillo identificar a encuestados con percepciones homogéneas. Los encuestados pueden agruparse con base en las calificaciones de los atributos. También es más fácil asignar una etiqueta a las dimensiones. Una desventaja es que el investigador debe identificar todos los atributos sobresalientes, lo cual es una tarea difícil. El mapa espacial obtenido depende de los atributos identificados.

Los enfoques directos se usan con más frecuencia que los enfoques basados en atributos. Sin embargo, sería mejor usar ambos enfoques de manera complementaria.

6.6. Ejemplo de aplicación del análisis multidimensional

El siguiente ejemplo cuenta con una base de datos de preferencias de marcas de automóviles y una base de datos de similitudes (proximidades) de los mismos. Tales bases de datos fueron extraídas del apunte *El análisis del escalamiento multidimensional, Natalia Vila López.* Para confeccionar la base de datos se recogió información sobre la forma en que se relacionan las 18 empresas siguiendo el método de categorización o clasificación. Este método consiste en solicitar a cada uno de los 211 profesionales de un sector industrial encuestados que reparta las 18 empresas en tantos grupos como él considere oportuno, basándose en la competencia que percibe entre ellas. Para tal fin se ha recurrido al empleo de tarjetas, cada una con el nombre de un competidor diferente.

La siguiente tabla resume esa información.

Tabla 6.1: Matriz cuadrada derivada a partir de los datos de categorización.

N°de Marca	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1	89.6	3.3	2.8	2.8	3.3	8.5	3.8	85.8	4.7	3.8	2.8	3.3	17.5	3.3	9.5	20.9	57.3
2		1	1.9	3.3	3.3	1.9	9.5	4.7	90.5	3.8	2.4	1.9	1.4	18	1.9	11.4	14.2	55
3			1	35.1	63.5	76.5	28.9	34.1	3.8	45	69.2	74.4	78.7	33.2	78.2	26.5	42.2	10.4
4				1	50.7	27.5	38.4	83.4	4.3	39.8	24.6	28	26.5	30.8	33.6	35.1	16.1	11.8
5					1	59.2	34.6	48.8	5.2	43.6	51.7	51.7	57.3	35.5	61.6	26.5	34.1	10.4
6						1	28.4	26.5	3.8	49.3	80.6	76.3	83.9	32.2	75.4	29.6	47.4	11.4
7							1	44.1	9	54	33.6	35.5	28.9	42.7	29.4	67.3	36	23.2
8								1	4.3	39.3	24.2	25.1	25.1	28	33.2	36	15.6	10.9
9									1	3.3	3.8	3.3	3.3	15.2	3.8	9.5	12.8	56.4
10										1	53.6	47.9	47.4	40.3	48.8	49.8	41.2	14.2
11											1	78.7	76.3	35.1	71.1	32.2	54	12.8
12												1	78.2	37.4	72.5	30.3	49.3	11.4
13													1	31.8	79.1	25.6	47.9	9.5
14														1	31.8	44.5	40.8	33.6
15															1	29.9	42.2	9
16																1	37.4	24.6
17																	1	32.7
18																		1

Fuente: N. Vila El análisis del escalamiento multidimensional, 2000

Leyenda:
1:Audi 2: BMW 3: Citroen 4: Daewoo 5: Fiat 6: Ford 7: Honda 8: Hyunday
9: Mercedes 10: Nissan 11: Opel 12: Peugeot 13: Renault 14: Rover 15: Seat
16: Toyota 17: Volkswagen 18: Volvo

El porcentaje de veces que dos empresas han sido agrupadas juntas es lo que se conoce como **coeficiente de similitud** o **coeficiente de proximidad**. Así por ejemplo, el coeficiente de similitud entre Audi (1 en la matriz) y BMW (2 en la matriz) es 89,6 %, lo que equivale a afirmar que de los 211 profesionales encuestados, 188 han colocado a Audi y BMW en la misma categoría competitiva (188/ 211 = 0,896), es decir ambas son competitivas en el mercado. Un caso contrario es observar el grado de competitividad que tienen Peugeot (15 en la matriz) con BMW (2 en la matriz), el que resulta ser bastante bajo (1,9 %).

A continuación, se procede a ingresar esta base de datos al programa SPSS y a la interpretación de las salidas que el procesamiento de estos datos arrojó.

Tabla 6.2: Medidas de estrés y ajuste.

Estrés bruto normalizado	,00231
Estrés-I	,04801[a]
Estrés-II	,08758[a]
S-Estrés	,00458[b]
Dispersión contada para (D.A.F.)	,99769
Coeficiente de congruencia de Tucker	,99885

Hay dos tipos de indicadores. Aquellos para los que el cero representa un ajuste perfecto. De este primer tipo serían los indicadores Stress bruto normalizado, Stress-I, Stress-II y S-Stress, que como se observa en la tabla 6.2, son todos bastante cercanos a cero. Aquellos valores cercanos a 1 indican que el ajuste perfecto. Esto es lo muestran los indicadores Dispersión explicada (D.A.F.) y Coeficiente de congruencia de Tucker.

