EDITORIAL DIGITAL
TECNOLÓGICO DE MONTERREY

ANÁLISIS MULTIVARIANTE PARA LA INTELIGENCIA DE MERCADOS

PILAR ESTER
ARROYO LÓPEZ
JULIO CÉSAR
BORJA MEDINA

Primera edición
De venta en: Amazon Kindle, Apple Books, Google Books y Amazon.
Fragmento editado, diseñado, publicado y distribuido por el Instituto Tecnológico y de Estudios Superiores de Monterrey.

Ave. Eugenio Garza Sada 2501 Sur Col. Tecnológico C.P. 64849 |
Monterrey, Nuevo León | México.

Editorial

El Tecnológico de Monterrey presenta su colección de eBooks de texto para programas de nivel preparatoria, profesional y posgrado. En cada título, nuestros autores integran conocimientos y habilidades, utilizando diversas tecnologías de apoyo al aprendizaje.

El objetivo principal de este sello editorial es el de divulgar el conocimiento y experiencia didáctica de los profesores del Tecnológico de Monterrey a través del uso innovador de la tecnología. Asimismo, apunta a contribuir a la creación de un modelo de publicación que integre en el formato eBook, de manera creativa, las múltiples posibilidades que ofrecen las tecnologías digitales.

Con la Editorial Digital, el Tecnológico de Monterrey confirma su vocación emprendedora y su compromiso con la innovación educativa y tecnológica en beneficio del aprendizaje de los estudiantes.

www.ebookstec.com

ebookstec@itesm.mx

Autores

Pilar Ester
Arroyo López

Es profesora de planta en el Tecnológico de Monterrey, Campus Toluca. Doctora en Administración, programa EGADE Campus Ciudad de México y Universidad de Texas en Austin, 1997. Quince años de experiencia docente como profesor en la Escuela de Negocios, instructor de los cursos Estadística Administrativa I y II, Investigación de mercados cuantitativa, Análisis multivariante y Seminario de Inteligencia de mercados. Actualmente profesor titular en la Escuela de Ingeniería, instructor del curso Diseño de experimentos a nivel licenciatura y Métodos estadísticos a nivel maestría. Autora de tres libros destinados a la docencia publicados por campus Toluca y coautora de otros dos más publicados por editorial Limusa. Coautora de tres libros de investigación sobre prácticas de comercio electrónicos en México; la Dra. Arroyo ha escrito también cuatro capítulos en libros de investigación y cinco casos estudio que están publicados en el Centro Internacional de Casos del Tecnológico de Monterrey. Autora de 17 artículos académicos publicados en revistas indizadas o arbitradas y de más de 40 artículos en memorias de congresos y revistas de difusión. Dirección de dos tesis de doctorado y de quince tesis de maestría. Líder académico de la cátedra de investigación “Logística de la cadena de suministros” y miembro del Sistema Nacional de Investigadores de México.

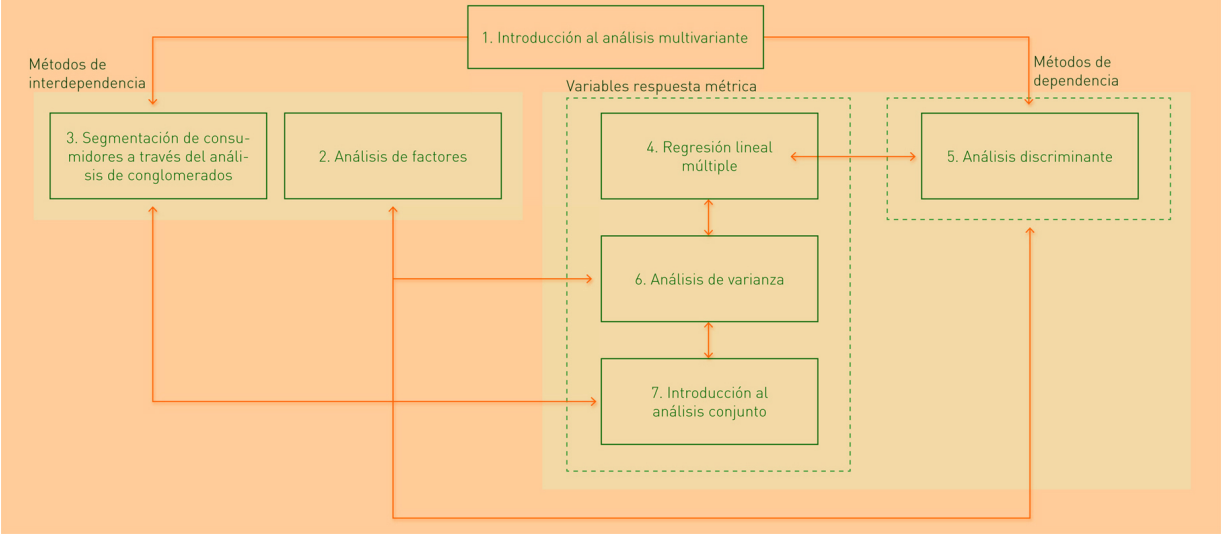
Julio César
Borja Medina

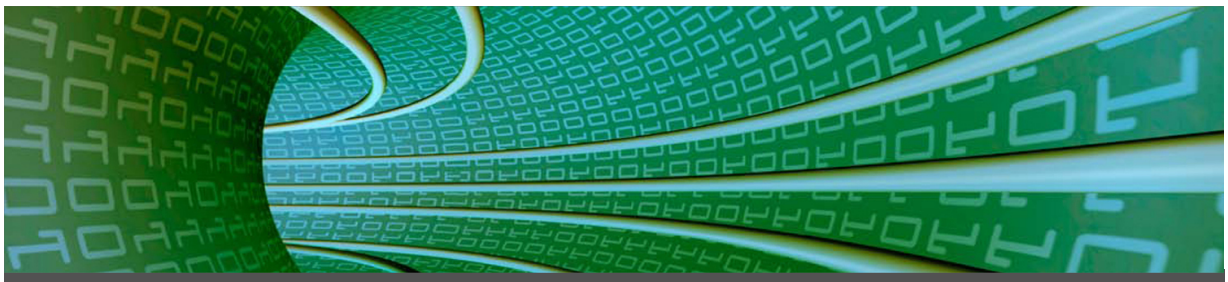
Es profesor del Tecnológico de Monterrey, Campus Toluca. Doctor en Administración por el ITESM Campus Ciudad de México. Profesor del Tec de Monterrey desde 1987; Es Maestro en Ciencias por el Centro de Investigación Científica y Educación Superior de Ensenada (CICESE); Ingeniero Físico Industrial por el Tecnológico de Monterrey, Campus Monterrey.

- Instructor del curso de “Análisis Multivariante”, “Pronósticos para la Toma de Decisiones”, “Estadística Administrativa”, “Seminario de Administración Estratégica” y “Seminario de Estrategia Internacional” a nivel licenciatura. En la Maestría en Administración ha impartido el curso de “Estrategia Competitiva, diseño y modelación de la organización”.

- Ha participado como codirector de tesis en la maestría en Ingeniería Industrial y participado como ponente en congresos nacionales. También ha participado como consultor para organizaciones gubernamentales y empresas nacionales.

Mapa de contenidos





Introducción del eBook

En esta era de la información, el interés de las empresas por generar datos para conocer mejor a su mercado va en aumento. La gran cantidad de información disponible y el desarrollo de software amigable para analizar los datos recolectados contribuyen a la creación de un sistema de inteligencia de mercados dinámico y accesible a todos los tomadores de decisiones. Un componente esencial de tal sistema es el módulo de procesamiento de datos, el cual incluye el uso de técnicas estadísticas que analizan de manera simultánea al conjunto de variables que describen a una empresa, un cliente, un producto o servicio.

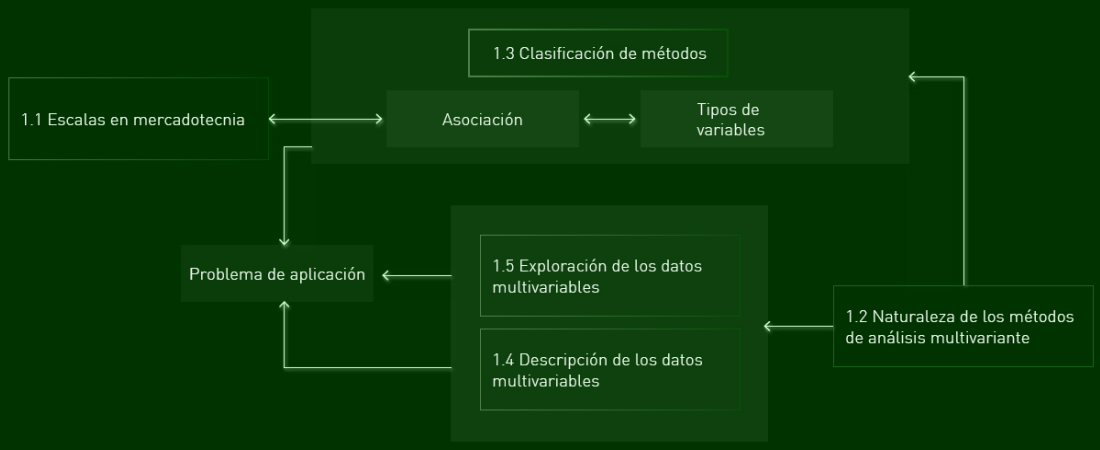
Conocer en qué consisten los métodos de análisis multivariable y cómo pueden aprovecharse para la efectiva toma de decisiones son aspectos importantes en la formación del profesionista en mercadotecnia. Este eBook tiene como propósito proporcionar a este futuro profesionista los conocimientos básicos sobre cinco métodos de estadística multivariada; múltiples ejemplos ilustran las aplicaciones de estas técnicas multivariadas para el diseño de escalas de medición, la segmentación de mercados, la clasificación de individuos y la construcción de modelos descriptivos y predictivos.

Este es un eBook introductorio al análisis multivariado que asume conocimientos básicos sobre métodos para la

recolección de datos, estadística descriptiva e inferencial. Como estos requisitos son cubiertos en cualquier programa de licenciatura en negocios o ingeniería, el eBook puede ser utilizado por una amplia audiencia.

Capítulo 1
Introducción al análisis multivariante

Organizador temático



1. Introducción al análisis multivariante

La inteligencia de mercados tiene como propósito reunir, clasificar y distribuir información oportuna y confiable sobre el mercado y el desempeño de la empresa dentro de éste. La información recolectada, debidamente procesada y analizada facilita la interacción entre personas, la sistematización de hechos, el estudio de relaciones y la comprensión de situaciones apoyando efectivamente el proceso de toma de decisiones.

La información en que se apoya la inteligencia de mercados proviene de cuatro fuentes principales: la propia empresa, su cadena de abastecimiento, la competencia y el consumidor. Cada una de estas fuentes de datos se asocia a una actividad de análisis para el mercado según se describe a continuación:

Dado que las actividades anteriores involucran el uso de información, la inteligencia de mercados requiere contar con procedimientos para la organización y análisis de datos y la estimación de modelos de asociación entre variables; además de mecanismos de consulta que permitan tomar decisiones en cuanto a precio, productos tanto nuevos como existentes, su distribución y promoción. El diagrama de la Figura 1.1 describe gráficamente las actividades que comprenden la inteligencia de mercados.

Figura 1.1 Componentes del sistema de inteligencia de mercados

Este eBook tiene como propósito principal describir y demostrar la aplicación de los métodos estadísticos que permiten analizar la información generada en las varias actividades de inteligencia de mercados.

¿Sabías qué?

Las agencias de investigación de mercados contratan estadísticos para apoyar sus actividades de consultoría. Esto es porque en general en todas las áreas de estudio se tiene bien identificada la necesidad de contar con estadísticos capacitados para ofrecer consultoría así como usuarios con distintos perfiles con conocimientos y habilidades para aplicar la estadística.

1.1 Escalas de mercadotecnia

La información que se genera en mercadotecnia puede involucrar la medición de variables tangibles, pero también de conceptos complejos que conllevan un alto grado de abstracción. Por ejemplo, si el interés son las ventas de un producto, esta variable se puede representar con relativa facilidad a través de medidas monetarias o número de productos vendidos. Sin embargo, también existen conceptos complejos tales como satisfacción, calidad del servicio, lealtad o imagen de la empresa. Estos conceptos requieren ser medidos para lograr el conocimiento deseado sobre el mercado de la empresa y diseñar una estrategia de mercadotecnia. Sin embargo la tarea de representarlos no es simple y requiere de un proceso estandarizado para asignar símbolos (generalmente números) a las características o propiedades de personas, objetos, eventos y situaciones, tal que estos símbolos tengan la misma relación relevante entre ellos que los objetos a los que representan. Estas ideas se ilustran a través del siguiente ejemplo:

En el diseño de campañas de mercadotecnia social, como es el caso de aquellas encaminadas a incrementar el reciclaje de papel, se ha observado que el interés del individuo por apoyar tales campañas depende de varios factores, entre ellos su nivel de escolaridad y su actitud hacia el reciclaje.

La variable “grado de escolaridad de un individuo” se puede definir como el máximo nivel de educación formal que haya alcanzado (certificado o título que lo avale) en una institución reconocida por la Secretaría de Educación Pública de México. Esta definición operacional deriva en la siguiente escala ordinal; esto es, una escala en donde los números más altos corresponden a un nivel educativo mayor, sin embargo la distancia entre estos números no es la misma:

1) Primaria

2) Secundaria

3) Enseñanza media superior

4) Enseñanza superior

5) Diplomado o especialidad

6) Posgrado

Por otra parte la “actitud hacia el reciclaje” es un concepto complejo que requiere capturar o representar las percepciones y sentimientos del individuo hacia esta actividad. Autores como Do

Valle et al. (2005) y Tonglet et al. (2007) han medido esta actitud utilizando múltiples reactivos (ítems) con un mismo formato, por ejemplo la siguiente multi-escala, extraída de los citados trabajos de Do Valle et al. y Tonglet et al., tiene cinco reactivos todos en una escala Likert.

	Totalmente de acuerdo				Totalmente en desacuerdo
El reciclaje de productos ayuda a ahorrar energía (luz y gas)	1	2	3	4	5
El reciclaje de productos ayuda a proteger y crear un mejor ambiente para las generaciones futuras	1	2	3	4	5
El reciclaje ayuda a reducir la cantidad de desechos que se va a los vertederos (tiraderos de basura)	1	2	3	4	5
El reciclaje ayuda a preservar los recursos naturales.	1	2	3	4	5
Los programas de reciclaje son un gasto de recursos	1	2	3	4	5

Las respuestas de un individuo para cada reactivo se pueden agregar (por ejemplo mediante el cálculo de un promedio) para obtener un único puntaje que engloba la actitud del individuo; tal puntaje corresponde a lo que se denomina una escala de **intervalo**. La Tabla 1.1 describe los cuatros tipos de escalas en que pueden representarse aquellas variables de interés para la inteligencia de mercados.

Escala	Caracteristicas	Ejemplos
Nominal	Los numeros identifican o clasifican objetos	Marcas, genero, tipo de canal de distribution
Ordinal	Los numeros indican una posicion relativa pero no la magnitud de la diferencia entre ellos	Preferencia por productos o mar¬cas Niveles socioeco- nomicos
Intervalo	Las diferencias entre elementos se pueden comparar, el cero es arbitrario	Escalas de actitud y opinion Escalas semanticas

		Numeros Índice
Razon	Las diferencias entre elementos se pueden comparar, el cero es arbitrario	Edad, ingreso neto, costos, ventas, participacion de mercado

Tabla 1.1 Tipos de escala en mercadotecnia.

El tipo de escala que se use para representar una variable es importante para decidir cuál método estadístico aplicar. Por ejemplo, para construir un modelo de regresión que permita describir el comportamiento de una variable (dependiente) en términos de una o más variables explicativas (variables independientes) se requiere que todas las variables de interés estén en una escala al menos de intervalo. El análisis de regresión lineal múltiple es un método de los métodos de estadística multivariante ya que involucra el análisis de múltiples variables.

Las variables de nivel educativo y actitud hacia el reciclaje son unidimensionales; sin embargo, hay conceptos que dada su naturaleza involucran más de una dimensión. Un ejemplo es el caso del concepto “calidad del servicio” para el cual Parasuraman et al. (1987) identificaron cinco subdimensiones: tangibles, confiabilidad, respuesta, empatía y compromiso. Estas subdimensiones son relativamente independientes entre sí y buscan englobar todo lo que el concepto representa. Parasuraman et al. diseñaron el instrumento de calidad de servicio conocido como SERVQUAL el cual incluye 30 reactivos, todos en una escala tipo Likert de 7 categorías. En la Tabla 1.2 se describen las dimensiones que integran el concepto y se proporcionan ejemplos de los reactivos sugeridos para evaluar cada una de ellas.

Dimensión	Descripción	Reactivo de ejemplo
Tangibles	Representa las instalaciones físicas, equipo y apariencia del personal que presta el servicio	El equipo con que cuenta esta compañía es moderno Total acuerdo 1 2 3 4 5 Total desacuerdo
Confiabilidad	Habilidad para realizar el servicio prometido de forma segura y adecuada	Cuando la empresa asegura que va a hacer algo en cierto tiempo, realmente lo cumple Total acuerdo 1 2 3 4 5 Total desacuerdo
Respuesta	Interés por ayudar a los clientes y darles pronta respuesta a sus peticiones de servicio	Los empleados en XYZ están siempre dispuestos a atender a los clientes Total acuerdo 1 2 3 4 5 Total desacuerdo
Empatía	Intención individual y preocupación por el cliente	Los empleados en XYZ son siempre corteses con los clientes Total acuerdo 1 2 3 4 5 Total desacuerdo
Compromiso	Empleados que entienden necesidades y transmiten confianza	Los empleados en XYZ son siempre corteses con los clientes Total acuerdo 1 2 3 4 5 Total desacuerdo

Tabla 1.2 Composición de la multiescala de calidad del servicio SERVQUAL.

¿Sabías qué?

La satisfacción del cliente es uno de los conceptos más estudiados en mercadotecnia pero su medición, especialmente

cuando se hace a través del autoreporte de los clientes, representa múltiples retos metodológicos. El diseño de índices de satisfacción se apoya en los métodos de estadística multivariable, aprende más sobre esto leyendo el siguiente artículo, disponible a través de la base de datos de revistas académicas latinoamericanas Redalyc.

Arroyo-López, P., Carrete-Lucero, L. y García-López, S. (2008). Construcción de un índice de satisfacción para clientes de supermercados mexicanos. Contaduría y Administración UNAM, No. 225, pp. 59-78. Recuperado de: http://redalyc.uaemex.mx/principal/ListaArticulosPorNombreAutor.jsp?aut=171980186191,120733,70431

Cuando un concepto es multidimensional no resulta conveniente reducirlo a un único puntaje, sino representarlo a través de múltiples respuestas; cada respuesta corresponde a alguna de las dimensiones o componentes que integran el concepto. Mientras los métodos estadísticos básicos asumen una única variable de interés, los métodos de análisis multivariante permiten trabajar simultáneamente con conjuntos de respuestas relevantes que describen ya sea conceptos complejos o las características de un objeto, individuo o empresa. Respecto a este último caso, considera el caso de un producto de consumo cualquiera, por ejemplo uno de higiene personal como un champú. Para el consumidor, este producto está representando por un conjunto de atributos que determinan su atractivo y posible selección, entre ellos su precio, aroma, consistencia, cualidades hidratantes y limpiadoras, y prestigio de la marca que lo respalda; el producto queda descrito entonces por seis variables interrelacionadas entre sí. Cuando se trabaja con múltiples respuestas, los métodos apropiados para analizar los datos recolectados son los de estadística multivariante, según establece la siguiente definición, la cual ofrece una idea general del objetivo asociado a estos métodos:

Definición 1. El análisis multivariante se refiere a aquellas técnicas estadísticas que permiten estudiar más de una respuesta de interés. Las variables que conforman esta

respuesta múltiple pueden estar correlacionadas entre ellas, tomándose en cuenta tales dependencias cuando se analizan los datos.

Consideremos el caso de una empresa que desea comparar un nuevo producto de champú, con el de mayor participación en el mercado. Los dos productos van a ser comparados respecto a las percepciones de los consumidores en cuanto a los atributos antes mencionados; en términos de inferencias estadísticas este problema corresponde a la comparación de dos medias. En el caso univariado, la prueba estadística básica para comparación de dos medias para muestras independientes se formula como: H_0: $\mu_1=\mu_2$. Si sólo hay una única variable de interés y ésta sigue la distribución normal, la hipótesis anterior se prueba utilizando la t-Student como estadístico de prueba, pero ¿qué pasa cuando las medias a contrastar incluyen múltiples variables que corresponden a los

atributos de un producto o dimensiones de un concepto? La extensión de las inferencias estadísticas básicas al caso de múltiples variables fue la motivación para el desarrollo del análisis multivariante, según lo expresa la siguiente definición propuesta sugerida por Anderson (1984):

> *Definición 2.* El análisis multivariante clásico es aquel análisis estadístico que pretende extender al caso de observaciones sobre varias variables las ideas y procedimientos estadísticos que han demostrado su efectividad en el caso univariado.

Bajo el supuesto de normalidad, el problema de comparación de dos medias planteado antes (H_0: $\mu_1=\mu_2$) utiliza como estadístico de prueba a la T-Hotelling que resulta ser la generalización al caso de la prueba de hipótesis univariada para contrastar dos medias. Como los métodos multivariantes toman en cuenta la interrelación entre las varias respuestas que en este ejemplo describen al producto champú, la complejidad de las técnicas se incrementa en relación a las técnicas univariadas.

Los métodos de Regresión Lineal Múltiple (RLM), Análisis Canónico y Análisis de Varianza Multivariante (MANOVA) que se presentan en este eBook, son técnicas multivariantes que se ajustan a este enfoque clásico y permiten la construcción de modelos estadísticos lineales que involucran múltiples variables.

Actividad de repaso

1.1 Escalas de mercadotecnia

1. El administrador de la cafetería de tu universidad desea evaluar la satisfacción de sus clientes con los servicios y paquetes que ofrece en el horario de comida. ¿Cómo medirías el concepto de "satisfacción"? ¿Es este un concepto multidimensional?

2. Ante las recomendaciones de la Secretaría de Salud por promover el consumo de alimentos saludables para combatir la obesidad y el sobrepeso, las cafeterías de tu Universidad han decidido incluir jugos y bebidas preparadas saludables en su menú. Para conocer las preferencias y actitud de los estudiantes hacia estas bebidas, se diseñaron las siguientes escalas. Selecciona, en cada caso, el tipo de escala en que se expresará la información (nominal, ordinal, intervalo o razón)

2.1. Indica la frecuencia con la que compras bebidas (de cualquier tipo) en esta cafetería

_____ Más de una vez al día

_____ Entre una a siete veces por semana

_____ Menos de una vez por semana

2.2. Marca aquella bebida que te gusta consumir más

____ Jugos de frutas envasados

____ Refrescos

____ Bebidas rehidratantes

____ Agua natural

____ Agua de frutas

____ Jugos naturales

____ Otra

2.3. ¿Cuánto dinero gastas semanalmente en la compra bebidas en la cafetería de tu Universidad?

2.4. Reporta el nivel de acuerdo con cada una de las siguientes declaraciones marcando sobre la escala correspondiente.

Las bebidas con gas son más dañinas para la salud que otras bebidas envasadas.

Totalmente de acuerdo 1 2 3 4 5 Totalmente en desacuerdo
Los productos rehidratantes (Gatorade, Energizer, etc.) sólo deben consumirse cuando haces ejercicio.