Observando los valores de unos y otros siempre se llega a la misma conclusión: que el ajuste del modelo es bueno o muy bueno en este caso. Esto es así porque el grado de error en los datos (distancias) es muy pequeño. En resumen, los cuatro primeros índices de ajuste deberían ser iguales a 0 y los dos últimos iguales a 1, para tener un buen ajuste; lo que en este caso está muy bien logrado.

En la siguiente tabla de resultados se tiene información de las coordenadas finales para las diferentes marcas. Por ejemplo, en la dimensión 1, los valores 1,206, 1,242 y 1,206 de las marcas Audi, BMW y Mercedes, respectivamente, indican una posición importante en la industria automotriz; en tanto que en la misma dimensión 1, las marcas Citroen y Fiat tienen valores -0,506 y -0,555 respectivamente, es decir, valores muy bajos. Por tanto, se podría inferir que la dimensión 1 está relacionada con el prestigio, el estatus que otorga una marca de automóvil.

Tabla 6.3: Coordenadas finales

	Dimensión	
	1	2
Audi	1,206	-,109
BMW	1,242	,030
Citroen	-,506	-,185
Daewoo	-,448	,512
Fiat	-,555	,122
Ford	-,465	-,254
Honda	-,026	,334
Hyundai	-,392	,534
Mercedes	1,206	-,035
Nissan	-,331	,101
Opel	-,444	-,278
Peugeot	-,459	-,224
Renault	-,498	-,262
Rover	,154	,018
Seat	-,502	-,223
Toyota	,055	,285
Volkswagen	-,016	-,341
Volvo	,779	-,025

Con respecto a la dimensión 2, se tiene que ser un poco más cuidadoso en las interpretaciones. Pero sin ir más lejos, podría tratarse de la relación calidad - precio. Se dejará al alumno sacar conclusiones sobre la dimensión 2.

Las coordenadas finales permiten deducir analíticamente la situación en la que se encuentra de cada marca en cuanto a su nivel de importancia y a su vez sacar algunas conclusiones respecto del significado que tiene cada dimensión. En este último punto, tiene vital importancia la experiencia y/ o conocimiento previo del investigador de mercados. Designar las dimensiones requiere que el investigador haga un juicio subjetivo. Los siguientes lineamientos ayudan a la tarea (Malhotra, 2004).

1. Incluso si se obtienen juicios directos de semejanza, pueden reunirse calificaciones de la marca en atributos proporcionados por el investigador. Gracias a los procedimientos estadísticos, como el de regresión, esos atributos pueden ajustarse en el mapa espacial. Los ejes se denominan luego según los atributos con los que se alinean más de cerca.

2. Después de que han proporcionado los datos de semejanza o preferencia directa, se pide a los encuestados que indiquen los criterios que usaron en sus evaluaciones. Tales criterios pueden relacionarse luego en forma subjetiva al mapa espacial para designar las dimensiones.

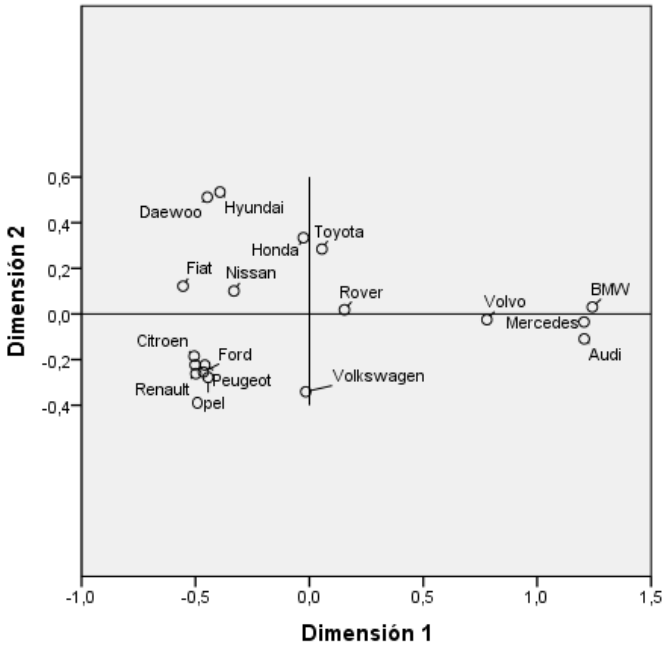
3. De ser posible, puede mostrarse a los encuestados sus mapas espaciales y pedirles que den nombre a las dimensiones al examinar las configuraciones.

4. Si se dispone de características objetivas de la marca (por ejemplo, caballos de fuerza o kilometraje por litro de gasolina en el caso de los automóviles), se recomienda usarlas como ayuda para interpretar las dimensiones subjetivas de los mapas espaciales.

Para complementar lo anterior, es relevante contar con el mapa de espacio común que se muestra en la figura 6.1.