Totalmente de acuerdo 1 2 3 4 5 Totalmente en desacuerdo
El agua natural es la mejor bebida que puede seleccionar una persona.

Totalmente de acuerdo 1 2 3 4 5 Totalmente en desacuerdo
Los jugos de frutas naturales no sólo proporcionan el agua necesaria sino también calorías y vitaminas.

Totalmente de acuerdo 1 2 3 4 5 Totalmente en desacuerdo

Respuestas:

1. Con una multiescala en la cual el cliente exprese hasta dónde los productos o servicios de la empresa cubren o incluso exceden sus expectativas.

Sí porque de acuerdo con la literatura de mercadotecnia hay varios componentes asociados a las características del producto o servicio que contribuyen a la satisfacción.

2.1 Ordinal

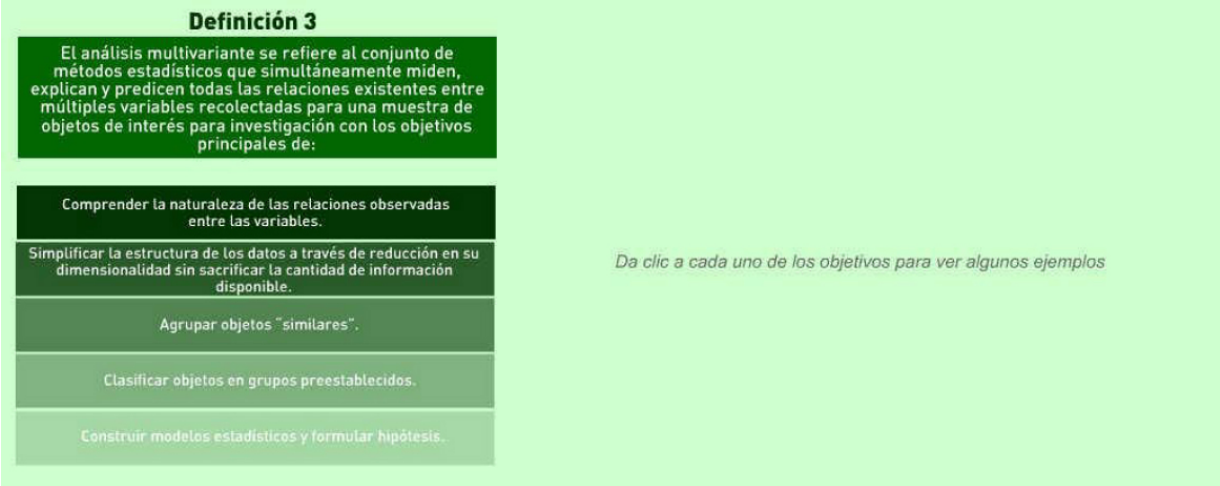
2.2 Nominal

2.3 Razón

2.4 Intervalo

1.2 Naturaleza de los métodos de análisis multivariante

Las definiciones anteriores ofrecen solo ideas generales sobre lo que tratan las técnicas de estadística multivariante, la siguiente definición, adaptada de Hair et al. (2002) e Iglesias-Antelo y Sulé-Alonso (2003) precisa los objetivos de los métodos de análisis multivariante:



Actividad de repaso

1.2 Naturaleza de los métodos de análisis multivariante

a. Distinguir a los clientes de las aerolíneas de bajo precio de los clientes de otras aerolíneas en términos de su estilo de vida y los beneficios que buscan de un servicio de transporte aéreo.

b. Evaluar tres diseños de empaque de helado en los cuales se han modificados colores, material y mecanismos de cierre según el nivel de practicidad y seguridad que ofrece a los consumidores.

c. Identificar los componentes relevantes al concepto “calidad de vida” entendido no sólo como capacidad de compra sino como bienestar global para un individuo y su familia.

d. Predecir cuál marca estadounidense de automóvil (Ford, Chrysler y GM) elegirá un comprador según sus percepciones sobre la calidad técnica, de diseño y de servicio de cada marca.

e. Expresar mediante un conjunto reducido de índices las percepciones de los usuarios sobre facilidad de uso, utilidad percibida y aplicabilidad en distintos contextos de un nuevo software CRM (CRM por sus siglas en inglés Customer Relation Management, en español se traduce como Administración de Relaciones con los Clientes).

Respuestas

a. Discriminación entre grupos y clasificación.

b. Construcción de modelos en los que los atributos del producto expliquen los niveles de practicidad y seguridad.

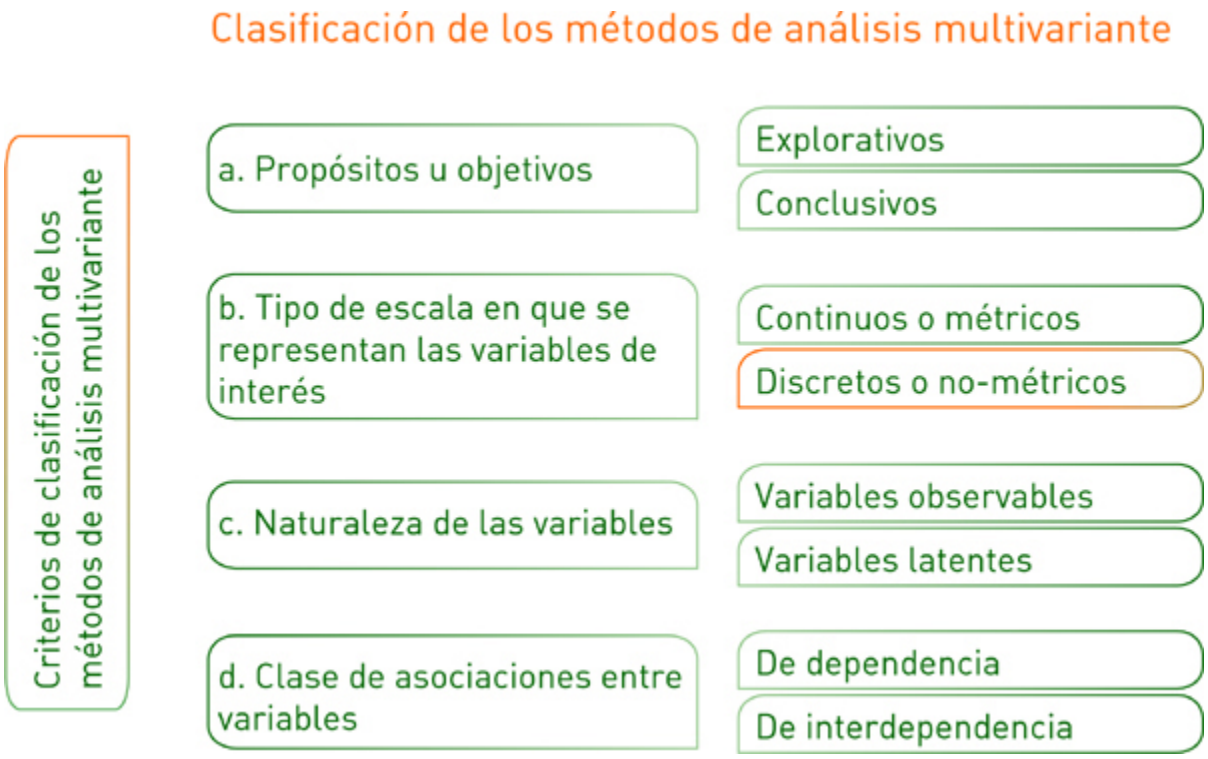
c. Comprensión de la naturaleza de las asociaciones entre variables.

d. Clasificación.

e. Simplificación o reducción de la estructura de los datos.

1.3 Clasificación de los métodos de análisis multivariante

Los métodos de análisis multivariante pueden clasificarse en base a varios criterios, entre ellos: los propósitos u objetivos del análisis, el tipo de escala en que se representan las variables de interés, la naturaleza de estas variables y la clase de asociaciones entre variables que se desean estudiar.



Los proyectos de inteligencia de mercados se pueden clasificar según sus propósitos en dos categorías: exploratorios y conclusivos. Los estudios exploratorios plantean objetivos como: comprender la estructura competitiva del mercado en base a las percepciones del consumidor sobre cuáles productos son sustitutos entre sí; identificar las dimensiones o componentes relevantes de un concepto; generar medidas para conceptos abstractos o generar una propuesta de segmentación de los clientes basada en su patrón de compras históricas. Los métodos multivariantes que permiten atender esta clase de objetivos se califican como exploratorios,

entre ellos se encuentran al análisis de factores, el análisis de conglomerados y el escalamiento multidimensional (MDS por sus siglas en inglés, Multi Dimensional Scaling). Con frecuencia los resultados de un análisis multivariante exploratorio son la entrada para otros análisis como es el caso cuando el análisis de factores se utiliza para resumir las respuestas múltiples en un único puntaje o "variante" (de *variate* en inglés) que se utiliza como variable de respuesta en otro tipo de análisis.

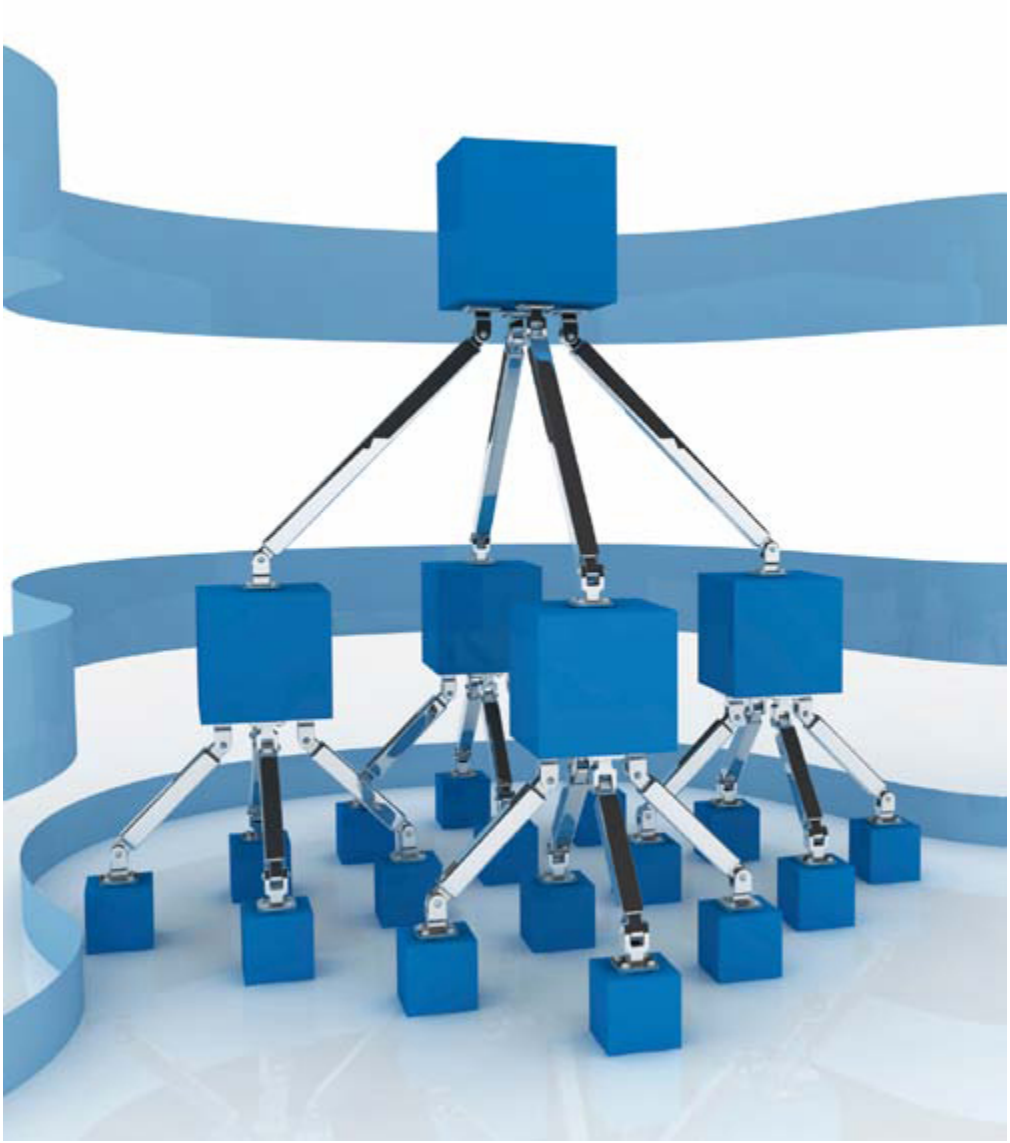
La otra categoría de análisis multivariantes según sus objetivos, son los análisis conclusivos, estas técnicas con frecuencia corresponden a las identificadas como métodos multivariantes clásicos; es decir, son extensiones a las técnicas estadísticas univariadas que permiten la estimación de modelos estadísticos o la prueba de hipótesis. Los objetivos de estos métodos son por ejemplo: seleccionar aquellos atributos de un producto que maximizan la intención de compra del consumidor; evaluar la calidad de un esquema de clasificación crediticia basado en indicadores económicos y demográficos; pro-bar que la publicidad influye en los hábitos alimenticios del consumidor o describir el efecto que la calidad del servicio tiene sobre la satisfacción de los clientes.

Respecto al tipo de escala en la cual se han medido las variables de in terés, los métodos multivariantes se clasifican como: continuos o métricos y discretos o no-métricos. Si el método es métrico, los datos están registrados en escalas de intervalo o de razón; mientras que si el método es no-métrico las variables de interés están en escalas nominales. Esta clasificación puede resultar un tanto ambigua ya que por ejemplo el análisis de regresión lineal simple requiere que todas las variables de interés sean métricas pero el análisis discriminante lineal, en su versión métrica, sólo requiere que las variables usadas para distinguir entre varios grupos de individuos sean métricas. Cuando estas variables discriminantes están medidas en escalas nominales, entonces el análisis discriminante califica como discreto. Los métodos de análisis discriminante se refieren al análisis de datos clasificados en múltiples categorías, por ejemplo consumidores clasificados en términos de la marca de automóvil que prefieren, su género, ocupación actual y la última marca de auto que compró.

En este eBook sólo se discuten los métodos multivariantes métricos, al lector interesado en métodos discretos se le recomienda consultar el libro de Bishop, Fienberg y Holland (2007).

Las variables que se pueden analizar con las técnicas multivariantes pueden diferir en cuanto a su naturaleza. Aquellas variables que son observables o corresponden a características “tangibles” del individuo, lo que per-mite registrarlas sin incurrir en errores de medición considerables, se denominan variables observables. Pero también se puede trabajar con variables que no son observables directamente sino que tienen que ser “inferidas” a partir de indicadores manifiestos. Esta clase de variables, denominadas latentes, se asocian a conceptos con un alto grado de abstracción y complejidad que sólo pueden ser aproximados por el efecto que inducen en variables que sí son medibles u observables. El análisis multivariante comprende tanto métodos que asumen que todas las variables son observables como métodos de análisis latente que permiten explicar las propiedades y relaciones de variables no-observables. Los modelos que se proponen en este tipo de análisis permiten analizar las relaciones entre variables latentes y explicarlas tomando en cuenta los errores asociados al expresarlas o medirlas a través de indicadores tangibles o manifiestos. El análisis de factores, el cual se discute en este eBook, es una técnica multivariante de análisis latente. Este análisis

pretende proponer estructuras lineales que describan las relaciones entre un conjunto de variables observables y latentes.



Un recurso importante que se utiliza en análisis latente son los diagramas de vías (path diagrams en inglés) los cuales permiten expresar gráficamente las asociaciones entre las variables latentes entre sí y con sus indicadores manifiesto. A partir de la estructura declarada, se busca estimar, probar y verificar las asociaciones entre variables.

La Figura 1.2 muestra un ejemplo de un diagrama de vías; las variables latentes son la capacidad de innovación y la posición

competitiva de la empresa. La flecha recta con una única punta que conecta a estas variables indica que la capacidad de innovación (denominada variable exó gena) influye o determina cuál será la posición competitiva de la empresa en su mercado (identificada como variable endógena).

Como estas variables latentes corresponden a concep tos de un alto grado de abstracción, no se pueden observar directamente pero sí deducir a partir de varios indicadores tangibles. Así, la capacidad de innovación se refleja o infiere en términos de cuántos productos nuevos ha desarrollado la empresa en los últimos cinco años (*X*1) y del porcentaje de las utilidades que la organización destina para proyectos de investigación y desarrollo (R&D por sus siglas en inglés, Research & Development). En tanto que la posición competitiva de la empresa se manifiesta en términos del porcentaje de participación que tiene en su mercado (*Y*1) y en su margen de utilidad (*Y*2). Dado que estos indicadores manifiestos son representaciones tangibles de un mismo concepto, se espera estén asociados. Tal asociación se re presenta en el diagrama con una flecha de doble curva de doble punta.

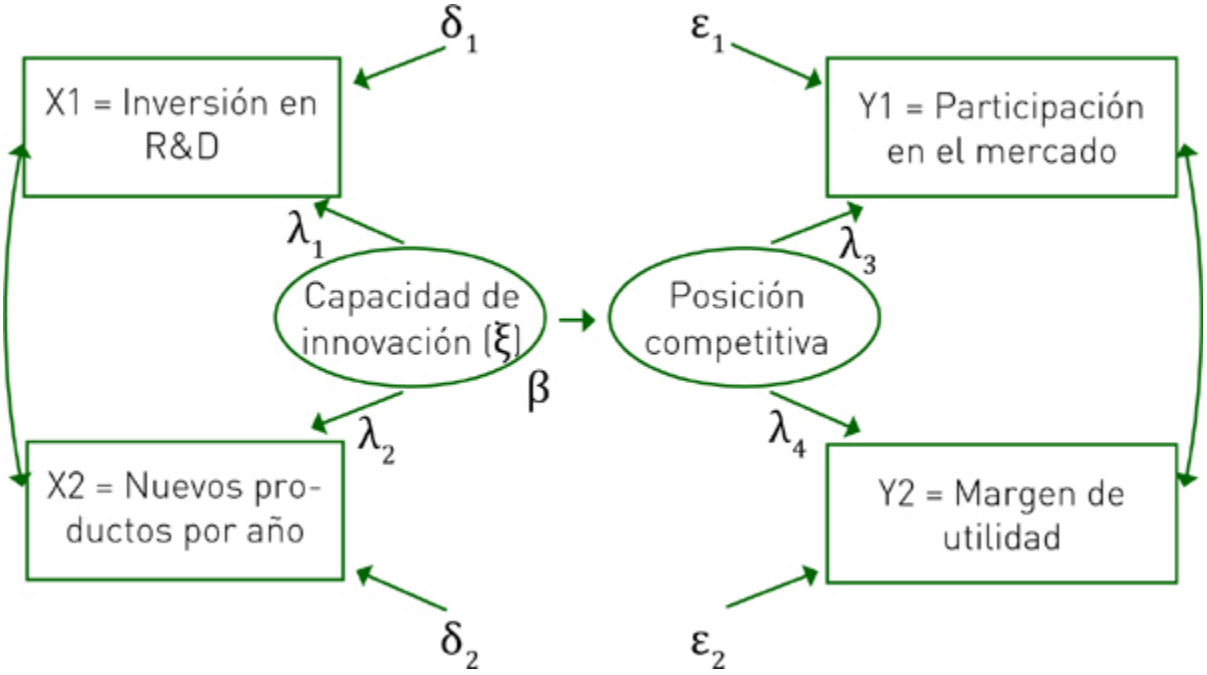


Figura 1.1 Componentes del sistema de inteligencia de mercados

El propósito central del análisis latente es estimar los parámetros del modelo de asociación propuesto, demostrar que describe bien las correlaciones observadas en los datos (buen ajuste) y probar varias hipótesis respecto a los parámetros del modelo. En el diagrama anterior, los parámetros están representados por las letras griegas β y $\lambda 1$-$\lambda 4$, las restantes letras δ's, ε's y ζ representan errores o términos de perturbación. El parámetro β "mide" el efecto que la capacidad de innovación de una firma tiene sobre su posición competitiva y es, por lo tanto, el de mayor interés; mientras las λ's indican hasta dónde los indicadores tangibles (*X*'s y *Y*'s) manifiestan o evidencian la presencia de la **variable latente**.

Dado que las variables latentes no son observables, no hay datos para ellas, por lo que no se pueden derivar expresiones explícitas para estimar los parámetros de interés. Se aplican por lo tanto métodos iterativos de búsqueda para definir los valores de los parámetros, requiriéndose, en general, de imponer restricciones sobre los parámetros para que se puedan obtener estimados únicos. Entre los métodos de estimación disponibles están los de máxima verosimilitud, mínimos cuadrados ponderados y mínimos cuadrados generalizados.

Las relaciones descritas en el diagrama de vías se pueden expresar algebraicamente en el formato de un modelo de **ecuaciones estructurales** lineales. La estructura general de estas ecuaciones, según la notación propuesta por Jöreskog y Sörbom (1979) es la siguiente:

$$\eta = \eta B + \xi \Gamma + \zeta$$

$$Y = \Lambda y \eta + \varepsilon$$

$$X = \Lambda x \xi + \delta$$

Donde:

($\boldsymbol{\eta}$) es un vector con las variables endógenas latentes; es decir, que incluye múltiples variables influenciadas o explicadas en

términos de un conjunto (ξ) de variables latentes exógenas o "predictoras" o incluso por otras latentes endógenas.

X y ***Y*** son vectores que contienen a todos los indicadores usados para "medir" las variables latentes endógenas y exógenas respectivamente.

Para el diagrama en la Figura 1.2, las ecuaciones estructurales correspondientes son:

$$\eta = 0 + \xi\gamma + \zeta$$

$$X_1 = \lambda_1\xi + \delta_1;\ X_2 = \lambda_2\xi + \delta_2$$

$$Y_1 = \lambda_3\eta + \varepsilon_1;\ Y2 = \lambda_4\eta + \varepsilon_2$$

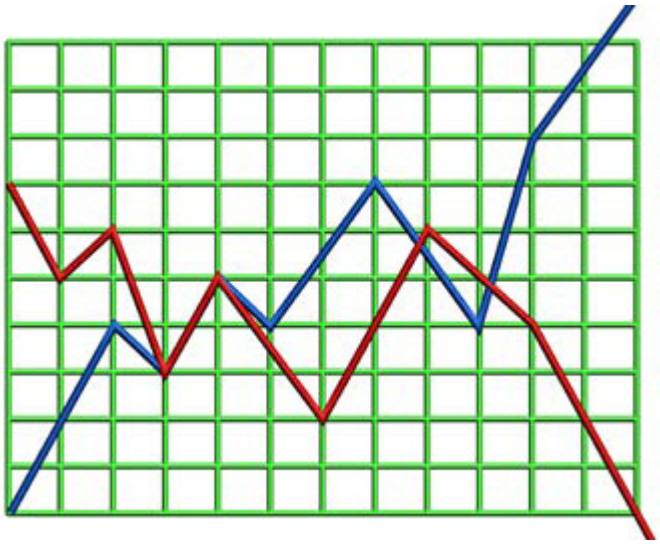
Donde η = posición competitiva de la empresa es la úni ca latente endógena la cual se propone está influenciada por una única latente exógena ξ = capacidad de innovación. Las ecuaciones del segundo renglón asocian a las variables latentes con sus indicadores manifiestos, dos indicadores por cada latente para este ejemplo.