En la figura 6.1 es más cómodo sacar conclusiones de las marcas de automóviles que son competidores inmediatos o más cercanos. Así se tiene que BMW, Mercedes Volvo y Audi son competidores directos, como también Hyundai y Daewoo. Se observa también que existe una competencia intensa entre las marcas Citroen, Renault, Opel, Ford y Peugeot. Cabe señalar que esta encuesta fue hecha en España. En el Continente Sudamericano quizas la encuesta arrojaría resultados muy diferentes.

Figura 6.1: Mapa de espacio común



A menudo las dimensiones representan más de un atributo. La configuración o el mapa espacial pueden interpretarse examinando de las coordenadas y posiciones relativas de las marcas. Por ejemplo, la competencia puede ser mayor entre las marcas localizadas muy cerca entre sí. Una marca aislada tiene una imagen única. Las marcas más alejadas en dirección de un descriptor tienen más fuerza en esa característica. De este modo, se comprenderán las fortalezas y debilidades de cada producto. Las brechas en el mapa espacial pueden indicar oportunidades potenciales para lanzar nuevos productos (Malhotra, 2004).

Para el caso estudiado, los pasos para llevar a cabo el escalamiento multidimensional en el programa SPSS son los siguientes:

1. Elija ANALIZAR de la barra menú de SPSS.
2. Haga clic en ESCALA y luego en ESCALAMIENTO MULTIDIMENSIONAL (PROXSCAL).
3. En el recuadro FORMATO DE DATOS elija LOS DATOS SON PROXIMIDADES y en recuadro NÚMERO DE FUENTES elija UNA FUENTE MATRICIAL y luego hacer clic en botón DEFINIR.
4. Traslade todas las marcas al recuadro PROXIMIDADES y después haga clic en MODELO.

5. En recuadro FORMA elija MAGTRIZ TRIANGULAR SUPERIOR y en recuadro PROXIMIDADES elija SIMILARIDADES. En recuadro TRANSFORMACIÓN DE LAS PROXIMIDADES elija ORDINAL y DESEMPATAR OBSERVACIONES EMPATADAS. Después haga clic en el botón CONTINUAR.

6. En el botón OPCIONES, en el recuadro CONFIGURACIÓN INICIAL usar SIMPLEX y en recuadro CRITERIOS DE ITERACIÓN tiquear USAR ITERACIONES RELAJADAS y después botón CONTINUAR.

7. En GRÁFICOS en recuadro GRÁFICOS tiquear ESPACIO COMÚN y después haga clic en CONTINUAR.

8. En RESULTADOS, en recuadro VISUALIZACIÓN tiquear COORDENADAS DE ESPACIO COMÚN y DIVERSAS MEDIDAS DEL ESTRÉS. Haga clic en CONTINUAR.

9. Seleccione OK.

CAPÍTULO VII

ANÁLISIS CONJUNTO

7.1. Objetivo

Comprender la técnica del análisis conjunto y aplicarla a problemáticas relacionadas con las ciencias de la administración.

7.2. Antecedentes

El Análisis de Conjunto es, precisamente, una técnica estadística que determina qué características de un producto (o servicio) son las preferidas por los consumidores y cuantifica estas preferencias. Las características de un producto incluyen atributos como la marca, el color, formas, precio y garantía y el análisis de conjunto mide las preferencias del consumidor por las características particulares de un producto (Pérez, 2008).

El análisis conjunto es una técnica estadística utilizada en estudios de mercado para determinar cuáles son los gustos o preferencias de los consumidores en relación a un bien o servicio, simulando su proceso de elección. Para ello se descompone el producto en atributos y niveles. Los atributos son sus características como: el tipo de envase, el precio, la marca, el sabor, etc. Los niveles son las alternativas que presenta una característica. Por ejemplo, en el caso del sabor pueden ser: chocolate, vainilla y frutilla. Combinando estos niveles y atributos se generan los productos hipotéticos que son presentados para que los encuestados elijan que opción prefieren. El análisis conjunto permitirá conocer:

- cuáles son las características más valoradas de un producto.
- cuál es la combinación de características que darían lugar al producto ideal.
- cuánto dinero de más está dispuesto a pagar el consumidor, por el cambio en una característica como, por ejemplo, el sabor.
- cuáles son los nichos de consumidores en donde un producto puede tener más éxito.

El análisis conjunto puede aplicarse a cualquier industria o a cualquier área de la administración, como: recursos humanos, economía, procesos productivos, etc.

El análisis de conjunto se basa en la suposición de que los consumidores toman la decisión de compra considerando simultáneamente todas las características del producto. Para hacer esto, los consumidores deben de buscar un equilibrio en términos de la relación calidad-precio, porque normalmente un producto no tiene todas las mejores características. Por ejemplo, el típico coche grande de lujo proporciona un mayor status, seguridad y tamaño, pero también cuesta más y tiene mayor consumo por kilómetro. El análisis de conjunto se utiliza para estudiar estos equilibrio (Pérez, 2008).