Tanto la formulación como la estimación de las ecuaciones estructurales se pueden realizar con el apoyo de software especializado. El software LISREL (*Linear Structural Relations*), diseñado por el psicometrista sueco Karl Jöreskog (Jöreskog y Sörbom, 1996) fue el software pionero en análisis latente por lo que LISREL y ecuaciones estructurales han sido usados como sinónimos. LISREL estima y prueba modelos de ecuaciones estructurales pero también se puede usar para realizar análisis de factores. Más recientemente han surgido en el mercado otros productos de software diseñados para trabajar con variables latentes, entre ellos está AMOS (*Analysis of Moment Structures*) que ofrece la posibilidad de un análisis a partir de la manipulación de los diagramas de vías (SPSS, 2009).

¿Sabías qué?

Existe un journal académico–Structural Equation Modeling: A Multidisciplinary Journal–dedicado a la publicación de artículos en los cuales se utilizan las ecuaciones estructurales para resolver problemas en varias disciplinas. En este journal también hay un “espacio del maestro” en el cual se proponen esquemas para la enseñanza del modelado con ecuaciones estructurales.

La discusión anterior sobre el tema de análisis latente pone de manifiesto que entre dos variables pueden darse diferentes tipos de asociaciones. Los métodos de estadís tica multivariante también se clasifican con respecto a la naturaleza de las asociaciones que se presuponen entre las variables, bajo este criterio las técnicas estadísticas pueden ser: de dependencia y de interdependencia.



Hay asociación estadística entre dos variables cuando estas tienden a moverse conjuntamente (covariar), esto implica que los cambios en una variable se acompañan por cambios sistemáticos en la otra variable. La intensidad de la asociación entre dos variables se mide en estadística a través de los coeficientes de correlación. El lector recordará que estos coeficientes asumen valores dentro del intervalo [-1; 1]. Cuando el coeficiente es cero las variables no están correlacionadas; valores positivos cercanos a 1 indican una fuerte relación directamente proporcional entre las

variables mientras valores negativos cercanos a -1 son evidencia de una alta relación inversamente proporcional entre las variables. Dos variables pueden covariar por diferentes razones, las más relevantes para la discusión de los métodos multivariantes que se presentan en este eBook son:

1. Ambas variables son indicadores de un mismo concepto. Considere el ejemplo del modelo de ecuaciones estructurales descrito previamente, las variables *X* = nuevos productos desarrollados y *X*2 = inversión destinada a R&D son ambos indicadores manifiestos del concepto “capacidad de innovación”. A medida que la capacidad innovadora de una firma se incrementa, ambos indicadores tenderán a asumir valores altos; esto es, se moverán conjuntamente dependiendo de la magnitud de la capacidad de innovación de la empresa. En el diagrama de vías, la asociación entre las variables está representada gráficamente por la flecha curva con dos puntas.

2. Las variables son indicadores manifiestos para las distintas dimensiones de un mismo concepto. Cuando un concepto es multidimensional como es el caso del concepto “calidad del servicio”, sus dimensiones o componentes son sólo relativamente independientes, por lo cual es razonable esperar cierta variación concomitante entre indicadores asociados a distintas dimensiones. Si una empresa de servicios hace esfuerzos por incrementar la calidad de su servicio, es de esperar que tales esfuerzos resulten no sólo en empleados más dispuestos a atender las necesidades de los clientes (dimensión Respuesta) sino también más enfocados a comprender cuáles son estas necesidades (dimensión Compromiso). Es decir que, al mejorarse globalmente la calidad del servicio, se incrementan simultáneamente los indicadores de todas las dimensiones dando como resultado asociaciones significantes entre parejas de indicadores definidos para dimensiones diferentes.

3. Ambas variables varían conjuntamente debido al efecto (directo o mediador) de una tercera variable. Esta situación incluye el caso de lo que se denominan relaciones falsas o espurias (en inglés *spurious correlations*). Las asociaciones aparentes entre variables

se prestan a malas interpretaciones ya que alguien podría suponer que un coeficiente correlación relativamente alto implica que una variable es influenciada o es consecuencia de otra variable.

Un ejemplo es el caso reportado de asociación entre la venta de helados y el número de personas ahogadas (*Examples of spurious correlations,* s.f.), suponer que si se restringe la venta de helado se disminuiría el número de personas ahogadas es una falacia ya que no hay un mecanismo que implique que el cambio en una variable conlleva un cambio en la otra.

Una explicación lógica para esta relación espuria es que durante los meses más cálidos, las personas buscan refrescarse consumiendo helando y/o nadando en albercas; como consecuencia del incremento en la temperatura ambiente ambas variables (venta de helados, número de ahogados) se incrementan.

Otro ejemplo de relación espuria es la reportada para las variables: largo del cabello de un estudiante de preparatoria y sus calificaciones (Burns, 1997). Nuevamente, presuponer que para mejorar las calificaciones hay que dejarse crecer el cabello es una falsedad.

Una explicación razonable para la ocurrencia de esta correlación falsa es que según la experiencia de los profesores de preparatoria, las mujeres tienden a ser más responsables y dedicadas con sus actividades escolares que los hombres, este compromiso y dedicación tienen como consecuencia mejores calificaciones. Como las mujeres suelen tener el cabello más largo que los hombres esto

explicaría el porqué de la relación espuria reportada. El lector debe asegurarse que las asociaciones que está explorando son asociaciones sustentadas por teorías que ofrecen una explicación robusta para las relaciones.

4. Relaciones causa-efecto. Este tipo de relaciones son asimétricas, es decir una variable conduce a la otra pero no a la inversa. En el diagrama de vías de la Figura 1.2, este tipo de relaciones se representó con flechas directas de una sola punta para remarcar que la relación es asimétrica, esto es va de "*X*" a "*Y*" pero no a la inversa. La mera variación conjunta entre dos variables no es condición suficiente para inferir causalidad. Se puede proponer que *X* es una de las "causas" de *Y* sólo si aparte de variar conjuntamente, si siempre que ocurre *X* acontece *Y*, y nunca se observa a Y si previamente no ha ocurrido *X*. Por ejemplo, se puede plantear que *X* = tamaño de la fuerza de ventas de una empresa es una de las causas de *Y* = volumen de ventas, si el coeficiente de correlación es significante (en términos de estadística) y si las ventas se incrementan cada vez que más agentes de ventas son contratados para promover y asesorar directamente a las industrias que son clientes de la empresa.

Un método de estadística multivariante es de interdependencia si se enfoca a estudiar relaciones de asociación simétricas; esto es, relaciones en las cuales ninguna de las dos variables se presupone como la causa de la otra. Bajo esta idea, los métodos de análisis de factores y de análisis de conglomerados son métodos de interdependencia. En contraste, un método de dependencia busca probar empíricamente la existencia de relaciones causa-efecto entre variables. El análisis de regresión, el análisis de varian za y el análisis conjunto son métodos de dependencia cuyo fin es modelar, relaciones causales, determinar cuáles variables tienen una influencia significativa sobre las respuestas de interés y pronosticar el comportamiento de estas respuestas.

Actividad de repaso (1era parte)

1.3 Clasificación de los métodos de análisis multivariante

1. Lee cuidadosamente cada enunciado e indica si la declaración es Falsa (F) o Verdadera (V).

___ El análisis factorial es un método de interdependencia.

___ La relación entre la calidad que tiene un producto y el interés de compra expresado por los consumidores se considera una relación de interdependencia.

___ Un gerente de mercadotecnia debe esperar una relación causal entre el porcentaje de cumplimiento en sus metas laborales y el promedio que logró en su licenciatura.

___ El análisis discriminante múltiple es un método de análisis latente.

___ El concepto lealtad está conformado por lealtad cognitiva, afectiva, de intención y conductual por lo cual es un concepto multidimensional.

___ Si se usan tres o más reactivos para evaluar un concepto éste se denomina multidimensional.

___ Un concepto abstracto como la lealtad se considera latente.

___ El margen de operación de una empresa es una variable observable.

___ La capacidad de mejora continua de una empresa es una variable observable.

___ Una relación espuria ocurre cuando no se cuenta con una explicación teórica razonable para el alto valor de su coeficiente de correlación.

Actividad de repaso (2da parte)

1.3 Clasificación de los métodos de análisis multivariante

Para cada uno de los proyectos de la actividad de repaso descritos en la sección 1.2 identifica si el proyecto requiere de aplicar métodos multivariables de interdependencia o de dependencia.

a. Dependencia

b. Dependencia

c. Interdependencia

d. Dependencia

e. Interdependencia

1.4 Descripción de los datos multivariantes

Los objetivos de un proyecto de inteligencia de mercados definen las necesidades de información. Una vez que la información requerida se ha recolectado ya sea a través de métodos de encuesta, observación o experimentación, el siguiente paso es presentarla debidamente organizada y resumida. Para ello se utiliza la notación matemática empleada en álgebra lineal puesto que se trabaja con múltiples variables.

Las variables que constituyen los datos de interés se representan como un vector columna, *x*, con *p* entradas. Para cada individuo, objeto, empresa o situación se miden todas las variables, por lo cual los datos consisten de una muestra de n vectores $x_1,x_2,\ldots x_n$, con entradas individuales x_{ij} = medición para el *i*-ésimo individuo en la j-ésima variable. Los *n* datos *u* observaciones recolectadas se pueden representar como en el arreglo matricial *nxp* siguiente; en este arreglo cada renglón *i*, contiene la información de un individuo u objeto en tanto cada columna *j* incluye la información disponible para la *j*-ésima variable medida.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \ddots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Las *p* variables de respuesta se consideran un conjunto de variables aleatorias cuyo comportamiento se describe en términos de su distribución de probabilidad conjunta $f(x_1,x_2,\ldots x_p)$, esta consideración sobre el modelo probabilístico que enlaza la información observada con el mecanismo que la genera, permite describir *y* modelar estadísticamente los datos multivariantes.

En el caso de los métodos estadísticos univariados, la función de probabilidad de la única respuesta de interés se asume normal para facilitar el trabajo de la estadística matemática para la construcción

de inferencias estadísticas. En el caso multivariante, la generalización de la función de densidad normal es también esencial en el desarrollo de los varios métodos de análisis. La distribución normal multivariante es, por lo tanto, la distribución de probabilidad conjunta que se asume para varias técnicas de estadística multivariante; la Figura 1.3 describe gráficamente esta función de densidad para el caso de dos variables (x_1,x_2).

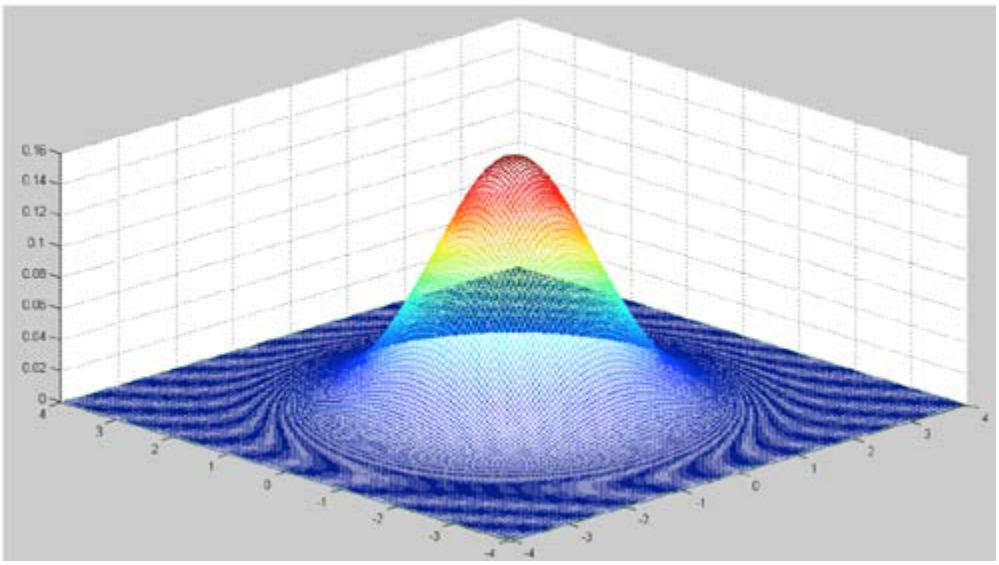


Figura 1.3 La distribución de probabilidad normal bivariada.

Si bien la distribución de probabilidad normal resulta atractiva porque simplifica el desarrollo matemático de los métodos estadísticos, su importancia radica en que para muchos de los datos disponibles esta distribución ofrece una aproximación razonable. Por otra parte, aún cuando la población a partir de la cual se generaron los datos no fuera cercana a la normal, de contarse con una gran cantidad de observaciones–como suele ser el caso cuando se aplican técnicas de estadística multivariante–las distribuciones de muestreo de los estadísticos multivariantes son asintóticamente normales, lo que valida la utilización de los métodos derivados bajo el modelo de probabilidad normal.

Puesto que los datos generados en inteligencia de mercados se representan con frecuencia en escalas ordinales o de actitud (por ejemplo escala Likert) para las cuales la distribución normal no

resulta una aproximación adecuada, el usuario deberá decidir si es válido analizarlos con métodos que asumen normalidad o bien optar por investigar otros métodos que no descansan en este supuesto.

Conviene resumir los datos multivariantes para fines de presentación y exploración. Los estadísticos descriptivos que se calculan dan información sobre la tendencia central y la variabilidad de los datos, así como sobre la intensidad de las asociaciones lineales entre variables. El promedio de las observaciones disponibles para la *j*-ésima variable de interés está dado por:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad (1)$$

Todos los promedios de las *p* variables son las entradas del vector llamado centroide que es la media muestral de los datos, el cual se representa como sigue:

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \qquad (2)$$

Para medir el grado de variabilidad de cada variable se utiliza la **varianza** muestral, cuyo estimador insesgado está dado por:

$$s_j^2 = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}{n-1} \qquad (3)$$

Para el conjunto de datos recolectados, la asociación lineal entre cualesquiera dos variables, x_j y x_k, se determina mediante el producto de las desviaciones de cada variable respecto a su media muestral, este estadístico es una medida absoluta del grado de asociación entre las variables y se denomina covarianza muestral.

Para calcular este estadístico a partir de *n* datos se usa la siguiente expresión:

$$s_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{n-1} \qquad (4)$$

La asociación entre dos variables implica un tipo de relación simétrica de donde $s_{jk} = s_{kj}$. Notar que si en la expresión de cálculo de la covarianza muestral se hace $j=k$, la expresión (4) corresponde a la del cálculo de la varianza (3), por lo cual en la notación de análisis multivariante sjj se utiliza para denotar a la varianza muestral de la j-ésima variable. Las varianzas y covarianzas para las *p* variables medidas se presentan en un arreglo matricial ***S*** denominado **matriz de varianza-covarianza** muestral, la cual contiene la información sobre el grado de dispersión de los datos multivariables y también aquélla referente a la magnitud y dirección de las asociaciones entre variables. La matriz ***S*** tiene la siguiente estructura:

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \ddots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \qquad (5)$$

La matriz de varianza-covarianza muestral es simétrica, es decir que su *i*-ésimo renglón es idéntico a su *j*-ésima columna. Recordando los conceptos de álgebra de matrices, S' o S^T representa a la matriz transpuesta de S, la cual se obtiene cuando el elemento s_{ij} de la matriz original *S* se convierte en el elemento *s*. Una matriz es simétrica cuan *ji* do coincide con su transpuesta, esto es cuando S'=S como se puede verificar fácilmente para el caso de la matriz de varianza-covarianza.

Para facilitar más la representación y exploración de los datos multivariados, la información de la matriz de varianza-covarianza muestral se puede resumir en un único número, el determinante de la matriz esto es |S|. Este estadístico, conocido como varianza generalizada se calcula de la siguiente forma:

$$S = |S| = s_{11} S_{11} + s_{12} S_{12} + \ldots s_{1p} S_{1p} \qquad (6)$$

Donde $S_{ij} = (-1)^{i+j}$ det M_{ij} es el cofactor i,j de S y M_{ij} es el menor, el cual corresponde a la matriz $(n-1)$x$(n-1)$ obtenida al eliminar el renglón i y la columna j de S. La varianza generalizada no proporciona información sobre la dirección de las asociaciones entre variables, por lo cual resulta útil para comparar la variabilidad y asociación de grupos de datos que covarían en la misma dirección. Si sólo se tienen dos variables, x_1 y x_2, la asociación y dispersión de los datos se puede representar gráficamente en un diagrama de dispersión como el de la Figura 1.4.

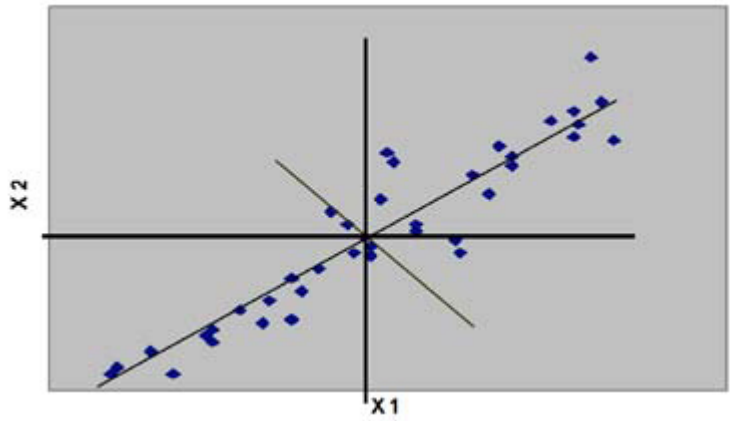


Figura 1.4 Caso bivariado: diagrama de dispersión y dirección del patrón de asociación.

La varianza generalizada |S| es proporcional a la magnitud de la dispersión de los puntos bivariados alrededor de su centroide $x = (x_1, x_2)$ que en el diagrama anterior se ha trasladado al origen para simplificar, esto es $x = (x_1, x_2) = (0, 0)$. El cuadrado de la distancia de un punto cualquiera $x = (x_1, x_2)$ al origen está dado por (7) donde S^{-1} es la inversa de la matriz de varianza-covarianza, esto es $SS^{-1} = S^{-1}$

S = I don-de I es la matriz idéntica, esto es una matriz con entradas de “1” en la diagonal y “0” por arriba y debajo de la diagonal. Los puntos (x_1,x_2) que distan a lo más una distancia arbitraria c del origen como se representa en (7):

$$x' S^{-1} s \leq c^2 \qquad (7)$$

definen un elipse cuyo eje mayor está en la dirección de máxima variabilidad de los datos. Dependiendo del valor que tome la constante *c*, un determinado porcentaje de datos quedarán contenidos en el elipse; si c^2 = 5.99, aproximadamente el 95% de las parejas de datos disponibles quedarán contenidos dentro del elipse (Johnson y Wichern, 1988, p. 134). El área de este elipse es proporcional a la magnitud de la varianza generalizada, es decir (área del elipse)^2 = (constante) (varianza generalizada muestral). Entre más dispersos estén los datos, mayor será el área del elipse que se trace para contenerlos lo cual implica un mayor valor de la varianza generalizada. Esta representación geométrica se puede extender a *p* dimensiones, es decir al caso de *p* variables, para *p*>2 la expresión (7) define hiper-elipsoides cuyo volumen es proporcional a la varianza generalizada de los datos.

Otro estadístico descriptivo muy utilizado en análisis multivariable es el **coeficiente de correlación**, que cuando es calculado a partir de datos muestrales se denota por. Este coeficiente mide la asociación lineal entre una pa $^r jk$ reja de variables pero es independiente de las unidades de medida, por lo que r_{j_k} es una medida estandarizada o adimensional que sólo toma valores en el intervalo de [-1;1] según se puede demostrar. Entre mayor el valor absoluto del coeficiente de correlación, mayor el grado de asociación entre las variables; el signo del coeficiente indica si la relación es directa o inversamente proporcional. El coeficiente de correlación muestral se calcula como sigue:

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}} = \frac{\sum_{i=1}^{n} x_{ij}x_{ik} - \frac{\left(\sum_{i=1}^{n} x_{ij}\right)\left(\sum_{i=1}^{n} x_{ik}\right)}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_{ij}^2 - \frac{\left(\sum_{i=1}^{n} x_{ij}\right)^2}{n}\right)\left(\sum_{i=1}^{n} x_{ik}^2 - \frac{\left(\sum_{i=1}^{n} x_{ik}\right)^2}{n}\right)}} \quad (8)$$

El coeficiente anterior también se puede visualizar como la **covarianza** muestral entre dos variables estandarizadas, y z_{ik} obtenidas a partir de las x's originales mediante la ^{z}ij transformación dada en (9). Las variables así transformadas tienen media cero y varianza de uno:

$$z_{ji} = \frac{(x_{ij} - \bar{x}_j)}{\sqrt{s_{jj}}} \quad para\ j = 1,2,\ldots,p \quad (9)$$

El conjunto de coeficientes de correlación se presenta en un arreglo matricial conocido como matriz de correlación, **R**. Esta matriz es también simétrica y todas las entradas en su diagonal son 1 ya que corresponden a la varianza muestral de las p variables estandarizadas z_{ij}, j = 1,2,…, p. La estructura de la matriz de correlación R se da en (10).

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \ddots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix} \quad (10)$$

Los estadísticos anteriores permiten no únicamente describir y resumir los datos multivariables, sino también estimar el valor de los parámetros que caracterizan a la población de donde se obtuvo la muestra. En la sección 1.4 se indicó que con frecuencia la distribución de probabilidad que se asume como generadora de los datos multivariables es la distribución normal, la forma de esta función es como sigue:

$$f(X) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} e^{-(x-\mu)'\Sigma^{-1}(x-\mu)/2}, \quad -\infty < x_i < \infty, i = 1,2,\ldots, p \qquad (11)$$

En la función anterior, μ es el parámetro de localización de la distribución y representa el valor esperado o media del vector x mientras que la matriz pxp, Σ es la matriz de varianza covarianza la cual contiene la información sobre la dispersión y patrones de asociación de x. Cuando las variables aleatorias que integran el vector x se estandarizan aplicando la transformación (9), la distribución de probabilidad normal multivariada anterior en lugar de Σ involucra a ρ la matriz de correlación poblacional, cuyas entradas son los coeficientes de correlación ρ_{ij} que expresan la asociación lineal de cada pareja de variables aleatorias.

Mientras que μ, Σ y ρ tienen entradas que son fijas y determinan exactamente el comportamiento de los datos, sus estimadores media muestral (centroide), matriz de varianza-covarianza y matriz de correlación muestrales son variables aleatorias puesto que sus entradas dependen de la muestra a partir de las cuales fueron calculadas. La media muestral y la matriz de varianza-covarianza son además independientes entre sí y tienen las siguientes distribuciones de muestreo:

$$\bar{X} \sim N_p(\mu, (1/n)\Sigma)$$
$$(n-1)S \sim \textit{Wishart con } n-1 \textit{ grados de libertad}$$
$$\textit{además}, n(\bar{X}-\mu)'S^{-1}(\bar{X}-\mu) \sim \chi^2_p$$

Las distribuciones normal y *ji*-cuadrada son conocidas para el lector de los cursos de estadística básica, en cuanto a la distribución

Wishart; ésta es otra distribución de probabilidad de importancia en el contexto multivariado y nombrada así en reconocimiento a su descubridor. Establecer el comportamiento de muestreo de los estadísticos es necesario para el desarrollo matemático de las técnicas multivariables que este eBook describe y aplica en el contexto de inteligencia de mercados. El cálculo e interpretación de los estadísticos multivariados se ilustra en el siguiente ejemplo, se usan pocos datos porque el objetivo es mostrar cómo se calculan los estadísticos descriptivos multivariados. Posteriormente se utilizará el *software* MINITAB para trabajar con grupos de datos de tamaño más razonable. Cabe aclarar que la complejidad de los cálculos de las técnicas de estadística multivariable requiere de *software* de apoyo.

EJEMPLO: En el curso de inteligencia de mercados, la evaluación de los estudiantes se hace en términos de varios indicadores.

Los tres más importantes son: el promedio de exámenes parciales los cuales consisten en resolver problemas prácticos proporcionados por el profesor, la calidad de un proyecto de inteligencia de mercados que desarrollan a lo largo del semestre bajo la supervisión del profesor y el examen final, el cual se diseña para evaluar conceptos y la comprensión de los resultados aportados por las varias técnicas multivariadas que se revisan en el curso. Los datos en la Tabla 1.3 corresponden a una muestra de estudiantes que tomaron el curso el año pasado.

Estudiante	X1 = promedio exámenes parciales prácticos	X2 = proyecto de inteligencia de mercados	X3 = clasificación del examen final
1	92	95	70
2	85	87	79
3	90	88	98
4	95	95	88
5	84	82	70
6	78	81	80
7	88	92	76

Tabla 1.3 Datos multivariantes que describen el desempeño de un estudiante.

Para los datos anteriores, calcular e interpretar el centroide, matriz de varianza-covarianza muestral, matriz de correlación muestral y varianza generalizada para los datos anteriores.

Respuesta:

a. *Media muestral.*

El cálculo del centroide requiere de obtener el promedio de los datos para cada una de las tres variables medidas, usando la expresión (1).

La media muestral de los datos es el siguiente vectors $\begin{bmatrix} 87.4286 \\ 88.5714 \\ 80.1429 \end{bmatrix}$

Las entradas del vector anterior representan hacia qué valor tienden a concentrarse los datos, comparando los tres promedios tenemos que las calificaciones más bajas son para el examen final y las mejores para el proyecto de inteligencia de mercados que desarrollaron los alumnos.

b. *Matriz de varianza-covarianza muestral.*

Para calcular la matriz de varianza-covarianza muestral, se tienen que calcular todas las varianzas y covarianzas usando las expresiones (3) y (4). Para la primera variable X1 = promedio de exámenes prácticos, el cálculo es como sigue:

$$s_1^2 = \frac{\sum_{i=1}^{7}(x_{i1} - \bar{x}_1)^2}{7-1} = \frac{(92-87.4286)^2 + (85-87.4286)^2 + \ldots + (88-87.4286)^2}{6} = 31.9524$$

Para las otras dos variables, el cálculo es similar y da como resultado los siguientes estimados:

$$s_2^2 = 32.9524 \; y \; s_3^2 = 100.8095.$$

El cálculo de las covarianzas muestrales amerita utilizar parejas de columnas de datos. En el caso de la covarianza entre las variables X1 y X2 se tienen que usar las primeras dos columnas de la Tabla 1.1 para hacer los siguientes cálculos:

$$s_{12} = \frac{\sum_{i=1}^{7}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{7-1} = 29.3810$$
$$= \frac{(92-87.4286)(95-88.5714) + (85-87.4286)(87-88.5714 + \ldots + (88-87.4286)(92-88.5714)}{6}$$

Una vez que todas las varianzas y covarianzas muestrales han sido calculadas, se utilizan como entradas de la matriz de varianza-covarianza muestral que se da enseguida:

$$S = \begin{bmatrix} 31.9524 & 29.3810 & 15.9286 \\ 29.3810 & 32.9524 & 5.0714 \\ 15.9286 & 5.0714 & 100.8095 \end{bmatrix}$$

Todas las covarianzas muestrales en la matriz S son positivas lo que indica que todas las variables covarían en la misma dirección. La covarianza mide la intensidad de la asociación lineal entre parejas de variables; de la matriz anterior se aprecia que s_{23} = 5.0714 es la menor de las covarianzas lo que revela que las

variables menos asociadas son X2 = calificación del proyecto de inteligencia de mercados y X3 = calificación del examen final.

El determinante de la matriz ***S*** es la varianza generalizada de los datos, para calcular el determinante se procede como sigue:

$$\det S = 5{,}798{,}245.172$$

$$(-1)^{1+1}(31.9524)\begin{vmatrix} 32.9524 & 5.0714 \\ 5.0714 & 100.8095 \end{vmatrix} + (-1)^{1+2}(29.2810)\begin{vmatrix} 29.3810 & 5.0714 \\ 15.9286 & 100.8095 \end{vmatrix} + (-1)^{1+3}(15.9286)\begin{vmatrix} 29.3810 & 32.9524 \\ 15.9286 & 5.0714 \end{vmatrix}$$

c. *Matriz de correlación*

Al igual que con la matriz de varianza-covarianza, el cálculo de la matriz de correlación requiere de calcular todos los coeficientes de correlación para las tres parejas de variables. Para calcular el coeficiente de una pareja en particular se usa la expresión (8), en el caso de las variables X1 y X2 el cálculo queda como sigue:

$$r_{12} = \frac{s_{12}}{\sqrt{s_{11}s_{22}}} = \frac{29.3810}{\sqrt{(31.9524)(32.9524)}} = 0.905$$

Los otros dos coeficientes de correlación se calculan de manera similar y en conjunto se integran a la matriz de correlación muestral R que se reporta enseguida.

$$R = \begin{bmatrix} 1 & 0.905 & 0.281 \\ 0.905 & 1 & 0.088 \\ 0.281 & 0.088 & 1 \end{bmatrix}$$

Note que el tamaño de los coeficientes de correlación es proporcional a la magnitud de las covarianzas de S, pero todos estos estadísticos son menores a 1 puesto que son medidas estandarizadas de asociación entre las variables.

Las medidas anteriores, vector de promedios (centroide) y matrices de varianza-covarianza y de correlación son los estadísticos descriptivos que resumen la información para un conjunto de datos multivariados. Estos estadísticos se utilizan no

únicamente para representar información, sino también como base para distintos análisis multivariantes según se apreciará en los capítulos subsecuentes.

1.5 Exploración de los datos multivariantes

Antes de empezar a procesar los datos multivariantes es conveniente realizar una exploración de los datos para determinar si estos satisfacen los supuestos básicos requeridos por los métodos estadísticos–especialmente los conclusivos–y también para identificar observaciones inusuales que puedan afectar la estimación y calidad de los modelos estadísticos.

Los datos atípicos o también llamados datos extremos son observaciones que se desvían del patrón de los datos en general; algunas observaciones pueden ser incluso marcadamente diferentes; esto es, inusualmente menores o mayores al resto de la muestra a la que pertenecen. En el caso univariado, estas observaciones extremas se pueden detectar utilizando gráficos exploratorios como los diagramas de puntos y los de bloques y líneas. Estos gráficos son recursos de amplio uso en estadística que proporcionan al analista una herramienta visual para detectar fácilmente los datos atípicos.

En el caso del diagrama de puntos, los datos extremos se muestran aislados del resto de los datos como se ilustra en la Figura 1.5 en la cual se ha representado el desempeño por territorio de ventas de una empresa fabricante de muebles. Las ventas están medidas como índices relativos, el 100 corresponde al promedio de las ventas para todos los territorios. La observación en la extrema derecha es un dato atípico que corresponde a un índice de ventas de aproximadamente 141 mientras que en el resto de los territorios de ventas los índices son inferiores a 122.

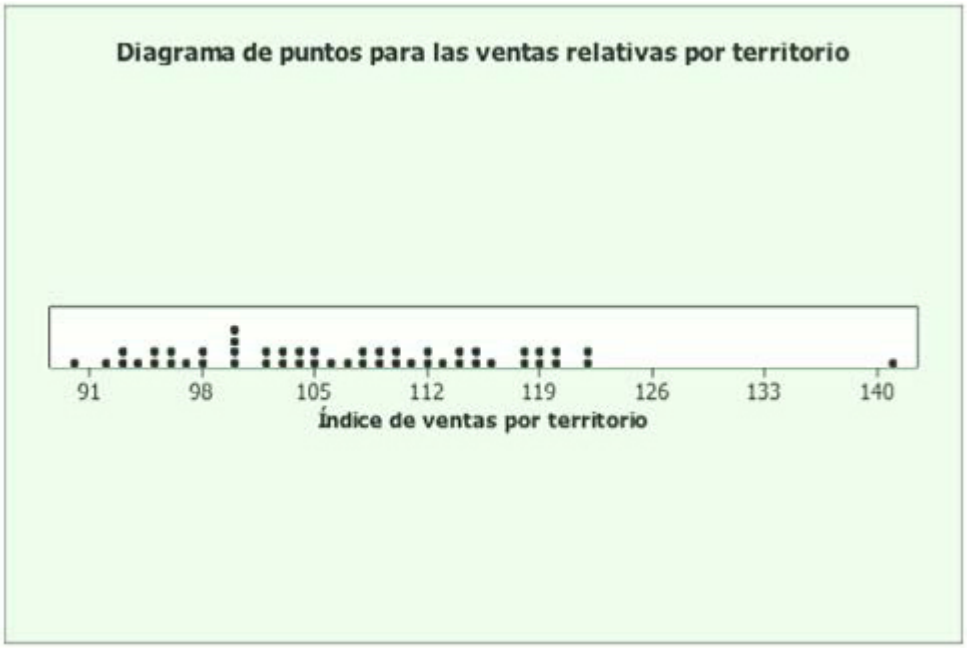


Figura 1.5 Detección de datos extremos en un diagrama de puntos.

En el diagrama de bloques y líneas de la Figura 1.6, este dato atípico está representado por un asterisco que dista considerablemente de la línea superior del diagrama.

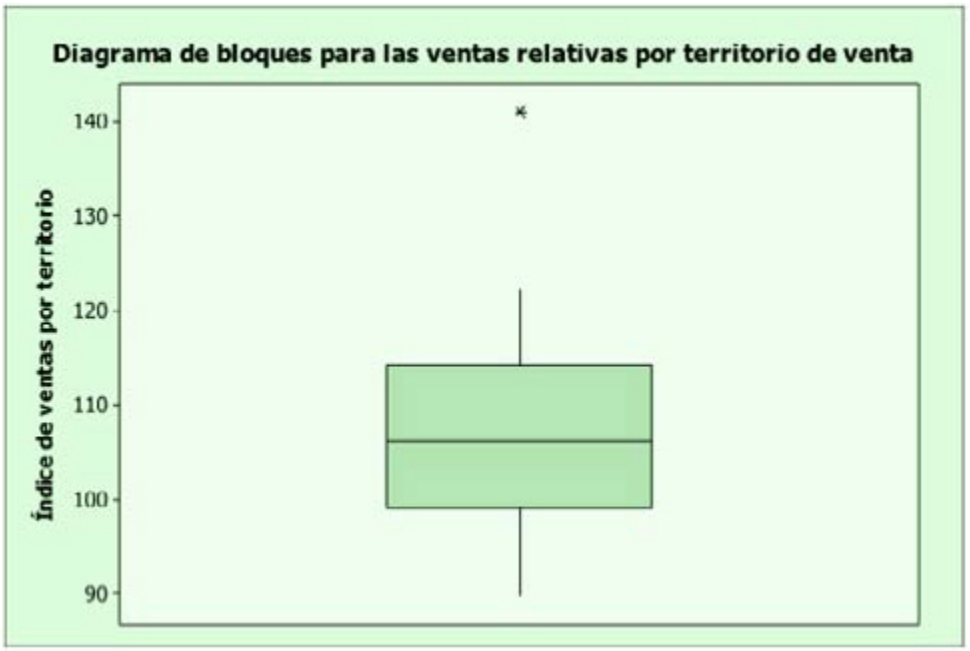


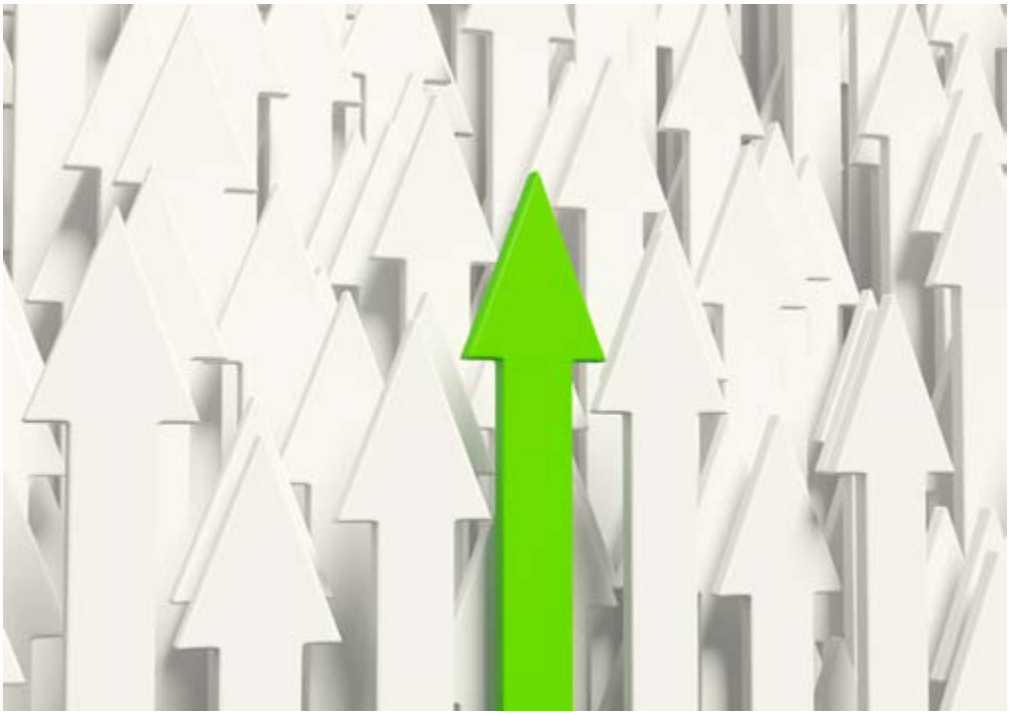
Figura 1.6 Diagrama de bloques para identificación de datos atípicos.

En aquellos casos en que se hayan ajustado modelos estadísticos para explicar una respuesta en términos de un conjunto de variables independientes como es el caso del análisis de regresión lineal múltiple–método que el lector ya conoce de su curso básico en estadística–lo más apropiado para identificar datos extremos es calcular los residuos del modelo y presentarlos gráficamente. Estos residuos se calculan como la diferencia entre los valores observados y aquellos que especifica o predice el modelo estadístico ajustado a los datos, esto es $e_i = Y_i - \hat{Y}_i$ donde el (^) indica que este valor fue generado o pronosticado en base al modelo estadístico.

Si los residuos son estandarizados aplicando la transformación descrita en (9), aquéllos que están fuera del intervalo de [-2; 2] se consideran atípicos puesto que bajo el supuesto de normalidad, el cual es común para todos los métodos multivariados de dependencia que describe este eBook, sólo el 5% de los residuos estarán fuera de este intervalo. Aquellos residuos que quedan fuera del intervalo de [-3; 3] son aún más inusuales, ya que el 99.75% de los residuos deberían estar dentro de los límites del intervalo cuando se asume normalidad. Es importante notar que un dato atípico o extremo no es necesariamente un dato incorrecto, si se tienen 100 datos estandarizados, un 5% se espera que queden fuera del intervalo de [-2; 2]. Por lo tanto, eliminar un dato atípico no siempre es justificable, el analista deberá determinar cómo fue obtenido el dato y el efecto que tiene sobre el análisis que va a realizar antes de decidir si conviene eliminarlo. Una opción para no eliminar los datos extremos es utilizar métodos más robustos que logren mejorar el ajuste de un modelo a los datos aún en presencia de observaciones extremas.

Una vez que se han calculado los residuos estandarizados, un gráfico de estos residuos versus el orden en que fueron generados los datos permite detectar fácilmente la presencia de datos atípicos. Si se utiliza software estadístico para estimar el modelo estadístico de interés, la construcción del gráfico de residuos versus el orden de los datos está disponible en las opciones. La gráfica de los residuos es la opción más deseable para visualizar los datos atípicos; sin embargo, estos datos también pueden ser observados en un

diagrama de dispersión. El problema con este diagrama es que sólo se pueden representar dos variables a la vez mientras que los residuos reflejan el ajuste de un modelo en el cual se han usado múltiples variables explicativas o predictoras.



Ejemplo: Arroyo-López y Borja-Medina (2011) describen un caso de aplicación de la regresión lineal múltiple en el cual la variación en el monto del gasto destinado a obras públicas y acciones sociales por parte de los gobiernos estatales se explicó en términos de variables como: inversión extranjera en el estado, ingresos del gobierno por concepto de impuestos, tasa de desocupación y producto interno bruto estatal. La gráfica de los residuos de este modelo versus el orden en que la información fue obtenida de los registros del Instituto Nacional de Estadística Geografía e Informática (INEGI) se reporta en la Figura 1.7.

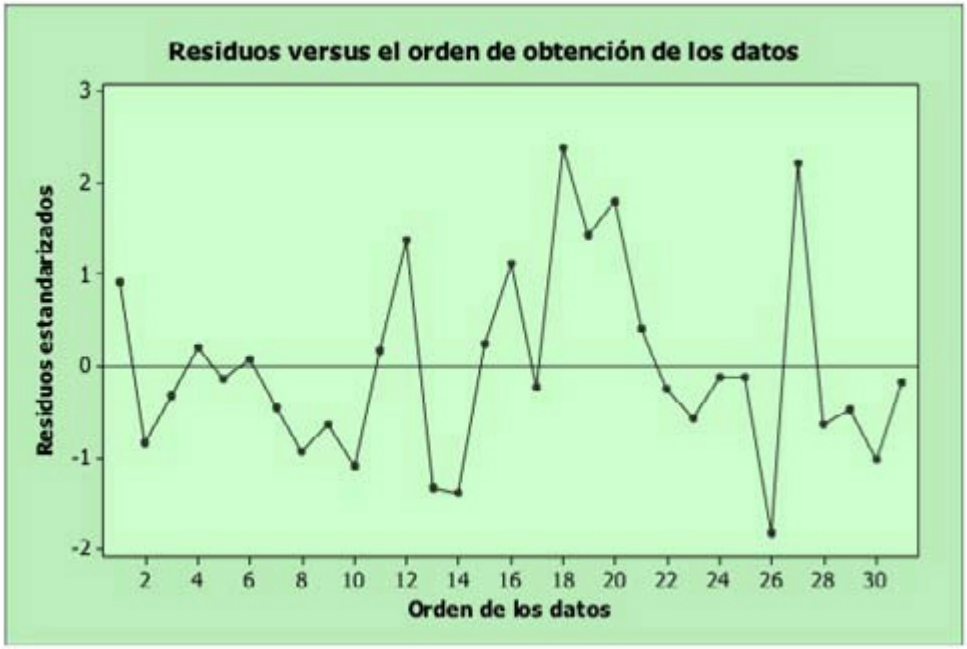


Figura 1.7 Residuos estandarizados para el modelo de regresión versus orden en que se obtuvieron los datos.

Puesto que en análisis multivariable se trabaja con múltiples respuestas, los gráficos de puntos o de bloques tendrían que construirse para cada variable, es decir serían gráficos marginales en los cuales los patrones de asociación entre variables no se consideran. Para corregir esta omisión, la detección de datos extremos multivariados se puede hacer a través del "cálculo" de la llamada distancia de Mahalanobis, la cual se calcula como sigue:

$$D^2 = (x_i - \bar{x})' S^{-1} (x_i - \bar{x}) \qquad (12)$$

Donde xi es un vector de datos cualquiera, $\bar{x}$ es el vector de promedios y S la matriz de varianza-covarianza muestral. Según la discusión en la sección anterior sobre la varianza generalizada, cuando los datos siguen la distribución normal el elipsoide definido por la distancia D^2 de Mahalanobis contiene o cubre aproximadamente el c^2 x100 de los datos. Por lo tanto aquellos datos cuya distancia sea mayor a c^2 califican como atípicos. Johnson y Wichern (1998, p. 204) establecen que para el caso bivariado aproximadamente un 95% de los datos quedarían dentro

del elipse dado por $D^2 = (x_i - \bar{x})' S^{-1} (x_i - \bar{x}) \leq \chi_2^2(.95)$ donde $\chi_2^2(.95)$ es el percentil 95 de la distribución ji-cuadrada con 2 grados de libertad. Aquéllos cuyas distancias de Mahalanobis D^2 sean superiores a este porcentil se consideran atípicos.

Uno de los principales supuestos para los métodos de análisis multivariable es la normalidad de los datos involucrados. Cuando este supuesto no se satisface, las inferencias estadísticas no serían válidas: es importante, por lo tanto, validar este supuesto antes de procesar los datos. Una técnica muy sencilla para examinar la normalidad de los datos son los histogramas. Estas gráficas demandan que el número de datos sea grande, para que el ancho de las clases no distorsione la representación visual (Hair et al., 1999) de los datos; el histograma de los datos se compara contra el de la distribución normal para determinar si hay discrepancias notables. Los datos de ventas por territorio de la empresa fabricante de muebles se utilizaron para construir el histograma de la Figura 1.8 el cual está inscrito dentro de la curva normal para facilitar la comparación entre la distribución empírica y el modelo normal que se asume.

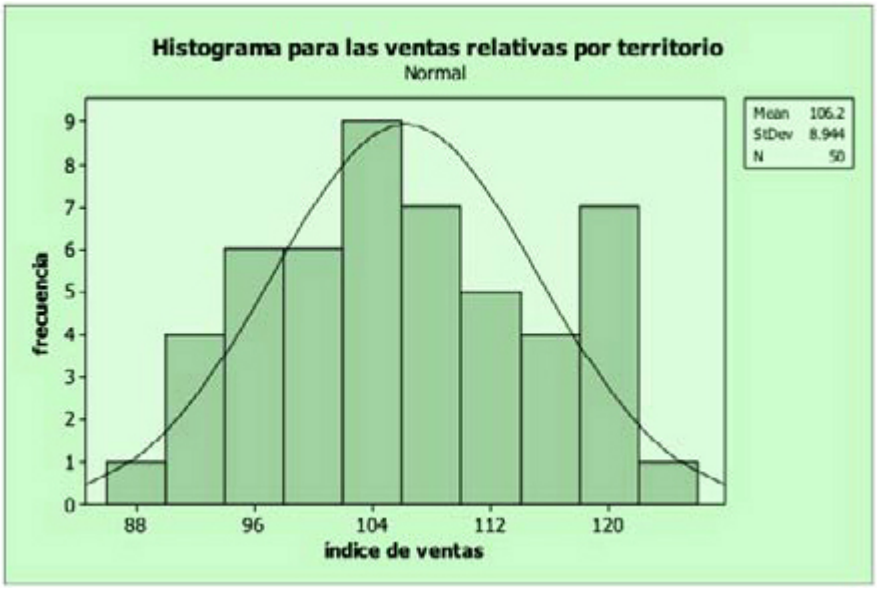


Figura 1.8 Histograma de las ventas por territorio para examinar normalidad

El histograma muestra que hay un grupo de observaciones con índices de ventas altas (barra a la derecha) que sobresale de la curva normal indicando que los datos se desvían de la normalidad.

Otra técnica gráfica empleada para examinar la normalidad de los datos, y en la cual la detección de datos extremos es más fácil, son los gráficos de probabilidad normal. En estos gráficos se comparan los percentiles de los datos muestrales con los percentiles correspondientes a la distribución normal. Cada pareja de percentiles es un punto en el sistema coordenado bidimensional; si los percentiles muestrales concuerdan con los de la distribución normal, los puntos se alinean alrededor de una línea recta en diagonal. El método visual para explorar normalidad consiste en examinar si los puntos están cerca de la línea recta en diagonal; toda desviación de la normalidad en los datos se reflejará en desviaciones respecto a la recta en diagonal. La Figura 1.9 muestra la gráfica de probabilidad normal de los residuos del modelo de regresión descrito previamente, el cual propuso explicar la inversión en obra pública y social de los estados mediante múltiples variables relacionadas con la generación de recursos del estado.