Desde mediados de los años 70 el análisis conjunto ha atraído una atención considerable tanto como método que representa de modo realista las decisiones de los consumidores como por los descartes entre multiatributos de productos o servicios. Ha ganado una amplia aceptación y uso en muchas industrias con tasas de utilización que aumentaron hasta diez veces más en los años ochenta (Pérez, 2008)

El análisis conjunto trata de determinar la importancia relativa que los consumidores asignan a los atributos sobresalientes y las utilidades que atribuyen a los niveles de atributos. La información se deriva de las evaluaciones que hacen los consumidores de marcas, o de perfi les de marcas compuestos por esos atributos y sus niveles. A los encuestados se les presentan estímulos que consisten en combinaciones de niveles de atributos. Se les solicita que evalúen esos estímulos en términos de su conveniencia. Los procedimientos conjuntos tratan de asignar valores a los niveles de cada atributo, de manera que los valores resultantes o las utilidades atribuidas a los estímulos concuerden, tanto como sea posible, con las evaluaciones de entrada proporcionadas por los encuestados. La suposición de base es que cualquier conjunto de estímulos, como productos, marcas o tiendas, se evalúa como un paquete de atributos (Malhotra, 2004).

Al igual que el escalamiento multidimensional, el análisis conjunto depende de las evaluaciones subjetivas de los encuestados. Sin embargo, en el EMD los estímulos son productos o marcas. En el análisis conjunto, los estímulos son combinaciones de niveles de atributos determinados por el investigador. La meta del EMD es desarrollar un mapa espacial que describa los estímulos en un espacio multidimensional de percepción o de preferencia. Por otro lado, el análisis conjunto busca desarrollar las funciones de valor parcial o utilidad que los encuestados confieren a los niveles de cada atributo. Las dos técnicas son complementarias (Malhotra, 2004).

7.3. Áreas de aplicación

En las ciencias de la administración y específicamente en la investigación de mercados, se hace análisis conjunto para diversos propósitos, como los siguientes:

1. El número y la naturaleza de las dimensiones que usan los consumidores para percibir diferentes marcas en el mercado.
2. El posicionamiento de las marcas actuales en tales dimensiones.
3. El posicionamiento de la marca ideal de los consumidores en esas dimensiones.

La información que arroja el análisis conjunto tiene variadas aplicaciones en ciencias administrativas, como las siguientes:

- *Medición de imagen.* Comparar las percepciones que se tiene de un conjunto de políticas de recursos humanos de quienes avalan la medida y de quienes no la respaldan con la percepción que tiene el propio departamento de recursos humanos, para identificar discrepancias perceptuales.
- *Desarrollo de nuevos productos.* Buscar discrepancias en el mapa perceptual, lo que indicaría posibles oportunidades para posicionar nuevos productos y marcas actuales a título de prueba, para determinar cómo perciben los clientes esos nuevos conceptos. La proporción de preferencias por cada producto es un indicador de su éxito.
- *Evaluar la eficacia de una campaña política.* Los mapas espaciales sirven también para determinar si la campaña política ha logrado convencer a los potenciales votantes.
- *Elaboración de escalas de opinión.* Las técnicas de EMD sirven para establecer las dimensiones apropiadas y la configuración del espacio de opinión.
- *Análisis de precios.* Pueden compararse mapas espaciales desarrollados con y sin información sobre los precios para determinar su impacto.

7.4. Estadísticas asociadas al análisis conjunto

- *Juicios de semejanza* Los juicios de semejanza son calificaciones en una escala tipo Likert de todos los pares posibles de marcas u otros estímulos en términos de su semejanza.
- *Ordenamientos de preferencia* Los ordenamientos de preferencias son rangos ordenados de las marcas u otros estímulos de los más a los menos preferidos. Por lo general se obtienen de los encuestados.
- *Estrés* Es la falta de ajuste de la medida: los valores más altos de estrés indican un ajuste más pobre.

- *R cuadrada* Es un índice de correlación elevado al cuadrado, que indica la proporción de varianza de los datos escalados en forma óptima, que puede explicarse mediante el procedimiento del EMD. Es una medida de la bondad del ajuste.
- *Mapa espacial* Las relaciones percibidas entre marcas u otros estímulos se representan en relaciones geométricas entre puntos en un espacio multidimensional llamado mapa espacial.
- *Coordenadas* Indican el posicionamiento de la marca o el estímulo en un mapa espacial.
- *Despliegue* Es la representación de marcas y encuestados como puntos en el mismo espacio.

7.5. Pasos para la realización de análisis conjunto

7.5.1. Planteamiento del problema.

Para plantear un problema de análisis conjunto, lo primero que debe hacer el investigador es identificar los atributos y los niveles de tales atributos que se emplearán en la elaboración de los estímulos. Los niveles de los atributos denotan los valores que éstos asumen. Desde un punto de vista teórico, los atributos elegidos deben tener una influencia considerable en la preferencia y elección del consumidor. Por ejemplo, en la elección de una marca de automóvil deben incluirse el precio, el kilometraje por litro de gasolina, el espacio interior, etcétera. Desde una perspectiva administrativa, los atributos y sus niveles deben ser procesables. No resulta útil decirle a un gerente que los consumidores prefieren un auto deportivo a uno de aspecto conservador, a menos que deportivo y conservador se definan en términos de atributos sobre los que el gerente tenga control. Los atributos pueden identificarse a partir de conversaciones con la administración y los expertos en la industria, el análisis de datos secundarios, la investigación cualitativa y encuestas piloto. Un estudio típico de análisis conjunto incluye seis o siete atributos (Malhotra, 2004).