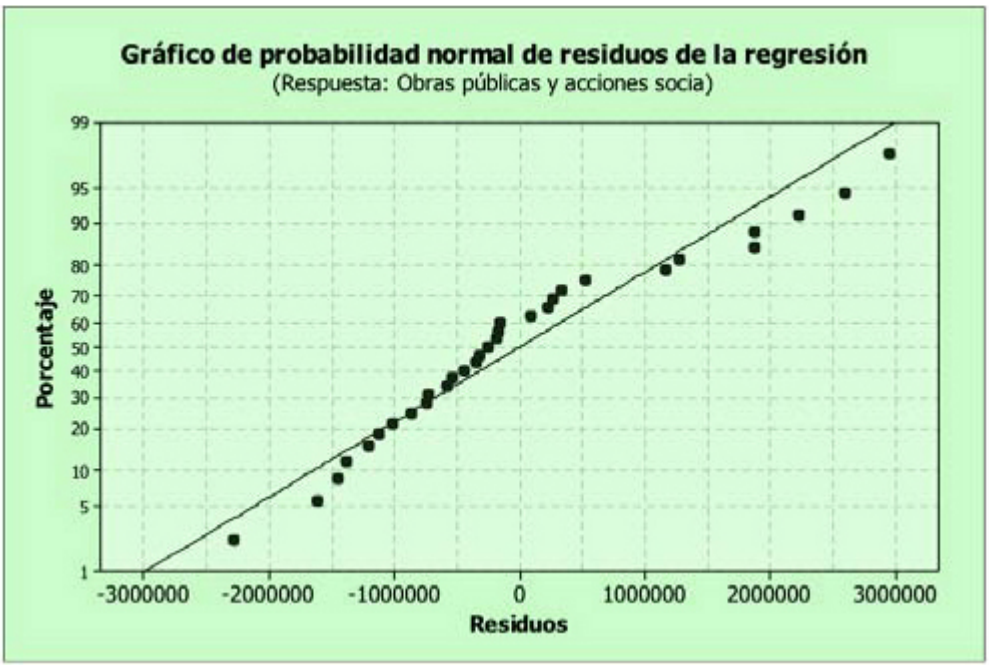


Figura 1.9 Gráfico de probabilidad normal para los residuos del modelo de regresión.

El gráfico muestra que los puntos no se alinean alrededor de la línea recta, por lo tanto el supuesto de normalidad es cuestionable. Cuando el gráfico no proporciona suficiente evidencia visual de falta de ajuste para el modelo normal se sugiere aplicar pruebas estadísticas formales para evaluar el supuesto. Estas pruebas contrastan la distribución empírica de los datos con la distribución normal y los detalles para su aplicación están disponibles en múltiples textos de estadística básica, entre ellos el de Devore (2007, sección 14.1). Las pruebas también están disponibles en el software estadístico comercial. En el caso de MINITAB que es el software que se utiliza como apoyo principal en este texto, hay que seleccionar los siguientes comandos del menú para realizar la prueba, previa captura de los datos (los residuos en el caso del ejemplo) en la hoja de trabajo:

Stat > Basic Statistics > Normality test

Hay varias opciones de pruebas de bondad de ajuste disponibles en MINITAB, entre ellas las de Kolmogorov-Smirnoff, Saphiro-Wilk y Anderson-Darling. Las hipótesis estadísticas correspondientes a estas pruebas son:

H_0: los datos se ajustan a la distribución normal

versus

H_1: los datos no siguen la distribución normal

Notar que la hipótesis alternativa no especifica la distribución de probabilidad de los datos, simplemente propone que ésta no es la normal. Para el ejemplo de los residuos del modelo de regresión sobre inversión en obra pública y social, el listado de salida es el gráfico de probabilidad normal de la Figura 1.10 el cual incluye en el cuadro superior derecho los resultados de la prueba estadística que para este ejemplo fue la de Anderson-Darling, la opción de default en MINITAB.

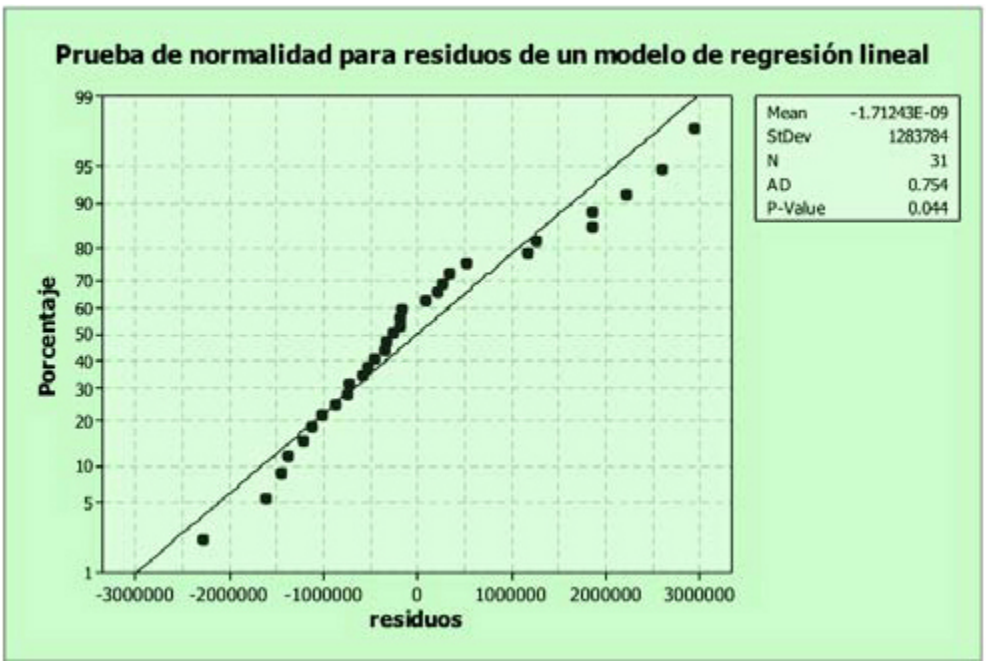


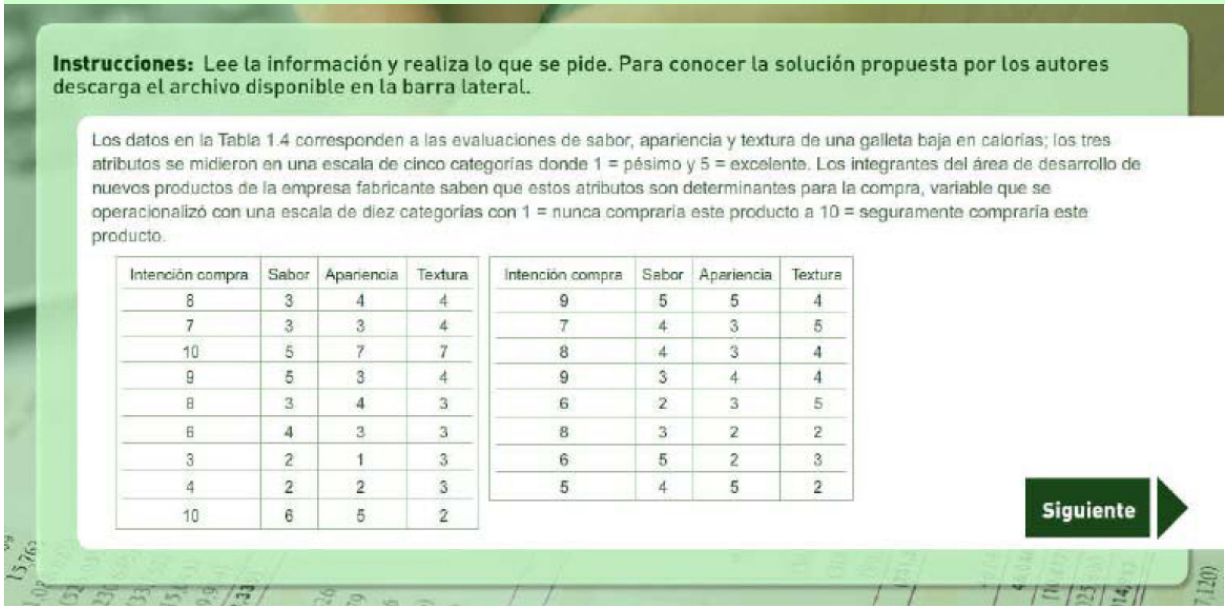
Figura 1.10 Resultados de la prueba de bondad de ajuste de Anderson-Darling.

La hipótesis nula se rechaza dado que el valor P = 0.044 asociado a la prueba indica que el riesgo de un error del tipo I–rechazo de la hipótesis siendo esta verdadera–es apenas del 4.4%. Puesto que los datos no son normales, las inferencias estadísticas construidas a partir del modelo de regresión lineal múltiple propuesto son cuestionables. El analista tendrá que decidir entre transformar los

datos para inducir normalidad–un recurso muy utilizado en estadística y en particular en regresión–o bien utilizar otros procedimientos que relajen el supuesto de normalidad. Una vez que se ha completado el examen exploratorio de los datos, y si estos tienen las características requeridas por los métodos multivariados que se desean utilizar, se procede a su análisis.

Ejercicio Integrador capítulo 1

1. Introducción al análisis multivariante



Instrucciones: Lee la información y realiza lo que se pide. Para conocer la solución propuesta por los autores descarga el archivo disponible en la barra lateral.

Los datos en la Tabla 1.4 corresponden a las evaluaciones de sabor, apariencia y textura de una galleta baja en calorías; los tres atributos se midieron en una escala de cinco categorías donde 1 = pésimo y 5 = excelente. Los integrantes del área de desarrollo de nuevos productos de la empresa fabricante saben que estos atributos son determinantes para la compra, variable que se operacionalizó con una escala de diez categorías con 1 = nunca compraría este producto a 10 = seguramente compraría este producto.

Intención compra	Sabor	Apariencia	Textura
8	3	4	4
7	3	3	4
10	5	7	7
9	5	3	4
8	3	4	3
6	4	3	3
3	2	1	3
4	2	2	3
10	6	5	2

Intención compra	Sabor	Apariencia	Textura
9	5	5	4
7	4	3	5
8	4	3	4
9	3	4	4
6	2	3	5
8	3	2	2
6	5	2	3
5	4	5	2

Siguiente

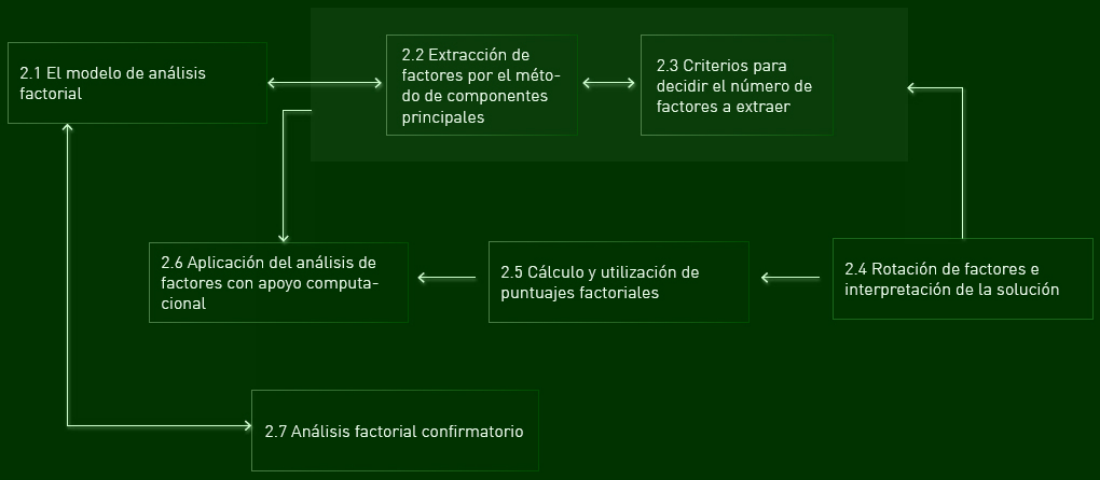
Conclusión capítulo 1

1. Introducción al análisis multivariante

En este primer capítulo se han expuesto los objetivos generales de los métodos de estadística multivariable así como los distintos esquemas para la categorización de tales métodos los cuales facilitan su comprensión. La representación y resumen de los datos multivariables se ha realizado empleando recursos de álgebra lineal; estos recursos seguirán siendo empleados a lo largo de este eBook. Si bien en este capítulo se realizó el cálculo de medidas descriptivas sólo con el apoyo de la calculadora, la complejidad de los métodos multivariables requiere del apoyo de software especializado. En los siguientes capítulos, en lugar de cálculos aritméticos detallados se aprovechará el recurso tecnológico orientándose la discusión de los análisis a la interpretación de los listados de salida proporcionados por el software estadístico.

Capítulo 2
Análisis de factores

Organizador temático



2. Análisis de factores

Introducción

"You can't fix what you can't measure" ésta es una de las citas famosas del Dr. Deming, reconocido académico en el área de control de calidad durante los años 80, quien a través de esta frase expresó su preocupación de que las decisiones administrativas se apoyaran en medidas válidas para las variables e indicadores críticos de la empresa. En el capítulo anterior se hizo mención del concepto de calidad del servicio, sin una métrica adecuada para este concepto ¿cómo puede una organización decidir si sus empleados de servicio requieren capacitación? ¿cómo puede evaluar sus políticas de servicio? ¿cómo puede compararse con su competencia en cuanto a la calidad del servicio que ofrece? El diseño de buenos instrumentos de medición es esencial en cualquier proyecto de inteligencia de mercados ya que los datos que

se generan a través de ellos, son la base para el diagnóstico, la evaluación y la elección de alternativas de acción.

El análisis factorial es uno de los métodos de estadística multivariable que apoya el diseño de medidas válidas para los conceptos de inteligencia de mercados. Costello y Osborne (2005) reportan que durante el período de 2002-2004 más de 1,700 estudios en las áreas de ciencias sociales utilizaron el análisis factorial para diseñar y evaluar instrumentos de medición. A través de este análisis es posible determinar si un concepto es multidimensional, identificar las dimensiones o componentes del concepto así como asegurar la validez convergente y divergente de multi-escalas elaboradas para evaluar conceptos complejos que pudieran confundirse entre sí. El estudio que se resume en el siguiente cuadro, es un ejemplo de cómo el análisis factorial ayudó al diseño de una métrica para la calidad del servicio hotelero en la Isla Mauricio (Juwaheer, 2004).

En el contexto de mercadotecnia, la calidad del servicio se ha definido como el cumplimiento de las expectativas de servicio de los clientes siendo el instrumento de medición conocido como SERVQUAL el recurso más difundido para su medición. Este instrumento contempla cinco dimensiones o componentes críticos que determinan la percepción de servicio de un cliente, sin

embargo estas dimensiones no son generalizables a todos los tipos de servicio.

En el caso de servicios de hospedaje, la seguridad de las instalaciones, la conveniencia de su ubicación, la facilidad de acceso y la identidad corporativa del hotel también han resultado ser componentes críticos de la calidad del servicio.

Una encuesta basada en el SERVQUAL y aplicada a 410 turistas internacionales que se hospedaron en los hoteles de las playas de la Isla de Mauricio, proporcionó datos multivariables que se utilizaron como entrada para un Análisis Factorial Exploratorio.

Gracias al uso de esta técnica estadística, fue posible identificar nueve dimensiones para el concepto de calidad del servicio de hotelería. Estas dimensiones o componentes fueron: confiabilidad del servicio, actitud del staff, atractivo de las habitaciones, lo llamativo de los alrededores y el ambiente del hotel, aseguramiento, servicios adicionales disponibles en la habitación, comunicación con el personal, empatía y servicios complementarios.

Los primeros cuatro componentes explican la mayor parte de la variabilidad en las impresiones general de calidad de servicio de los huéspedes así como su satisfacción con el servicio ofrecido. Los administradores de los hoteles en Isla Mauricio podrán utilizar estos resultados para evaluar el servicio que brindan al turismo internacional y también enfocar sus esfuerzos en cumplir las promesas hechas a clientes y responder efectivamente a sus peticiones puesto que es esta dimensión de confiabilidad la que el cliente más valora.

2.1 El modelo de análisis factorial

El análisis factorial o de factores del tipo exploratorio (AFE = Análisis Factorial Exploratorio) es un procedimiento complejo que incluye varias etapas y que permite agrupar variables con base en sus patrones de asociación. Tal agrupamiento se realiza con los siguientes objetivos:

Explorar los patrones de soporte o relación de un gran número de variables

Determinar si la información respecto a este “gran” número de variables observables se puede condensar o resumir en un número menor de componentes o factores latentes

Comprobar la multi-dimensionalidad de conceptos complejos en las áreas de ciencias sociales y administración

Apoyar en la determinación de la validez de escalas de medición

Dada la naturaleza exploratoria de sus objetivos, el AFE requiere que el analista se apoye en su experiencia práctica y conocimientos sobre el problema de inteligencia de mercados para interpretar y utilizar sus resultados. El modelo que presupone el análisis factorial exploratorio se describe gráficamente en el diagrama de vías de la Figura 2.1.

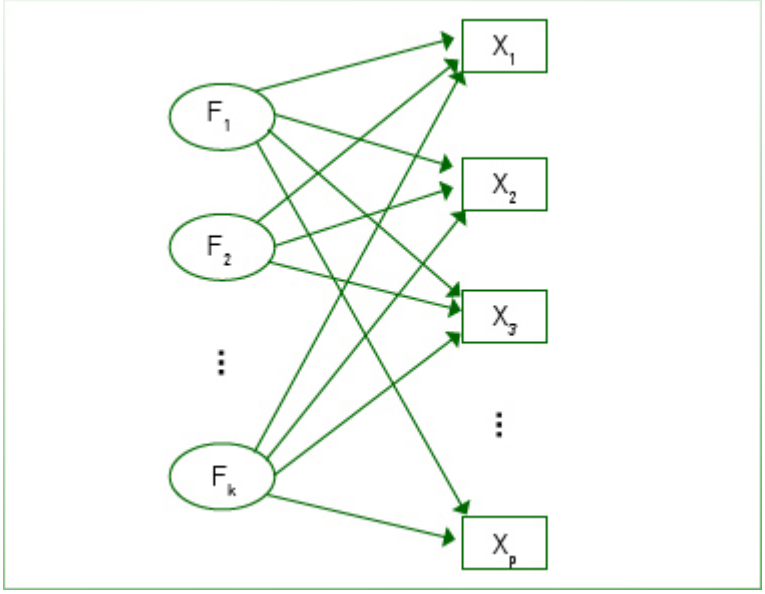
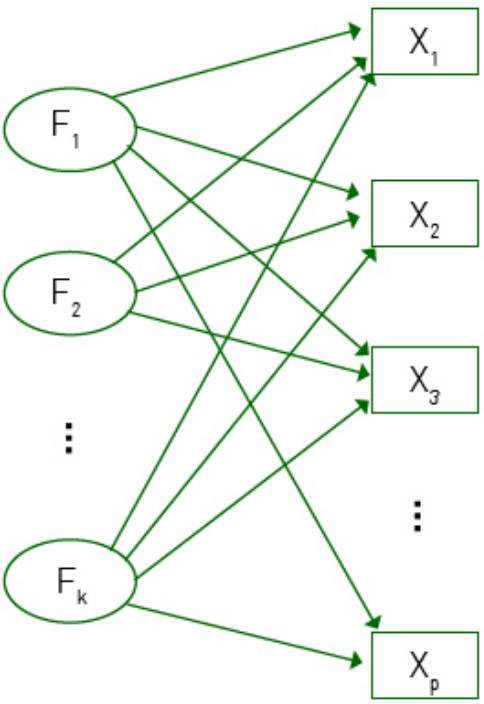


Figura 2.1 Diagrama de vías para el análisis factorial exploratorio



El modelo anterior propone que las p variables registradas (X_i) son indicadores tangibles de un mismo conjunto de variables/dimensiones latentes o factores comunes (F_j) cuyo número k<<p no está especificado sino que es sugerido a través del análisis.

Algebraicamente, el modelo anterior se expresa como sigue:

$$x = \mu + LF + \varepsilon \quad (1)$$

Donde

x = vector de Xi variables observables (i = 1,2,...,p)

μ = vector px1 de las medias de las variables observables

L = matriz de "cargas" o grado de asociación de la variable i con el factor j

F = vector de Fj factores no observables o latentes (j= 1,2,...,k ↓↓ p)

ε = vector de p errores aleatorios

Para definir qué variables están asociadas a cuáles factores no observables o latentes es necesario estimar la **matriz factorial** *L*. Esta matriz tiene como entradas a las "cargas" l_{ij} (*loadings* en inglés) las cuales equivalen a coeficientes de correlación entre las variables observables y los factores latentes que se desea identificar. Si una

variable tiene una carga alta (cercana a 1 en valor absoluto) en un factor, se tomará como un indicador tangible de éste. Todas aquellas variables con cargas relevantes en el mismo factor constituyen su grupo de indicadores manifiestos. Por tanto, estimar las cargas es la primera etapa del análisis factorial.

El método de mínimos cuadrados ordinarios, que se utiliza generalmente para estimar los coeficientes de un modelo de regresión lineal, no es aplicable en el caso del modelo factorial ya que en la expresión (1), las entradas para la matriz ***F***, es decir los factores no son directamente observables, solo se tienen datos para los indicadores manifiestos que integran la matriz ***X*** ¿cómo proceder entonces? Para resolver el problema de estimación de cargas se considera la estructura de la matriz de varianza-covarianza de las *X*, la cual, de acuerdo a lo descrito en el primer capítulo, contiene toda la información sobre los patrones de asociación entre estas variables manifiestas. Del modelo de análisis factorial en (1) se puede determinar la siguiente estructura para la matriz de varianza-covarianza:

$$Var(x) = \sum = LL' + \Psi \qquad (2)$$

De la expresión anterior se deduce que para una única variable X_i, $Var(X_i)$ = donde l_{ij} es la carga de la variable X_i en el factor *Fj*, la suma de cuadrados de todas las cargas para la *i*-ésima variable sobre todos los factores se denomina comunalidad mientras que el término ψ_i se conoce como la varianza específica. La **comunalidad** es aquella porción de la varianza de X_i que es común o compartida con el resto de las variables *X*'s (1, 2, … *i*-1, *i* + 2, … *p*) debido a que todos los indicadores tangibles se asocian con los mismos factores comunes. La comunalidad se interpreta por tanto como un coeficiente de determinación múltiple entre el indicador X_i y las otras *p*-1 variables manifiestas. Cuando la comunalidad de un variable observada *X* es muy baja, el analista infiere que esta variable no covaría con el resto porque no es un indicador manifiesto de los mismos factores comunes a los que están asociadas las otras variables. Por otra parte, el componente de varianza específica

representa aquella porción de la variabilidad de X_i que es particular a esta variable y que resulta de errores de medición y de la variación en los datos de muestra a muestra.

Una vez explicada la estructura de la varianza de las X se regresará al problema de cómo estimar las cargas. Varios métodos han sugerido para completar esta parte del análisis, los más difundidos son Componentes Principales (CP) y **Máxima Verosimilitud**. En la siguiente sección se describen ambos métodos con énfasis especial en el **análisis de componentes principales** por ser el método más simple y rápido de usar, de carácter netamente exploratorio y además la opción por default en la mayor parte de los paquetes estadísticos.