Una vez que se hayan identificado los atributos sobresalientes, deben elegirse sus niveles apropiados. El número de niveles de un atributo determina el número de parámetros que se calculará y también influye en el número de estímulos que los participantes evaluarán. Para minimizar la tarea de evaluación de los encuestados y aun así calcular los parámetros con precisión razonable, es conveniente restringir el número de niveles del atributo. La utilidad o función de valor parcial para los niveles de un atributo puede ser no lineal. Por ejemplo, un consumidor quizá prefiera un carro de tamaño mediano a uno pequeño o uno grande. De la misma manera, tal vez la utilidad del precio no sea lineal. La pérdida de utilidad al

pasar de un precio bajo a uno mediano puede ser mucho menor que la pérdida de utilidad al pasar de un precio mediano a uno alto. En tales casos deben usarse al menos tres niveles. No obstante, algunos atributos se presentan naturalmente en forma binaria (dos niveles): un automóvil tiene o no tiene techo corredizo (Malhotra, 2004).

Los niveles del atributo seleccionados impactarán las evaluaciones de los consumidores. Si el precio de una marca de automóviles varía de $10,000 a $12,000 o $14,000 dólares, el precio será relativamente poco importante. Por otro lado, si el precio varía de $10,000 a $20,000 o a $30,000, será un factor de importancia. Por consiguiente, el investigador debería tomar en consideración los niveles de los atributos que son comunes en el mercado y los objetivos del estudio. Usar niveles del atributo que estén fuera del rango reflejado en el mercado disminuirá la credibilidad de la tarea de evaluación, pero incrementará la precisión con que se calculan los parámetros. La norma general es elegir niveles del atributo de modo que los rangos sean algo más grandes que los que predominan en el mercado, aunque no tan grandes que tengan un impacto adverso en la credibilidad de la tarea de evaluación (Malhotra, 2004).

Para ilustrar el análisis conjunto se considera el problema de cómo evalúan los estudiantes los zapatos deportivos. La investigación cualitativa identificó tres atributos destacados: la suela, la parte superior (pala) y el precio. Como se muestra en la tabla 7.1, cada uno se definió en términos de tres niveles. Tales atributos y sus niveles se usaron para elaborar los estímulos del análisis conjunto. Advierta que, para mantener la simplicidad del ejemplo, se utiliza un número limitado de atributos, es decir, sólo tres (Malhotra, 2004).

Figura 7.1: Atributos y niveles de zapatos deportivos

Atributo	Nivel N°	Descripción
Suela	3	Hule
	2	Poliuretano
	1	Plástico
Parte superior (pala)	3	Lona
	2	Cuero
	1	Nylon
Precio	3	$ 30
	2	$ 60
	1	$ 9o

7.5.2. Composición de los estímulos.

Existen dos procedimientos generales para elaborar los estímulos del análisis conjunto: por pares y de perfiles completos. En el procedimiento por pares, llamado también evaluaciones de dos factores, los encuestados evalúan dos atributos a la vez, hasta que se hayan evaluado todos los pares de atributos posibles (Malhotra, 2004). En el procedimiento de perfiles completos, llamado también evaluaciones de factores múltiples, se construyen perfiles completos de marcas para todos los atributos. Por lo general, cada perfil se describe en una tarjeta separada. En la tabla 7.2 se ilustra este enfoque en el contexto del ejemplo de los zapatos deportivos.

Figura 7.2: Método de perfiles completos para reunir datos conjuntos

Atributo	Descripción del nivel
Suela	Hecha de hule
Parte superior (pala)	Hecha de nylon
Precio	$30

El ejemplo de los zapatos deportivos sigue el procedimiento del perfil completo. Dados tres atributos, definidos cada uno en tres niveles, pueden construirse un total de $3 \times 3 \times 3 = 27$ perfiles. Para reducir la tarea de evaluación del encuestado se utilizó un diseño factorial fraccional y se elaboró un conjunto de nueves perfiles que constituían los conjuntos de estímulos de estimación (véasela tabla 7.3). Para fines de validación se construyó otro conjunto de nueve estímulos. Se obtuvieron datos de entrada tanto para los estímulos de estimación como de validación. Sin embargo, antes de que pudieran obtenerse los datos de entrada fue necesario tomar una decisión sobre su forma.

7.5.3. Elección de la forma de los datos de entrada.

Como en el caso del escalamiento multidimensional, los datos de entrada del análisis conjunto pueden ser métricos o no métricos. Para los datos no métricos, por lo general, se solicita a los participantes que den evaluaciones de rangos ordenados. En la técnica del perfil completo, ordenan todos los perfiles del estímulo. Los ordenamientos implican evaluaciones relativas de los niveles del atributo. Los defensores del ordenamiento de los datos creen que éstos reflejan con precisión la conducta de los consumidores en el mercado.