¿Sabías qué?

Charles Spearman (1863-1945), psicólogo inglés, es considerado el creador del análisis factorial gracias a su trabajo pionero sobre la Teoría Bifactorial de la Inteligencia. Después de observar las fuertes correlaciones positivas entre los resultados obtenidos en distintas pruebas de habilidad mental–habilidad matemática, artística, lógica, verbal–aplicadas a niños, Spearman formuló la hipótesis de que estos resultados podían explicarse por un “factor” general de inteligencia al que llamó g.

2.2 Extracción de factores por el método de componentes principales

Los componentes principales son aquellas combinaciones lineales no correlacionadas en las que sus varianzas son lo más grande posible. Bajo el supuesto de normalidad, la dirección de estos componentes principales corresponde a los ejes de un elipsoide de densidad constante. Las direcciones de los ejes coinciden con la dirección de máxima variabilidad de los datos mientras que la suma de las distancias perpendiculares al cuadrado de los datos, respecto a estos ejes de máxima variabilidad, es un mínimo. En dos o tres dimensiones, la condición de que los componentes principales estén no correlacionados implica que estos son perpendiculares u ortogonales, es decir que entre ellos hay un ángulo de 90°. En la Figura 2.2 se describen gráficamente estas características de los componentes principales

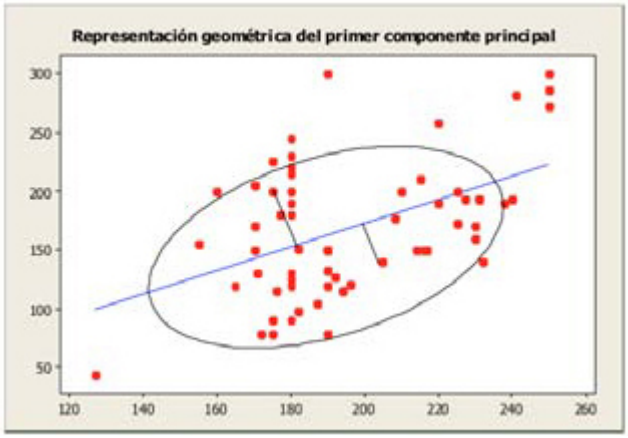


Figura 2.2 Geometría de los componentes principales

El problema de encontrar las combinaciones lineales $Y = l'\boldsymbol{x}$ de máxima varianza, denominados componentes, principales se formula como sigue:

$$max\ Var(Y) = max\ l'Cov(x)l$$

$$sujeto\ a\ \ l'l = 1 \qquad (3)$$

La solución al problema anterior es: $Y_1 = e_1{}'\boldsymbol{x}$, con $Var(Y_1) = \lambda_1$ donde λ_1 es el **eigen-valor** más grande de la matriz de varianza-covarianza $Cov(\boldsymbol{x})$ y e_1 el eigen-vector normalizado, esto significa que el vector tiene longitud o norma igual a uno como lo establece la condición dada en (3)- asociado al primer (el mayor) **eigen-valor**. El término “eigen” es una palabra en alemán que al español se traduce como propio, por lo cual la pareja (λ_1, e_1) también se conoce como el primer valor y vector propios o característicos de la matriz $Cov(\boldsymbol{x})$. En álgebra lineal, los vectores propios (e) son vectores que al ser transformados, es decir cuando se multiplican por una matriz cuadrada cualquiera A, no cambian su dirección sino que solo resultan multiplicados por un escalar usualmente denotado por λ y que es denominado el eigen-valor o valor propio o característico. En términos algebraicos esto queda expresado como $Ae = \lambda e$.

¿Sabías qué?

Los eigen-valores y eigen-vectores son de utilidad no solo en el área de estadística sino también en las de ingeniería y ciencias, por ejemplo en el estudio y diseño de estructuras y la identificación de pozos petroleros.

Una matriz de orden pxp como es el caso de *Cov(x)* tiene hasta p componentes principales. Cada uno de estos componentes corresponde a los eigen-vectores ortonormalizados, esto es de longitud uno y ortogonales entre sí, de la matriz y sus varianzas decaen en el orden de magnitud de sus eigen-valores $\lambda_1<\lambda_2<\lambda_3<\ldots<\lambda_p$. La estructura de la matriz de varianzacovarianza $\Sigma=Cov(x)$ se puede explicar o reconstruir totalmente a través de estos p componentes, sin embargo el objetivo del análisis de componentes principales (ACP) es explicar la estructura de Σ con solo unos cuantos de ellos ($k<<p$). Esto significa que el método se orienta a lograr una reducción de la dimensionalidad sin pérdida importante de la información contenida en la matriz de varianza-covarianza.

La proporción de la varianza total asociada o "explicada" por el *j*-ésimo componente principal está dada por:

$$\frac{j\text{-ésimo eigen-valor de } Cov(\boldsymbol{x})}{\text{suma de eigen-valores (traza de la matriz)}} \qquad (4)$$

Es importante notar que el análisis de componentes principales no es un auténtico de análisis factorial en el sentido de que el ACP se realiza sin asumir un modelo de medida como el descrito en la Figura 2.1. Los componentes se calculan considerando la varianza de los indicadores manifiestos bajo el supuesto de que la varianza específica es solo variabilidad aleatoria por lo cual solo la varianza compartida o comunalidad figura en la solución (Fabrigar, Wegener, MacCallum y Strahan, 1999). Por esto el ACP tiende a resultar en un porcentaje de varianza explicada o reconstruida mayor que la reportada con otros métodos.

Ejemplo 1: Cabe recordar el ejemplo del capítulo anterior en el cual se calcularon estadísticos descriptivos para los datos multivariables que representan el desempeño de una muestra de estudiantes en un curso sobre métodos multivariados. Usando el software estadístico MINITAB (Minitab, 2003) se calcularon los componentes principales de la matriz de correlación **R.**

MINITAB muestra en una misma pantalla la ventana de la sesión en la cual se reportan los resultados de los análisis estadísticos y la hoja que contiene las columnas con los datos a procesar (ver Figura 2.3). Desde el menú de opciones el usuario puede elegir la siguiente secuencia de comandos para completar el análisis de componentes principales:

Una vez abierta la ventana de diálogo correspondiente, la cual se incluye en la Figura 2.3, basta con declarar cuáles son las columnas que contienen a los datos. El programa se encarga de calcular primero la matriz de correlación para después extraer sus componentes principales. Los siguientes resultados se despliegan en la ventana de la sesión de MINITAB según se muestra, también, en la Figura 2.3.

Variable	PC1	PC2	PC3
Examen práctico	0.696	-0.077	0.714
Proyecto	0.669	-0.291	-0.684
Examen final	0.260	0.954	-0.151
Eigenvalue	1.9758	0.9505	0.0737
Proportion	0.659	0.317	0.025
Cumulative	0.659	0.975	1.000

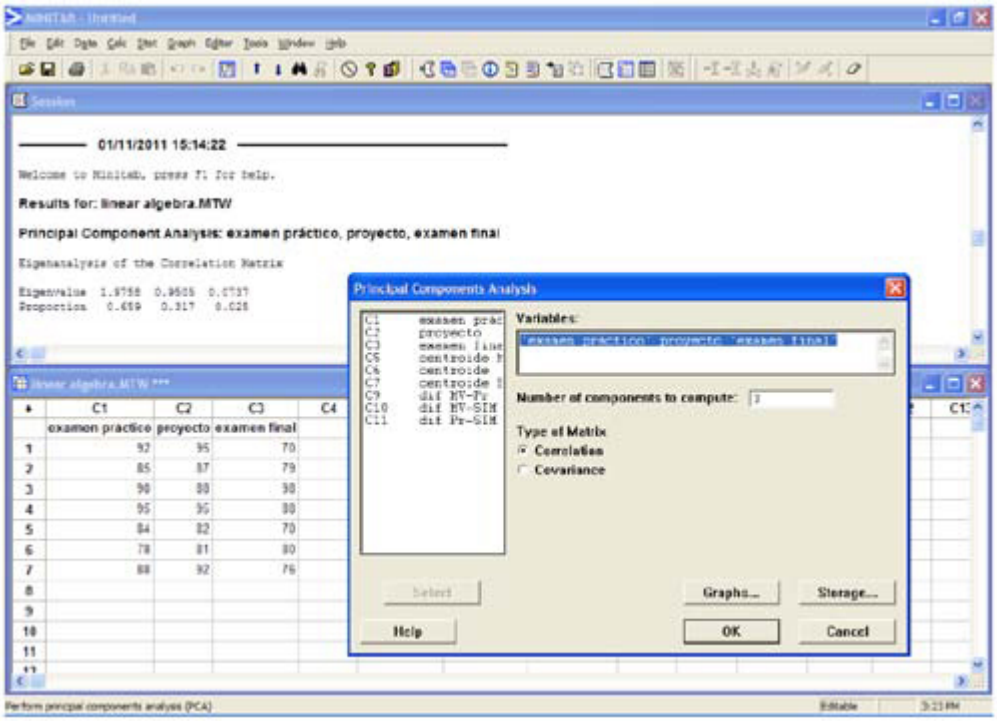


Figura 2.3 Análisis de componentes principales empleando MINITAB

De los resultados anteriores se puede apreciar que el uso de un único componente principal (*PC*1 = Y_1) reconstruye el 65.9% ($\lambda_1/3$ x 100 = 1.9758/3 x 100) de la varianza original de los tres indicadores. Al utilizar dos componentes principales (*PC*1 *y PC*2) se reproduce hasta un 97.5% de la varianza de los tres indicadores estandarizados. Esto implica que la matriz de correlación original se puede aproximar cercanamente mediante el siguiente producto: R~PΛP' donde *P* es la matriz que tiene por columnas a los dos primeros componentes principales extraídos; *Λ* es una matriz diagonal cuyas entradas son los eigen-valores asociados con los dos componentes principales y *P'*es la transpuesta de la matriz *P*. Este producto resulta en la siguiente aproximación para la matriz de correlación:

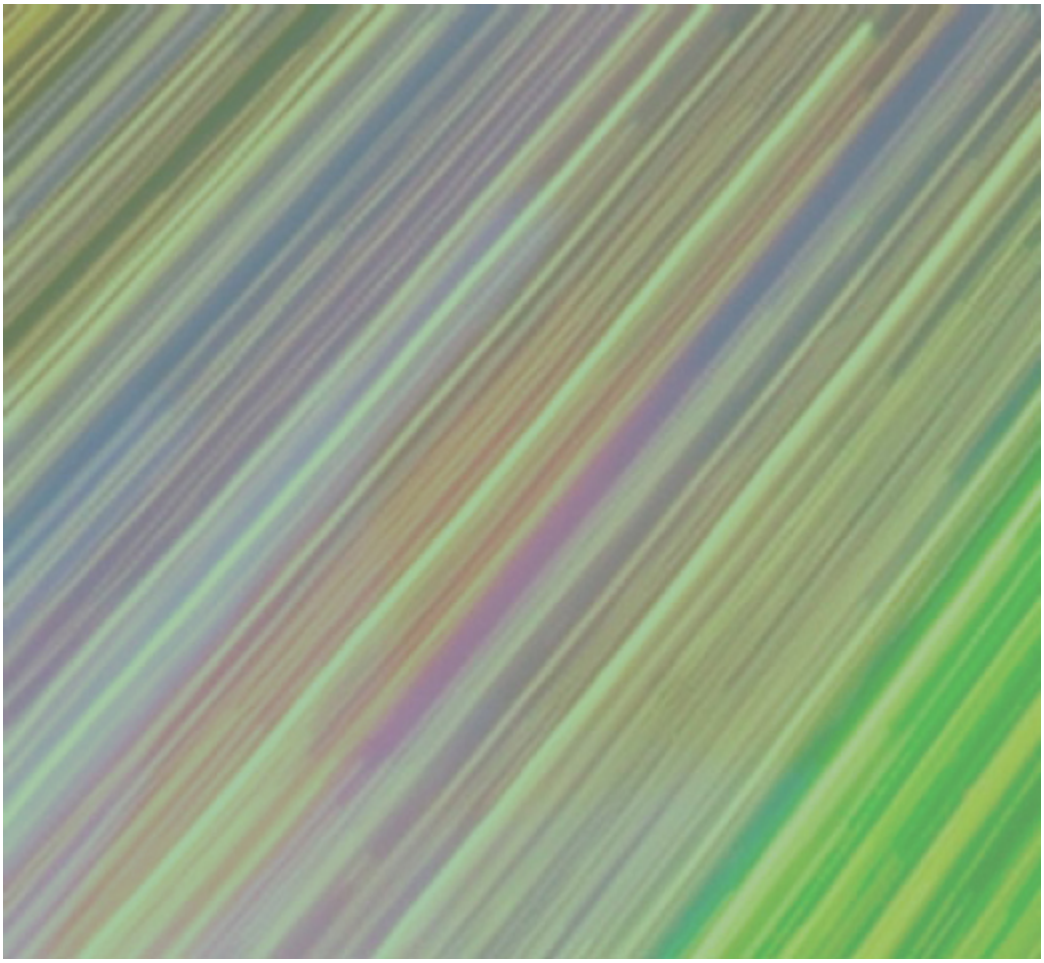
$$R = \begin{bmatrix} 1.000 & .905 & .281 \\ .905 & 1.000 & .088 \\ .281 & .088 & 1.000 \end{bmatrix} \cong \begin{bmatrix} .696 & -.077 \\ .669 & -.291 \\ .260 & .954 \end{bmatrix} \begin{bmatrix} 1.9758 & 0 \\ 0 & .9505 \end{bmatrix} \begin{bmatrix} .696 & .669 & .260 \\ -.077 & -.291 & .954 \end{bmatrix}$$

$$R \cong \begin{bmatrix} .940 & .904 & .352 \\ .856 & .823 & .320 \\ .568 & .546 & .212 \end{bmatrix}$$

A través del APC se logra también una simplificación en la estructura de los datos ya que la información de cada individuo (estudiante) quedaría representada no por tres variables tangibles (calificación de parciales, calificación del proyecto y examen final) sino por solo dos combinaciones lineales (los componentes principales) de estas variables las cuales, bajo el modelo factorial, se asocian con dimensiones latentes.

Otro método que se puede utilizar para estimar o extraer los factores subyacentes a un conjunto de indicado res manifiestos es el método estadístico de **Máxima Verosimilitud** (MV). Este método propone encontrar aquellos valores posibles de los parámetros de la distribución de probabilidad que generó los datos de la muestra. Con MV, los estimadores de los parámetros de interés son aquellas funciones de los datos muestrales que maximizan la llamada función de verosimilitud; para construir esta función se requiere de especificar la distribución de probabilidad de los datos. En el primer capítulo de este eBook se indicó que los métodos multivariantes asumen con frecuencia que esta distribución de probabilidad es la normal. Bajo el supuesto de normalidad para el vector $\boldsymbol{x}$ la función de verosimilitud se especifica como $L(\boldsymbol{\mu}, \sum \mid \boldsymbol{x})$ donde $\boldsymbol{\mu}$ y $\sum$ son los parámetros que se desea estimar. De acuerdo con el modelo factorial descrito en (1), el cual asume que hay exactamente k factores y bajo el supuesto de que las variables observadas ($\boldsymbol{x}$) y latentes (F) siguen la distribución normal multivariada, se tiene que $\sum = LL' + \Psi$. El objetivo es entonces estimar a la matriz factorial o de cargas $\boldsymbol{L}$ de tal forma que se tenga una alta probabilidad de reproducir las entradas observadas en la matriz de varianza-covarianza muestral o en la de correlación (Kim y Mueller, 1978).

Los estimadores obtenidos bajo este método tienen propiedades es tadísticas deseables como la de eficiencia, por lo cual MV se recomienda sobre componentes principales siempre y cuando el supuesto de normalidad se satisfaga, lo cual, lamentablemente, no ocurre con mucha frecuencia para datos recolectados en estudios de inteligencia de mercados. En MV la varianza específica (ψ) de cualquier variable observable se asume como resultado de variación aleatoria por lo cual el método asignará mayor peso a las variables con mayores comunalidades (Johnson y Wichern, 1998). Maximizar la función de verosimilitud es una tarea difícil y no hay fórmulas cerradas disponibles, la estimación de las entradas de la matriz de cargas *L* se realiza mediante un procedimiento iterativo de búsqueda, es decir se proponen valores iniciales para las entradas de esta matriz y éstos se refinan con cada iteración hasta que se alcancen criterios de precisión preestablecidos.



Mínimos cuadrados	Propone minimizar las diferencias entre las correlaciones observadas y aquélla reproducida por los factores extraídos. El procedimiento inicia asumiendo k factores y proponiendo valores iniciales para las comunalidades para después extraer los factores que reproduzcan mejor la matriz de correlación. Las comunalidades son reestimadas sucesivamente hasta que no se puedan decrecer más las diferencias.
Método de extracción alfa	Este método tiene un enfoque psicométrico por lo que asume que las variables observadas son solo una muestra de una población de indicadores que se observa para una población de individuos. En el método alfa, el número de factores a utilizar se determina con base en el criterio de que los eigenvalores sean mayores de uno porque esto es equivalente a proponer que el coeficiente de generalización alfa para el universo de variables es mayor de uno. Los pesos factoriales se estiman de tal manera que los factores extraídos tengan correlación máxima con el total de factores comunes.
Ejes principales	Este es un método iterativo de estimación similar al de componentes principales pero en lugar de usar una matriz de correlación que tiene "1s" en su diagonal, sustituye estas entradas por las comunalidades de los indicadores manifiestos. A partir de esta matriz de correlación reducida se calculan los eigenvalores y eigenvectores los cuales se usan para re-estimar las comunalidades; las iteraciones terminan cuando estos estimados ya no cambian.

El lector puede ampliar esta descripción a través de la consulta de la monografía de Kim y Mueller (1978) y el capítulo elaborado por Vicente y Oliva y Manera-Bassa (2003).

Actividad de repaso

2.2 Extracción de factores por el método de componentes principales

RELACIONAR COLUMNAS

En análisis factorial con componentes principales, esta cantidad corresponde a la varianza de cada uno de los factores extraídos	○	○	Componentes principales
Es la porción de la varianza que una variable tiene o comparte en común con otras variables debido a la existencia de factores o variables latentes comunes	○	○	Eigen-valor/Suma de eigenvalores
Conjunto de combinaciones lineales de las variables observables ortogonales entre sí tal que sus varianzas sean lo más grande posible	○	○	Ortogonales
En términos geométricos significa que los componentes principales guardan entre sí un ángulo de 90°	○	○	Máxima verosimilitud
Matriz que contiene la información sobre la variabilidad y las asociaciones entre los indicadores manifiestos	○	○	Eigen valor
Es la contribución que hace cada componente principal para explicar la varianza original de los indicadores manifiestos	○	○	Simplificación de la estructura de datos
Método estadístico para la extracción de factores que asume normalidad para los datos	○	○	Comunalidad
Son equivalentes a coeficientes de correlación entre las variables observables y los factores latentes	○	○	Varianza-covarianza
Enfoque básico del método de componentes principales	○	○	Eigen vector
Es un vector que al ser multiplicado por una matriz no cambia su dirección sino que solo resulta multiplicados por un escalar	○	○	Cargas

2.3 Criterios para decidir el número de factores a extraer

Como el análisis factorial es un método exploratorio, la decisión respecto al número de factores a extraer se apoya en varios criterios heurísticos. Solo en el caso en que Máxima Verosimilitud haya sido el método de estimación, se cuenta con una prueba estadística para determinar la calidad del ajuste alcanzado para una solución en k factores.

Los criterios recomendados para decidir cuántos factores utilizar son:

- Gráfico de sedimentación o scree-test.
- Valores propios mayores de uno.
- Porcentaje de varianza explicada.

El primer criterio utiliza como base un gráfico conocido como gráfico de sedimentación o *scree plot* en inglés. En el eje horizontal del gráfico se muestra el número de factor que se ha extraído (k = 1, 2, 3,… p) y en el vertical el eigen-valor o valor propio correspondiente a cada factor. La varianza de los primeros factores, la cual está medida por sus eigen-valores, es muy grande en comparación a la de los últimos factores, de tal manera que el gráfico dará la impresión de consistir de dos líneas: una con pendiente negativa, decreciente, que describe el decaimiento gradual en las varianzas de los primeros factores y otra casi horizontal en la que se alinean las varianzas de los últimos factores. El criterio consiste en examinar el gráfico e identificar el punto anterior a aquel donde hay un “quiebre” en la gráfica, este punto de ruptura o quiebre es donde la recta inclinada se corta con la aparente línea horizontal del gráfico; el punto inmediato anterior corresponde al número factores a extraer. La tarea de identificar este punto se dificulta cuando hay muchos puntos alrededor del punto de quiebre (Costello y Osborne, 2005).

El segundo criterio utilizado para definir el número de factores a extraer es el del eigen-valor mayor de uno ($\lambda_1 < 1$) este criterio es comúnmente utilizado cuando se usan componentes principales como método de estimación. Solo aquellos factores cuya varianza sea mayor de uno serán considerados en la solución. La lógica del criterio es que si la varianza de un indicador manifiesto estandarizado es igual a uno, la varianza de una combinación lineal de los indicadores, es decir de un componente principal asociado a un factor latente, deberá ser mayor que la del indicador individual ya que, de otra manera, no tiene sentido considerar un factor que explica menos varianza de la que contiene un indicador individual. Este criterio es el default en paquetes estadísticos como SPSS, sin embargo es un criterio poco preciso y que tiende a sugerir más factores de los que conviene a la estructura del modelo de medida de la Figura 2.1 (Fabrigar et al. 1999; Costello y Osborne, 2005).

El último criterio empírico es el de varianza explicada, bajo este criterio el objetivo es que el conjunto de factores explique una porción relevante de la variabilidad original de los indicadores manifiestos ya que la variabilidad y asociaciones de estos indicadores observables se atribuye a los factores latentes. No hay un valor específico para establecer el porcentaje de varianza que debería explicar la solución en k factores; el analista tendrá que considerar los objetivos del análisis para justificar su decisión.

Como se mencionó previamente, aparte de la calidad estadística de los estimados que produce, el método de máxima verosimilitud permite probar qué tan apropiada, en términos de si ajusta bien a los

datos, es la solución en k factores. La formulación de esta prueba es como sigue:

H_0: La estructura en exactamente *k* factores describe *Cov(x)* versus

H_1: La estructura en exactamente *k* factores no describe a *Cov(x)*

$$-2ln\boldsymbol{\Lambda} = n\,(ln\,|\boldsymbol{LL'} + \boldsymbol{\psi}\,| - ln|\boldsymbol{S}| + tr(\boldsymbol{S}(\boldsymbol{LL'} + \boldsymbol{\psi})\,-1\,-p) \sim \chi 2\,(v) \qquad (5)$$

donde

n = total de datos

S = matriz de varianza-covarianza muestral

L = matriz de cargas estimadas

ψ = matriz que en su diagonal contiene a las comunalidades de cada variable.

x^2 se refiere a la distribución *ji*-cuadrada con $v = ½\,[(p\text{-}m)\,2\text{-}(p+m)]$ grados de libertad.

La región de rechazo para un nivel de significancia predeterminado igual a α son todos aquellos valores del estadístico de prueba χ^2 que excedan el porcentil 1-*α* de la distribución *ji*-cuadrada con v grados de libertad. Si la hipótesis nula se rechaza, se procede a incrementar el número de factores y repetir la prueba. Esta prueba estadística está disponible en software comercial como SPSS (Pardo-Merino y Ruíz-Díaz, 2005).

Ejemplo 2: Se realizó una encuesta entre 100 egresados de las universidades ubicadas en la zona metropolitana de la ciudad de Toluca con el objetivo de identificar sus aspiraciones profesionales y las características de su trabajo ideal. Once reactivos se utilizaron para determinar el perfil de este trabajo ideal; cada respondiente indicó en una escala de diez categorías (1 = nada importante a 10 = indispensable) qué tan importante era la característica descrita en

su elección de un puesto de trabajo. El contenido resumido de los reactivos empleados es el siguiente:

*X*1 = Un trabajo desafiante donde pueda desarrollar mis competencias

*X*2 = Un trabajo en el cual mis superiores faciliten mis tareas

*X*3 = Un lugar de trabajo con un ambiente de colaboración

*X*4 = Un trabajo en el cual los objetivos de desempeño estén claramente definidos

*X*5 = Un trabajo en que haya oportunidades de concursar por otros puestos

*X*6 = Un trabajo que ofrezca prestaciones adicionales atractivas

*X*7 = Un lugar de trabajo en donde el personal de apoyo sea confiable

*X*8 = Un trabajo donde la selección del personal se la adecuada

*X*9 = Un trabajo con responsabilidades bien definidas

*X*10 = Que las demandas del trabajo no afecten negativamente mi calidad de vida

*X*11 = Un trabajo donde haya oportunidades de desarrollo profesional

En este ejemplo se presupone un conjunto de dimensiones subyacentes al concepto de “trabajo ideal” las cuales se desea identificar a través del análisis factorial. En este ejemplo el Análisis Factorial se realizará empleando el software SPSS. Es importante que el lector se familiarice con diferentes paquetes estadísticos, los métodos de análisis multivariante requieren de apoyo computacional pero su aplicación no debe estar supeditada al dominio de un

paquete estadístico. Para realizar el análisis en SPSS hay que usar la siguiente secuencia de comandos:

Analyze → Data reduction → Factor Analysis

Abierta la ventana de diálogo que se muestra en la Figura 2.4 hay que declarar las columnas que contienen los datos de las variables observables; a partir de estos datos el programa calculará la matriz de varianza-covarianza o la de correlación. Se recomienda factorizar la matriz de correlación ya que los indicadores observables tienen distintos grados de variabilidad y pueden estar expresados en diferentes unidades, al estandarizarlos se facilita la interpretación de sus relaciones y la evaluación del efecto de observaciones extremas (Fung y Kwan, 1995). En general los resultados del análisis si se usa la matriz de varianza-covarianza o la de correlación serán diferentes.

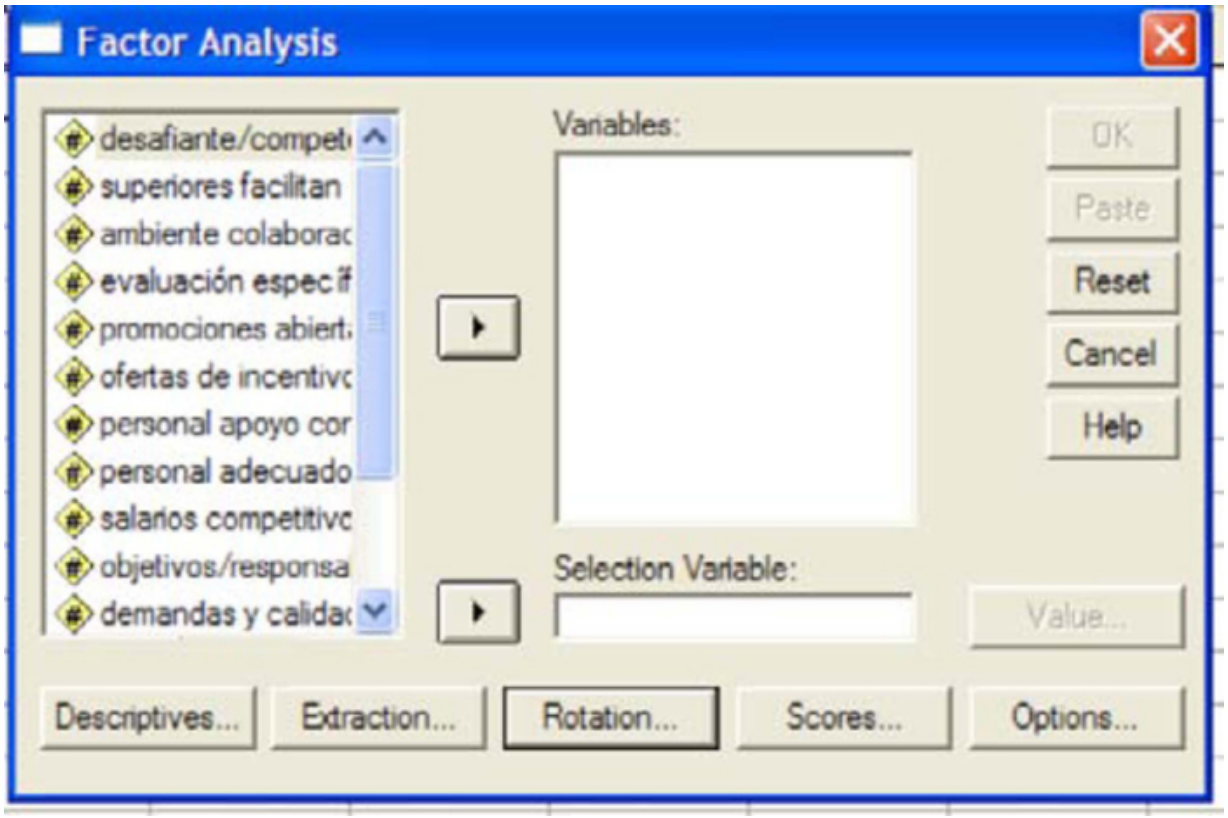
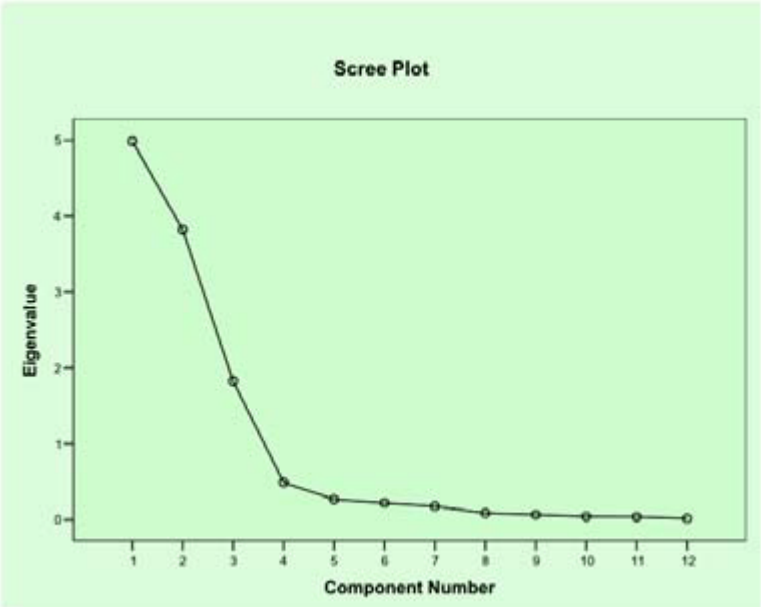


Figura 2.4 Ventana de diálogo básico para El Análisis Factorial en SPSS

Como puede observarse al acceder a la opción Extraction en la ventana de diálogo anterior, el método de extracción de factores que está predeterminado en SPSS es el de Componentes Principales, si se desea utilizar Máxima Verosimilitud solo hay que elegir la opción correspondiente. El default del programa es extraer únicamente aquellos factores con eigen-valores o varianzas mayores de uno; dados los inconvenientes de esta opción, que se comentaron antes, se sugiere elegir el número de factores en términos del referente teórico del estudio, esto es ¿cuántas dimensiones latentes integran el concepto estudiado de acuerdo con la literatura sobre el tema? o bien especificar k = p/2 para así asegurarse de haber extraído todos los factores relevantes y a partir de ahí proceder a definir un número razonable de factores.

En la ventana de diálogo principal de SPSS también aparece la opción de **Rotación** para la solución, *Rotation*, esta parte del análisis factorial se discute en la siguiente sección. Para fines de este ejemplo es apropiado aceptar la opción *None* que es el default ya que rotar sin haber es pecificado el número de factores no tiene sentido. Al elegir la opción Scree plot se obtendrá el gráfico siguiente, a partir del cual se puede sugerir el número de factores a extraer. porción numérica del listado se muestra a continuación:



Interpretación:

Después de analizar el scree-plot, se recomienda una solución en k = 3 factores.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.983	41.522	41.522	4.983	41.522	41.522
2	3.817	31.809	73.330	3.817	31.809	73.330
3	1.822	15.184	88.514	1.822	15.184	88.514
4	.488	4.070	92.584			
5	.264	2.202	94.786			
6	.217	1.811	96.597			
7	.174	1.454	98.051			
8	.084	.698	98.749			
9	.063	.523	99.272			
10	.039	.326	99.597			
11	.035	.290	99.887			
12	.014	.113	100.000			

La primera parte del listado numérico muestra el porcentaje de varianza que se logra reproducir o explicar a medida que se van extrayendo más factores. La primera columna son los eigen-valores o varianzas estimadas de cada componente principal, las siguientes dos columnas son las contribuciones porcentuales netas y acumuladas que cada factor hace la varianza total. De acuerdo con el criterio del eigen-valor, el número de factores que conviene extraer es de *k* = 3 lo que coincide con el scree-plot. Si

se utilizan solo tres factores, se logra reproducir o explicar hasta un 88.5 % de la varianza original (41.5 + 31.8 + 15.2) lo cual es bastante aceptable.

	Component		
	1	2	3
desafia	.073	-.145	.491
supfacil	.731	.557	.195
colabora	.364	.652	-.033
evalua	-.824	.473	.133
promueve	.643	-.717	-.006
incentivo	-.769	.460	.167
apoyo	.599	.720	.086
personalok	.586	.699	.091
salario	-.776	.555	.004
objetivo	-.667	-.631	-.012
calidad	-.049	-.260	.915
oportunidad	.777	-.577	-.015

Extraction Method: Principal Component Analysis.

Esta segunda tabla del listado es la matriz factorial L, que contiene las cargas estimadas para cada variable en los tres factores extraídos; cada columna de esta matriz es un componente principal que, conceptualmente, se vincula con un factor latente potencial.

Una vez definido el número de factores, es importante analizar las comunalidades ya que éstas indican hasta dónde las variables

observables comparten una porción importante de variabilidad entre sí. Las comunalidades están dadas en la siguiente tabla del listado y son iguales a la suma de los cuadrados de las cargas sobre todos los factores extraídos. Así para la primera variable X1 = trabajo desafiante, su comunalidad es igual a: (-.073)2 + (-.145) 2 + (-.941) 2 = 0.911. La solución en k =3 factores resulta en un conjunto de comunalidades mayores a 0.8 las cuales indican que las variables observadas comparten entre ellas una gran porción de su varianza.

	Communalities	
	Initial	Exctraction
desafia	1.000	.911
supfacil	1.000	.882
colabora	1.000	.828
evalua	1.000	.921
promueve	1.000	.927
incentivo	1.000	.832
apoyo	1.000	.884
personalok	1.000	.840
salario	1.000	.910
objetivo	1.000	.843
calidad	1.000	.907
oportunidad	1.000	.937

Extraction Method: Principal Component Analysis.

En caso de que uno de los indicadores manifiestos tenga baja comunalidad el analista puede optar por eliminarlo bajo la consideración de que el indicador no covaría con el resto de las variables observables porque no está relacionada con el dominio del concepto de interés, es decir, no se relaciona con los mismos factores comunes para el resto de los indicadores (Fabrigar et al., 1999). La pregunta que surge es ¿cuál es un valor aceptable de comunalidad? Puesto que cargas por arriba de 0.3 se consideran importantes (ver sección 2.4) y la suma de cuadrados de ellas es

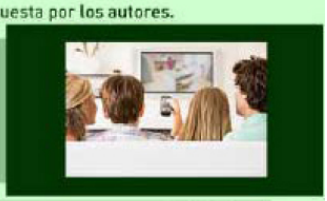
igual a la comunalidad, esto puede ayudar a establecer una cota mínima aceptable. En el ejemplo anterior 0.32 x 3 (factores) = 0.27 por lo que indicadores con comunalidades menores de 0.27 serían candidatos para eliminación.

Actividad de repaso

2.3 Criterios para decidir el número de factores a extraer

Revisa el caso. Responde a las preguntas. Da clic en Respuesta para revisar la solución propuesta por los autores.

En un estudio realizado para determinar las percepciones sobre la programación de la televisión comercial, se realizó una encuesta entre 100 televidentes elegidos al azar de un panel de medios integrado por familias toluqueñas. La encuesta consistió de cinco reactivos todos en una escala Likert de cinco categorías donde 1 = total acuerdo a 5 = total desacuerdo con la declaración. Se aplicó un análisis factorial a los datos usando como método de extracción componentes principales. Resultados:



Variables	Factor 1	Factor 2
1. La T.V. no promueve valores significativos	.594	.216
2. La T.V. es la forma de entretenimiento más accesible	.409	.741
3. La T.V. fomenta la creación de estereotipos	.580	.337
4. La T.V. hace buena difusión de eventos relevantes	.547	.425
5. La T.V. contribuye a decrecer el interés en otras actividades recreativas	.668	.408
Eigen-valor	1.60195	1.0564

a) Los Eigen-valores asociados a los restantes factores son: .911, .745, .686. Construir un scree-plot ¿cuántos factores se sugiere extraer?

Ver scree-plot Ver respuesta

b) Calcular el porcentaje de varianza explicada por la solución en dos factores.

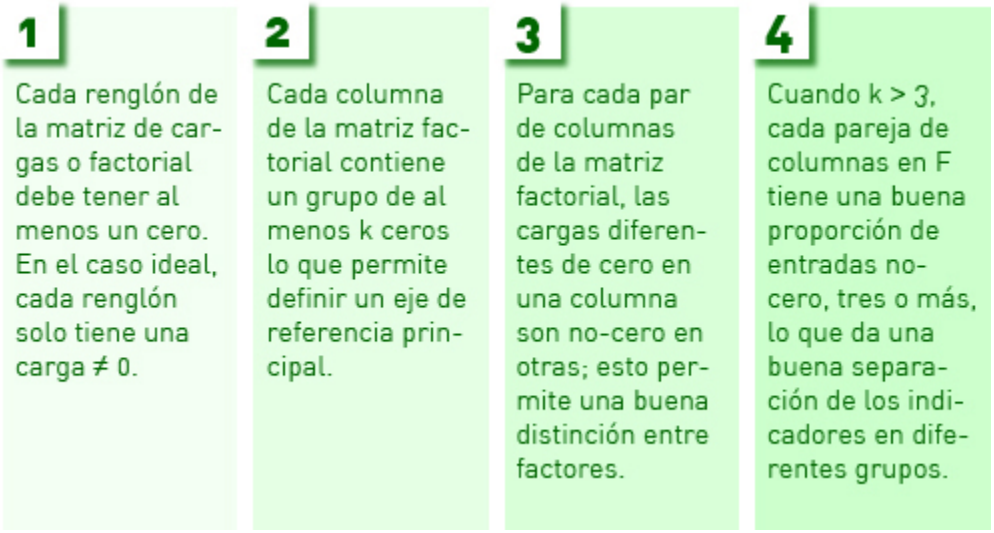
Ver respuesta

c) Calcular las comunalidades, ¿alguno de los reactivos debería ser considerado para su eliminación en la multi-escala?

Ver tabla Ver respuesta

2.4 Rotación de factores e interpretación de la solución

La siguiente parte del análisis de factores es la rotación de la solución; el objetivo de este paso es tener una solución más simple que ayude a comprender la estructura de los datos sin que se modifiquen ni el número de factores ni las comunalidades de las variables. Ya sea que se utilicen componentes principales o máxima verosimilitud, los factores que se ex traen difícilmente permiten identificar cuáles son los indicadores manifiestos que tienen asociados ya que usualmente todas las variables tienen cargas altas en el primer factor y cargas bipolares (negativas y positivas) en los restantes. Es importante recordar que la identificación de las dimensiones subyacentes a un concepto se hace en función de sus indicadores mani fiestos, por lo que es necesario determinar grupos de indicadores a través de los cuales inferir la presencia de los factores latentes. Para generar una Cada renglón de Cada columna la matriz de car-de la matriz facgas o factorial torial contiene debe tener al un grupo de al menos un cero. menos k ceros En el caso ideal, lo que permite cada renglón definir un eje de solo tiene una referencia princarga ≠ 0. cipal. estructura más simple que facilite la interpretación, Thurstone, citado por de Vicente y Oliva y Manera Bassa (2003), propuso los siguientes criterios:



Hay dos tipos de rotaciones que se pueden utilizar: ortogonales y oblicuas. En una rotación ortogonal los factores, que son los ejes de referencia para las variables observadas, se giran de tal modo que los ángulos entre ellos sean siempre ángulos rectos o de 90°. Esta rotación cambia las car-gas de las variables en los factores pero la solución es equivalente a la norotada en el sentido de que se explica la misma cantidad de varianza. Las rotaciones ortogonales producen factores que están no-correlacionados entre ellos lo cual es relevante cuando se desea demostrar la validez de una multi-escala y cuando después del análisis factorial se utilizan métodos de dependencia en los cuales los llamados puntajes factoriales, ver siguiente sección, se utilizan como datos de entrada. La Figura 2.5 describe gráficamente en qué consiste el procedimiento de rotación.

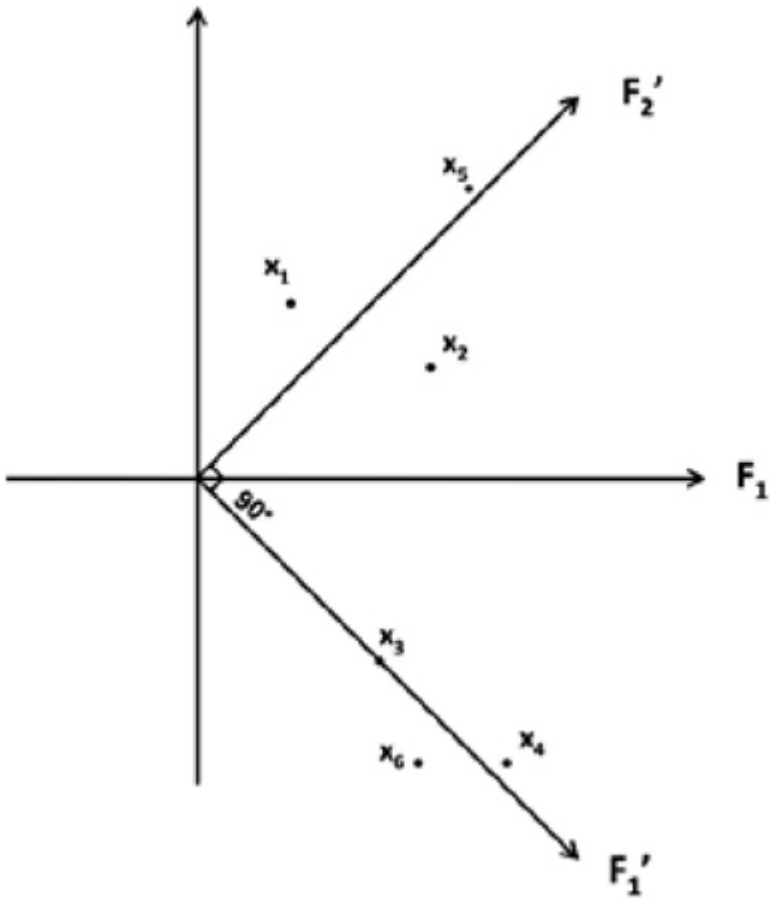


Figura 2.5 Rotación de los ejes coordenados del espacio de los factores

Quartimax

Propone maximizar la variabilidad entre las cargas para cada renglón de la matriz factorial, es decir, se busca que por cada variable se tengan cargas distintas de cero en un único factor.

Varimax

Se maximiza la variabilidad entre cargas por columna, lo que en general resulta en una mejor separación entre los factores. Este tipo de rotación es la más popular porque se obtienen soluciones simples en las que cada factor solo tiene un pequeño número de variables con cargas distintas de cero.

Equimax

Compromiso entre los dos criterios anteriores, esto es max(cargas por renglón vs. cargas por columna). Este tipo de rotación es poco utilizada.

Orthomax

En este caso el usuario define el tipo de compromiso entre maximizar variabilidad por renglón o por columna. En MINITAB, $\gamma \in (0;1)$ es el parámetro que permite decidir el sentido en el cual se desea maximizar la variabilidad entre cargas. Se tienen los siguientes casos: 1) $\gamma = 1$ equivale a rotación Varimax; 2) $\gamma = 0$ equivale a rotación Quartimax; $\gamma = k/2$ corresponde a rotación Equimax. Cuando γ es fijada por el usuario en valores distintos a los anteriores se tiene rotación Orthomax. Para obtener más detalles consultar Abdi, 2003.

Se pueden aplicar distintas formas de rotación ortogonal las cuales se explican a continuación:

Cuando se utiliza rotación oblicua los nuevos ejes pueden asumir cualquier posición en el espacio de los factores, es decir que el ángulo entre los ejes ya no tiene que ser un ángulo recto; sin embargo se buscan ángulos cercanos a 90° para que el grado de correlación entre los factores no sea muy grande ya que una alta correlación entre factores complica la interpretación del modelo de medida. Según Costello y Osborne (2005, p.3) no hay un método de rotación oblicua predominante por lo que se recomienda utilizar las opciones disponibles y valores de default del *software* comercial. Si

el analista cuenta con el paquete estadístico MINITAB no podrá realizar rotaciones oblicuas, pero sí con otros paquetes estadísticos como SPSS. La rotación oblicua en SPSS se puede hacer a través de la opción Oblimin; el parámetro δ define que tan oblicua es la rotación.

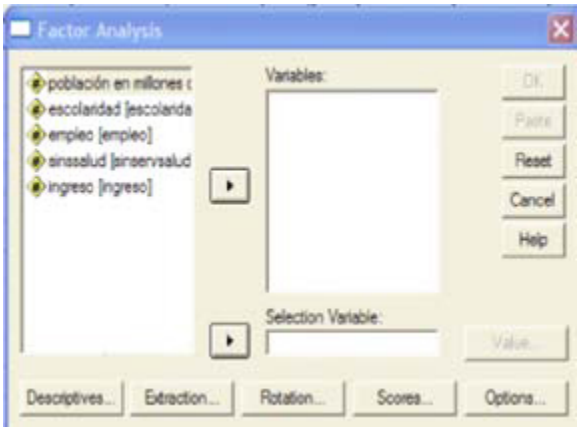


Figura 2.6 Ventana de diálogo para realizar rotación de la solución factorial en SPSS

Si δ se elije muy cercana a cero la rotación es muy oblicua, mientras que a medida que δ → 0.8 la rotación se hace progresivamente menos oblicua. La Figura 2.6 muestra la ventana de diálogo principal de SPSS y la ventana que abre el botón Rotation para que el analista elija las opciones de rotación.