En la forma métrica los encuestados no brindan ordenamientos sino calificaciones. En este caso, los juicios por lo general se hacen de forma independiente. Los defensores de los datos de calificación creen que son mucho más convenientes para los participantes y que son más fáciles de analizar que los ordenamientos. En años recientes se ha vuelto cada vez más común el uso de calificaciones (Malhotra, 2004).

En el análisis conjunto la variable dependiente suele ser la preferencia o intención de compra. En otras palabras, los encuestados proporcionan calificaciones u ordenamientos en términos de su preferencia o intención de compra. Sin embargo, la metodología conjunta es flexible y tiene cabida para otras variables dependientes que incluyen a la compra o elección reales.

En la evaluación de perfiles de los zapatos deportivos, se solicitó a los encuestados que proporcionaran calificaciones de preferencia para cada zapato descrito por los nueve perfiles en el conjunto de estimación.

Para obtener esas calificaciones se utilizó una escala Likert de nueve puntos (1= no se prefieren, 9= los favoritos). En la tabla 7.3 se muestran las calificaciones obtenidas de un encuestado.

Tabla 7.3: Perfiles de zapatos deportivos y sus puntuaciones

Niveles de atributos[a]				Puntuación de referencia
Perfil N°	Suela	Forro	Precio	
1	1	1	1	9
2	1	2	2	7
3	1	3	3	5
4	2	1	2	6
5	2	2	3	5
6	2	3	1	6
7	3	1	3	5
8	3	2	1	7
9	3	3	2	6

a : Los niveles de atributos corresponden a los de la tabla 7.1

7.5.4. Elección de un procedimiento para el análisis conjunto.

El modelo de análisis conjunto básico se representa con la siguiente fórmula:

$$U(X) = \sum_{i=1}^{m} \sum_{j=1}^{k_i} \alpha_{ij} X_{ij}$$

donde,

$U(X)$ = utilidad general de una alternativa
α_{ij} = contribución del valor parcial o utilidad asociada con el nivel j - *ésimo* (Con j, tal que $j = 1, 2, ..., k_i$) del *i-ésimo* atributo (Con i, tal que $i = 1, 2, ..., m$)
k_i = número de niveles del atributo i
m = número de atributos
X_{ij} = variable artificial, tal que $X_{ij} = 1$ si está presente el *j-ésimo* nivel del *i-ésimo* atributo, o bien $X_{ij} = 0$ en otro caso.

La importancia de un atributo I_i se define en términos del rango de los valores parciales, α_{ij}, en todos los niveles de ese atributo:

$$I_i = \{m\acute{a}x(\alpha_{ij}) - m\acute{i}n(\alpha_{ij})\}, \forall i$$

La importancia del atributo se normaliza para evaluar su significación en relación con otros atributos, W_i

$$W_i = \frac{I_i}{\sum_{i=1}^{m} I_i}$$

de modo que

$$W_i = 1$$

Existen varios procedimientos diferentes para calcular el modelo básico. El más sencillo, cuya popularidad va en aumento, es la regresión con variables ficticias (*dummy*). En este caso, las variables predictivas son las variables ficticias para los niveles del atributo. Si un atributo tiene k_i niveles, se codifica en términos de $k_i - 1$ variables ficticias. Si se obtienen datos métricos, las calificaciones constituyen la variable dependiente, siempre que se ajusten a una escala de intervalo. Si los datos son no métricos, los ordenamientos pueden convertirse en 0 ó 1 realizando comparaciones pareadas entre las marcas. En este caso, las variables predictivas representan las diferencias en los niveles de los atributos de las marcas comparadas (Sandor y Wedel, 2001).

El investigador también debe decidir si los datos se analizarán a nivel del encuestado individual o a nivel conjunto. A nivel individual los datos de cada participante se analizan por separado. Si el análisis se va a realizar a nivel conjunto, debe idearse algún procedimiento para agrupar a los encuestados. Un enfoque común consiste en calcular primero las funciones de valor parcial o de utilidad a nivel individual. Los encuestados se agrupan luego con base en la semejanza de sus valores parciales. Después se realiza el análisis conjunto para cada conglomerado. Debe especificarse un modelo apropiado para estimar los parámetros (Arora y Allenby, 1999).

Para analizar los datos presentados en la tabla 7.3 se utilizó la regresión de mínimos cuadrados (RMC) ordinarios con variables ficticias. La variable dependiente fueron las calificaciones de preferencia.

Tabla 7.4: Datos de zapatos deportivos codificados para una regresión

Calificaciones de preferencia	Atributos					
	Suela		Parte superior (pala)		Precio	
Y	X_1	X_2	X_3	X_4	X_5	X_6
9	1	0	1	0	1	0
7	1	0	0	1	0	1
5	1	0	0	0	0	0
6	0	1	1	0	0	1
5	0	1	0	1	0	0
6	0	1	0	0	1	0
5	0	0	1	0	0	0
7	0	0	0	1	1	0
6	0	0	0	0	0	1

Las variables independientes o predictivas fueron seis variables ficticias, dos para cada variable. En la tabla 7.4 se muestran los datos transformados. Como los datos conciernen a un solo participante, se realizó un análisis a nivel individual. En la tabla 7.5 se presentan las funciones de valores parciales o de utilidad calculadas para cada nivel, así como la importancia relativa de los atributos (Malhotra, 2004).

Tabla 7.5: Resultados del análisis conjunto

Atributo	Nivel		Utilidad	Importancia
	N°	Descripción		
Suela	3	Hule	0,778	
	2	Poliuretano	- 0,556	
	1	Plástico	- 0,222	0,286
Parte superior (pala)	3	Cuero	0,445	
	2	Lona	0,111	
	1	Nylon	- 0,556	0,214
Precio	3	30 dólares	1,111	
	2	60 dólares	0,111	
	1	90 dólares	- 1,222	0,500

El modelo estimado se presenta como:

$$U = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6$$

donde:

X_1, X_2 = Variables ficticias asociadas a los niveles del atributo suela
X_3, X_4 = Variables ficticias asociadas a los niveles del atributo parte superior
X_5, X_6 = Variables ficticias asociadas a los niveles del atributo precio

En cuanto a la suela, los niveles de este atributo se codificaron como sigue:

Nivel	X_1	X_2
1: Hule	1	0
2: Poliuretano	0	1
3: Plástico	0	0

En la primera fila los valores "1" y "0" significa que la suela está hecha de hule (1) y no de poliuretano (0). En la última fila los valores "0" y "0" significa que la suela no estaría hecha hule ni de poliuretano y por tanto estaría conformada por el tercer nivel; en este caso plástico.

Los niveles de los otros atributos se codificaron de manera semejante.

Para la obtención de los estimadores paramétricos, se procesaron los datos de la tabla 7.4 en Excel y haciendo uso del paquete *análisis de datos*, que provee dicho programa, se efectuó la regresión lineal, obteniendo los siguientes resultados:

$$b_0 = 4,222$$
$$b_1 = 1,000$$
$$b_2 = -0,333$$
$$b_3 = 1,000$$
$$b_4 = 0,667$$
$$b_5 = 2,333$$
$$b_6 = 1,333$$

Dada la codificación de las variables ficticias, donde el nivel 3 es el básico, los coeficientes pueden relacionarse con los valores parciales. Cada coeficiente de las variables ficticias representa la diferencia en el valor parcial de ese nivel menos el valor parcial del nivel básico. En el caso de la suela, se tiene lo siguiente:

$$\alpha_{11} - \alpha_{13} = b_1$$
$$\alpha_{12} - \alpha_{13} = b_2$$

Para resolver los valores parciales es necesaria una restricción adicional. Los valores parciales se calcularon en una escala de intervalo, por lo que el origen es arbitrario. Por ende, la restricción adicional que se impone es de la forma:

$$\alpha_{11} + \alpha_{12} + \alpha_{13} = 0$$

Estas ecuaciones para el primer atributo, la suela, son:

$$\alpha_{11} - \alpha_{13} = 1,000$$
$$\alpha_{12} - \alpha_{13} = -0,333$$
$$\alpha_{11} + \alpha_{12} + \alpha_{13} = 0$$

Resolviendo las ecuaciones, se obtiene:

$$\alpha_{11} = 0,778$$
$$\alpha_{12} = -0,556$$
$$\alpha_{13} = -0,222$$

Los valores parciales para los demás atributos reportados en la tabla 7.5 se calculan de forma similar. Para la parte superior tenemos:

$$\alpha_{21} - \alpha_{23} = b_3$$
$$\alpha_{22} - \alpha_{23} = b_4$$
$$\alpha_{21} + \alpha_{22} + \alpha_{23} = 0$$

Para el tercer atributo, el precio, se tiene:

$$\alpha_{31} - \alpha_{33} = b_5$$
$$\alpha_{32} - \alpha_{33} = b_6$$
$$\alpha_{31} + \alpha_{32} + \alpha_{33} = 0$$

Al resolver estos dos sistemas, por separado se obtuvieron los siguientes valores:

$$\alpha_{21} = 0,445$$
$$\alpha_{22} = 0,111$$
$$\alpha_{23} = -0,556$$

$$\alpha_{31} = 1,111$$
$$\alpha_{32} = 0,111$$
$$\alpha_{33} = -1,222$$

Valores que ya se registran en la tabla 7.5.

La importancia relativa de los pesos se calculó con base en los rangos de los valores parciales, de la siguiente manera:

$$S = \sum_{i=1}^{3} I_i$$
$$S = (\alpha_{11} - \alpha_{12}) + (\alpha_{21} - \alpha_{23}) + (\alpha_{31} - \alpha_{33})$$

en donde:

S = suma de extensiones de los valores parciales e $I_i = \{máx(\alpha_{ij}) - mín(\alpha_{ij})\}, \forall i$
Luego:

$$S = (0,778 - (-0,556)) + (0,445 - (-0,556)) + (1,111 - (-1,222)) = 4,668$$

$$\text{Importancia relativa de suela} = \frac{I_1}{S} = \frac{0,778 - (-0,556)}{4,668} = \frac{1,334}{4,668} = 0,286$$

$$\text{Importancia relativa de pala} = \frac{I_2}{S} = \frac{0,445 - (-0,556)}{4,668} = \frac{1,001}{4,668} = 0,214$$

$$\text{Importancia relativa de precio} = \frac{I_3}{S} = \frac{1,111 - (-1,222)}{4,668} = \frac{2,333}{4,668} = 0,500$$

El cálculo de valores parciales y los pesos de la importancia relativa proporcionan la base para interpretar los resultados.

7.5.5. Interpretación de los resultados.

Para interpretar los resultados se utiliza la tabla 7.5. Como se observa, en la evaluación de los zapatos deportivos este encuestado mostró mayor preferencia por las suelas de hule. La segunda preferencia fue por las suelas de plástico y la menos preferida fue la de poliuretano. Prefería que la parte superior fuera de cuero. Seguida de la de lona y nylon. Como se esperaba, el precio de $30 tenía la mayor utilidad, y el de $90 la menor. Los valores de la utilidad reportados en la tabla 7.5 sólo tienen propiedades de la escala de intervalo y su origen es arbitrario. En términos de la importancia relativa de los atributos, vemos que el precio es el primero, seguido de la suela y luego por la parte superior. Dado que el precio es por mucho el atributo más importante para este participante, podría designarse como sensible al precio.

7.5.6. Evaluación de la confiabilidad y la validez.

Existen varios procedimientos para evaluar la confiabilidad y la validez de los resultados del análisis conjunto (Andrews, R. L., 2002). A continuación se dan algunas recomendaciones:

1. Debe evaluarse la bondad del ajuste del modelo estimado. Por ejemplo, si se usa la regresión con variables ficticias, el valor de R^2 indicará el grado en que el modelo se ajusta a los datos. Hay que dudar de los modelos con mal ajuste.

2. En una etapa posterior de la entrevista, se pide a los participantes que evalúen de nuevo algunos estímulos seleccionados. Los dos valores de esos estímulos se correlacionan luego para evaluar la confiabilidad de la primera y segunda prueba.

3. Las funciones estimadas de los valores parciales permiten predecir las evaluaciones de los estímulos de retención o de validación. Las evaluaciones pronosticadas pueden luego correlacionarse con las obtenidas de los encuestados para determinar la validez interna.

4. Si se ha realizado un análisis a nivel conjunto, la muestra de estimación puede dividirse de varias formas y en cada submuestra es factible realizar un análisis conjunto. Los resultados se comparan entre las submuestras para evaluar la estabilidad de las soluciones del análisis conjunto.

Al efectuar el análisis de regresión sobre los datos de la tabla 7.5 se obtuvo un coeficiente de determinación R^2 igual a 0,934, lo que indica un buen ajuste. El coeficiente de correlación fue de 0,967, lo que indica que en las seis variables en conjunto están altamente correlacionadas con la preferencia.

Como en este caso se trata sólo de las preferencias dadas por un solo individuo, para estudios más completos y acabados se recomienda realizar este cuestionario a una muestra superior a 30, de un segmento específico de una población. Ideal sería realizar un análisis discriminante previo para llevar a cabo un análisis conjunto.

Andrews, R. L., (2002). *Hierarchical Bayes Versus Finite Mixture Conjoint Analisis: A Comparison of Fit, Prediction and Partworth Recovery*, Journal of Marketing Research, pp. 87-98, 39, núm. 1.

Arora, N., Allenby, G.M., (1999). *Measuring the Influence of Individual Preference Structures in Group Decision Making*, Journal of Marketing Research, pp. 476-487, 36, núm. 4.

De la Garsa, J., Morales, B., González, B.(2013). *Análisis Estadístico Multivariante: Un enfoque teórico y práctico*, Editorial McGraw Hill, México.

Guisande, C., Bahamonde, A., Barreiro, A.(2011). *Tratamiento de Datos con R, Statistica y SPSS*, Editorial Díaz de Santos, España.

Johnson, D. (2000). *Métodos Multivariados Aplicados al Análisis de Datos*, Editorial Thomson, México.

Malhotra, N.(2004). *Investigación de Mercados: Un enfoque aplicado*, Editorial Pearson, México.

Montgomery, D.(1991). *Diseño y Análisis de Experimentos*, Grupo Editorial Iberoamericana.

Pedret, R., Sagnier, L., Camp, F.(2000). *Herramientas para Segmentar Mercados y Posicionar Productos*, Editorial Deusto S. A., España.

Peña, D. (2002). *Análisis de Datos Multivariantes*, Universidad Carlos III de Madrid, España.

Pérez, C. (2008). *Técnicas de Análisis Multivariante de Datos: Aplicaciones con SPSS*, Editorial Pearson, España.

Sandor, Z., Wedel, M., (2001). *Designing Conjoint Choice Experiments Using Managers Prior Beliefs*, Journal of Marketing Research, pp. 430-444, 38, núm. 4.

Vila, N. (2000). *El análisis del escalamiento multidimensional*, Universidad de Valencia, España.