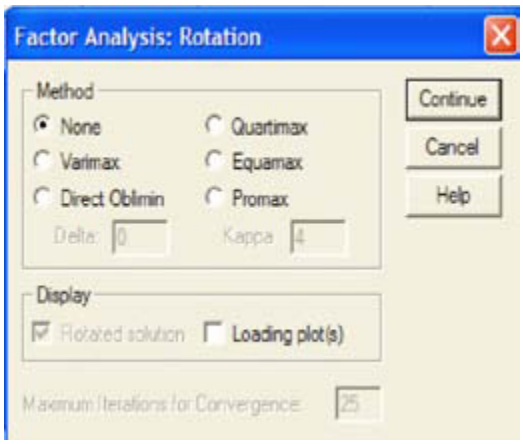


Figura 2.6 Ventana de diálogo para realizar rotación de la solución factorial en SPSS

Con rotación oblicua se generan dos matrices que se usan para realizar la interpretación de la solución: la matriz de estructura y la

matriz de patrones. La matriz de estructura contiene las correlaciones entre cada variable observable y cada eje rotado, es decir que es equivalente a la matriz de cargas rotada generada con una rotación ortogonal. La matriz de patrones, por su parte, tiene como entradas a las llamadas “cargas beta” que son equivalen tes a coeficientes estandarizados entre cada indicador y un factor latente, y reproducen los valores de las variables a partir de los puntajes factoriales. Hay cierta controversia respecto a cuál de las dos matrices utilizar para interpretación, la de estructura es relevante puesto que enfatiza el objetivo de asociar indicadores manifiestos a variables latentes por otro lado, la de patrones ofrece una estructura más simple e inequívoca que facilita la interpretación de la solución.

Lo deseable es que ambas matrices concuerden y permitan una clara identificación de los factores.

Una vez que se ha obtenido la **matriz de factores rotada**, se procede a interpretar la solución. Para completar esta interpretación se utilizan las cargas que constituyen las entradas de la matriz: si una variable observable tiene una carga fuerte en un factor, se considera indicador manifiesto para esa variable latente. Stevens (1986) ofrece una buena discusión de cómo analizar las cargas, la cual se resume enseguida:

1. Las cargas l_{ij} son equivalentes a coeficientes de correlación entre el indicador manifiesto X_i y el factor latente F_j, por lo que su significancia se puede probar si se utiliza la prueba disponible para el coeficiente de correlación rij. Por ejemplo para $n = 100$, si $l_{ij} > .19$ se rechaza la hipótesis de no asociación con un nivel de significancia $\alpha = .05$. Cuando $l_{ij} > .26$ la hipótesis de no asociación se rechaza con un nivel de significancia $\alpha = .01$.

2. El criterio para significancia se aumenta gradualmente al ir de F_1 a F_2 a F_3 ... F_k. Es decir un mayor valor de la carga es requerido para declararla significante en el factor 3 respecto al factor 1.

3. El valor requerido para declarar una carga significante es menor a medida que aumenta el tamaño de muestra (n) y/o el número de variables observables (p).

Como regla empírica también se ha sugerido considerar una carga significante si ésta es mayor de 0.3 y muy importante cuando es superior a 0.5. Además de utilizar las cargas es muy importante que el analista se apoye en la literatura apropiada para vincular la solución propuesta con sus referentes teóricos. El análisis factorial exploratorio se ha criticado debido a que la interpretación de la solución es hasta cierto punto subjetiva en el sentido de que el analista identifica los factores en términos de su experiencia, habilidad y conocimientos sobre el tema que estudia sin que la calidad de su interpretación se pueda juzgar cuantitativamente (Stewart, 1981). Ciertamente la interpretación de la solución factorial está guiada por los propósitos del proyecto de inteligencia de mercados y fuertemente sustentada en las habilidades del analista, sin embargo, la buena comprensión de las etapas de la técnica y de sus limitaciones asegura una buena identificación de factores y apoya en la comprensión de conceptos complejos y el desarrollo de medidas válidas.

Ejemplo 3. Considere la encuesta aplicada a los egresados de las universidades ubicadas en la zona metropolitana de la ciudad de Toluca y cuyo propósito fue identificar las características de un trabajo ideal. Después de aplicar rotación Varimax, la matriz factorial rotada tiene la siguiente estructura, que permite identificar fácilmente los indicadores asociados a cada factor.

Stat > Multivariate > Factor Analysis

El listado que se exhibe fue obtenido con MINITAB a través de la secuen cia de comandos:

Rotated Factor Loadings and Communalities

Varimax Rotation

Variable	Factor1	Factor2	Factor3	Communality
desafiante/competencias	-0.049	0.043	-0.953	0.911
superiores facilitan	-0.181	0.915	-0.113	0.882
ambiente colaboración	-0.070	0.898	0.127	0.828
evaluación específica	0.944	-0.163	-0.052	0.921
promociones abiertas	-0.948	-0.124	-0.110	0.927
ofertas de incentivos	0.898	-0.133	-0.088	0.832
personal apoyo confiable	0.013	0.940	0.021	0.884
personal adecuado	0.010	0.916	0.013	0.840
salarios competitivos	0.946	-0.084	0.088	0.910
objetivos/responsabilidades	0.104	-0.908	-0.080	0.843
demandas y calidad vida	-0.032	-0.126	-0.943	0.907
oportunidad desarrollo	-0.962	0.067	-0.080	0.937
Variance	4.4712	4.2810	1.8696	10.6217
% Var	0.373	0.357	0.156	0.885

En la matriz anterior se han marcado las cargas altamente significantes o muy importantes, a partir de los grupos de indicadores se identifican los tres factores:

F1: Retribución y desarrollo, se refiere a aquellos atributos del trabajo relacionados con la evaluación del desempeño, las remuneraciones y las oportunidades de crecimiento laboral.

F2: Ambiente colaborativo, incluye atributos que tienen que ver con la disponibilidad de un recurso humano que facilita y apoya las tareas laborareportan a continuación.

F3: Exigencia, dimensión latente inferida a través de las exigencias del trabajo y cómo estas afectan la calidad de vida del trabajador.

Ejemplo 4. Un grupo de médicos y administradores formó una sociedad civil que tiene el interés de abrir una clínica de especialidades en una ciudad de la zona centro de México. Entre las varias decisiones a tomar está la de elegir la ciudad en dónde ubicar la clínica. La ciudad elegida debe tener un alto potencial de mercado, es decir, debe contar con una cantidad suficiente de individuos que tengan necesidad de servicios médicos de calidad, que reconozcan la importancia de contar con tales servicios y que puedan pagarlos. Para apoyar la decisión de ubicación de la clínica, uno de los licenciados en mercadotecnia del grupo consultó las publicaciones del Instituto Nacional de Estadística Geografía e Informática (INEGI, 2011) y obtuvo datos para 14 ciudades candidatas, todas ubicadas en la zona central de México. Los datos recolectados incluyen información para los siguientes indicadores:

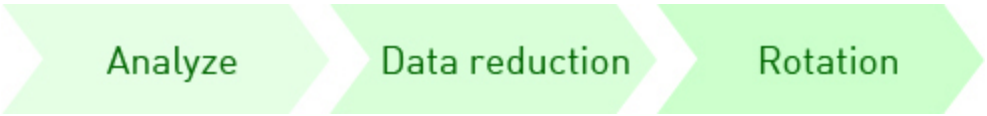
$X1$ = población total de la ciudad (en miles de habitantes)

$X2$ = Promedio de escolaridad de la población (en años)

*X*3 = Tasa de desempleo (en porcentaje)

*X*4 = Porcentaje de la población que no cuenta con servicios de salud

*X*5 = Ingreso promedio en la zona (en múltiplos de salarios mínimos)



En este ejemplo los datos fueron procesados con SPSS usando la secuencia de comandos Dentro de la opción de rotación, la selección del procedimiento Varimax generó la siguiente matriz factorial:

En la matriz anterior, las cargas del indicador X4 = porcentaje de individuos sin servicios de salud en los dos factores son significantes; para simplificar la solución se aplicó una rotación oblicua con δ = 0.7. Las matrices de patrones, estructura y asociación entre los dos factores extraídos se reportan a continuación.

	Componente	
	1	2
Población en millones de habitantes	.279	.920
Nivel de escolaridad	-.230	.650
Tasa de desempleo	.882	.055
% de la población sin servicios de salud	.424	.746
Ingreso promedio	-.931	-.123

Extraction Method: Principal Component Analysis
Rotation Method: Varimax with Kaiser Normalization.
A Rotation converged in 3 iterations.

	Componente	
	1	2
Población en millones de habitantes	.183	.906
Nivel de escolaridad	-.305	.686
Tasa de desempleo	.892	-.040
% de la población sin servicios de salud	.350	.713
Ingreso promedio	-.933	-.024

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.
Rotation converged in 5 iterations.

Matriz de patrones

	Componente	
	1	2
Población en millones de habitantes	.375	.945
Nivel de escolaridad	-.160	.621
Tasa de desempleo	.883	.149
% de la población sin servicios de salud	.501	.787
Ingreso promedio	-.933	-.222

Componente		
	1	2
1	1.000	.212
2	.212	1.000

Matriz de estructura
Component Correlation Matrix

F1: potencial demográfico esto es la disponibilidad de una gran cantidad de individuos que podrían requerir los servicios hospitalarios porque reconocen que los necesitan, están bien educados por lo que buscan cuidar su salud, y no cuentan con los servicios actualmente, un alto porcentaje de la población no tiene seguros de salud.

F2: potencial económico, que implica que las personas cuentan con los recursos económicos, un empleo formal y buen ingreso, para pagar por los servicios.

Analizando la matriz de patrones se aprecia un diferencial mayor entre las cargas del indicador X3 con rotación oblicua respecto a rotación ortogonal. Los siguientes grupos de variables pueden formarse (X1, X2 y X4) y (X3, X5), a partir de estos grupos se identifican a los factores como sigue:

El número de datos usados en este ejemplo es solo de 14, es conveniente introducir una nota de precaución respecto a un tamaño de muestra así de pequeño. El uso de reglas empíricas para definir el tamaño de muestra, n >100 o cinco datos por cada variable observada, no se aplica con mucha frecuencia porque se reconoce que el tamaño de muestra está determinado por la naturaleza de los datos. Si las comunalidades son altas, el número de factores no se ha sobre determinado, los indicadores no tienen cargas grandes en múltiples factores y la solución se mantiene para otros conjuntos de datos realmente no es necesario aumentar el tamaño de muestra (Fabrigar et al. 1999). Sin embargo Costello y Osborne (2005) verificaron que cuando la razón número de datos a número de variables observables es menor de 10, el número de indicadores mal asignados a los factores extraídos y los errores de estimación en cargas y eigen-valores se incrementan significativamente. Por tanto una recomendación conservadora sería tener n > 5p.

¿Sabías qué?

La cantidad de trabajos de investigación en los cuales se aplicó el análisis factorial mostró un crecimiento casi exponencial durante el período de 1960 a 2005.

2.5 Cálculo y utilización de puntajes factoriales

En la primera sección del capítulo se mencionó que entre los objetivos del análisis factorial está reducir o condensar la información sobre un gran número de variables. Esto se logra al calcular agregados de las variables observadas, conocidos como puntajes factoriales, que corresponden a estimados de los factores no observables. Estos puntajes o scores se pueden usar como datos de entrada para otros métodos y cuando son ortogonales resultan ser muy útiles para evitar problemas de multi-colinealidad o asociación intensa entre variables. Los puntajes factoriales también son útiles para identificar datos extremos o para comparar objetos entre sí ya que al representar la información original en el espacio reducido de los factores es más fácil detectar datos inusuales o apreciar qué tan distantes están dos objetos entre sí.

Para calcular los puntajes factoriales, se considera de nuevo el modelo de análisis factorial exploratorio $\boldsymbol{x - \mu = LF + \varepsilon}$. Inicialmente, solo los valores del vector *x* están disponibles, pero después del análisis, se cuenta con la matriz que contiene los estimados de las cargas, esto es ***L***. Si las entradas de esta matriz se tratan como valores verdaderos y a los componentes específicos ε como errores aleatorios se puede estimar a ***F***, es decir a los factores latentes (Johnson y Wichern, 1998). Para realizar esta estimación se pueden aplicar los siguientes métodos:

Los puntajes factoriales calculados por cualquiera de los procedimientos anteriores corresponden a una combinación lineal de las variables observadas según se muestra en la expresión (9).

$$\hat{f}_j = \sum_{i=1}^{p} w_i X_{ij} \qquad (9)$$

Los términos w_j se conocen como los coeficientes para el puntaje o score factorial y representan la ponderación o peso que cada variable tiene en el puntaje. Estos coeficientes son diferentes a las cargas pero si una variable tiene una carga alta en un factor también tendrá un coeficiente relativamente grande en el puntaje factorial correspondiente. Los coeficientes wj son reportados en la última sección del listado regular para un análisis factorial en MINITAB y puede requerirse su impresión en el caso de SPSS según se describe en el siguiente ejemplo:

Ejemplo 5. Considere el problema del ejemplo 4 en el cual se desea seleccionar la ciudad con mayor potencial de mercado en la zona centro para ubicar ahí una clínica de especialidades médicas.

Para visualizar mejor el potencial de mercado de las ciudades candidatas la información de cada una se resumió en dos puntajes factoriales: supotencial demográfico y su potencial económico. Los puntajes fueron calculados eligiendo la opción Scores en la ventana de diálogo principal (ver Figura 2.5); el método usado fue el de regresión que es la opción de default disponible en SPSS. Para guardar los puntajes en la hoja de datos y desplegar los valores de los coeficientes basta con elegir las opciones: “Save as variables” y “Display factor score coefficients matrix”. La matriz de coeficientes es la siguiente:

	Factor	
	1	2
Población en millones de habitantes	.078	.493
Nivel de escolaridad	-.170	.383
Tasa de desempleo	.460	-.042
% de la población sin servicios de salud	.167	.384
Ingreso promedio	-.481	.007

Los puntajes factoriales se calculan multiplicando los coeficientes por los valores de los indicadores manifiestos estandarizados por ciudad $Z_i = \frac{(X_i - \bar{X})}{s_i}$ por ejemplo el primer puntaje factorial (f_1) que representa el potencial económico se calcula como sigue:

$$f_1 = .078\ Z_1 - .17\ Z_2 + .46\ Z_3 + .167\ Z_4 - .481\ Z_5$$

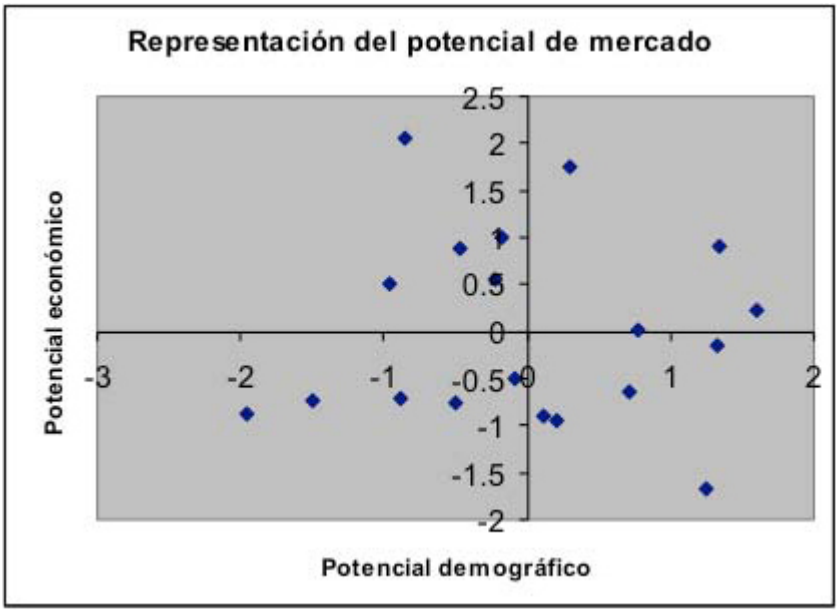
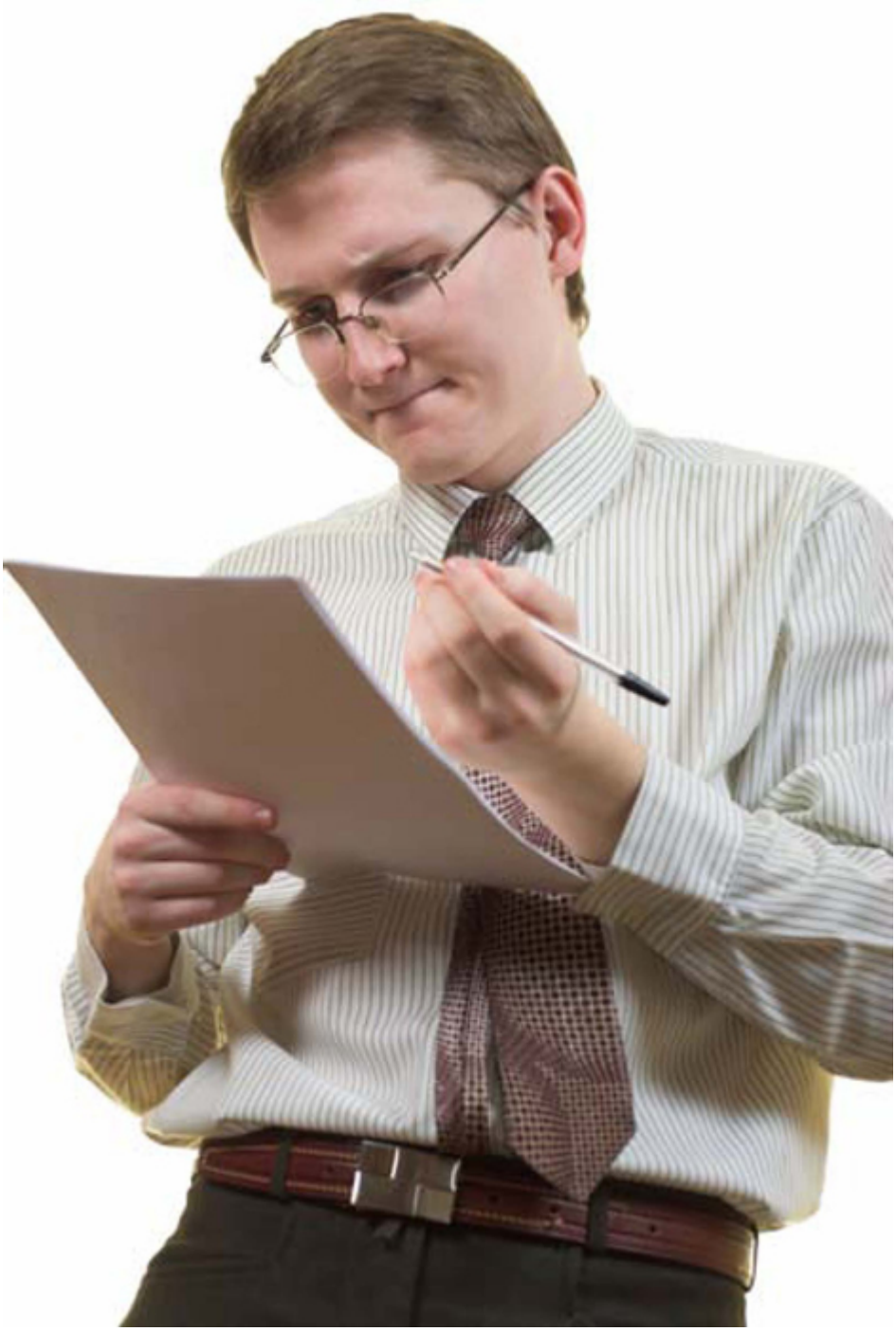


Figura 2.7 Representación del potencial de mercado de ciudades en el espacio factorial

Estos puntajes se utilizaron para construir el gráfico de la Figura 2.7. En el diagrama los tres puntos ubicados en el cuadrante noreste, derecho superior, representan a aquellas ciudades con el mayor potencial demográfico y económico y que por tanto serían las mejores opciones para localizar la nueva clínica de salud. Una de ellas destaca por su gran potencial económico (1.73) y otra por ser la de mayor potencial demográfico (1.62).



Actividad de repaso (1era parte)

2.5 Cálculo y utilización de puntuajes factoriales

Calificar cada declaración como falsa o verdadera:

En una rotación oblicua se relaja la condición de que los factores sean ortogonales.

Los puntajes factoriales son aquellas combinaciones lineales de las variables observables con máxima varianza.

Las cargas son equivalentes a los coeficientes de correlación entre variables observables y los factores latentes.

El objetivo de la rotación es reducir el número de indicadores a utilizar para representar a un individuo.

Los puntajes factoriales permiten representar los datos en el espacio reducido de la solución en k factores.

Una carga se considera significante si es mayor de cero.

La matriz de patrones que se genera en una rotación oblicua es equivalente a la matriz de cargas rotada obtenida con Varimax.

El método de mínimos cuadrados ponderados es una opción para el cálculo de los puntajes factoriales.

Actividad de repaso (2da parte)

2.5 Cálculo y utilización de puntuajes factoriales

Instrucciones: Revisa el ejemplo y contesta las siguientes preguntas. Dar clic en "Respuesta" para conocer la solución propuesta por los autores.

El área de recursos humanos de una multinacional realizó un proyecto cuyo objetivo fue definir aquellas competencias que debían tener los ejecutivos de mercadotecnia de empresas globales. Como parte del proyecto se aplicó una encuesta a 90 ejecutivos con experiencia. Se utilizó una multi-escala de 4 ítems, cada ítem está en una escala Likert donde 1 = total acuerdo a 5 = total desacuerdo en la declaración. Después de aplicar un análisis factorial a los datos disponibles, se obtuvo la siguiente matriz de cargas rotada.

Declaración	Solución varimax	
	factor 1	factor 2
Tengo buenas capacidades para comunicarme con colaboradores extranjeros	.716	.194
Tengo sensibilidad para comprender el ambiente de negocios en otros países	.841	-.209
Poseo conocimientos de alta calidad sobre mi área	.137	.680
Soy capaz de buscar nueva información para actualizar mis conocimientos	.025	.747
Eigenvalor	1.25	1.10

a) Con base en las cargas agrupar a los indicadores manifiestos.

Ver respuesta

b) Interpretar la solución obtenida

Ver respuesta

2.6 Aplicación del análisis de factores con apoyo computacional (MINITAB)

Un análisis de factores incluye varias etapas, en cada una de las cuales hay que tomar decisiones que incluyen ¿cómo estimar los factores? ¿cuántos factores extraer? ¿qué tipo de rotación usar?, entre otros. En esta sección se desarrollan todas las etapas a partir de un ejemplo de aplicación en el contexto de inteligencia de mercados, el análisis se lleva a cabo empleando MINITAB.

La satisfacción del cliente es uno de los conceptos críticos en mercadotecnia ya que se considera como variable mediadora crítica para el desarrollo de la lealtad. Uno de los estudios más interesantes en relación a este concepto es el desarrollo del índice de satisfacción americano para evaluar los servicios de empresas en varios sectores.

Este índice se ha validado en el contexto estadounidense y se considera el estándar para reportar los niveles de satisfacción del consumidor americano. Sin embargo, su aplicabilidad en otros contextos no está garantizada y en particular su uso dentro del sector de restaurantes es limitado. La CANIRAC (La Cámara

Nacional de la Industria de Restaurantes y Alimentos Condimentados) ha recibido la propuesta de manejar un índice de satisfacción para restaurantes y es necesario explorar cuáles son las dimensiones de satisfacción que incluye la multi-escala de diez reactivos que se ha porpuesto. Todos los reactivos están en una escala tipo Likert con siete categorías (1= en desacuerdo, 7 = en acuerdo). La descripción de los reactivos es la siguiente:

*X*1 = Este restaurante cuenta con personal que parece bien entrenado y competente

*X*2 = Los empleados en este restaurante hacen que usted se sienta confortable y seguro en su trato con ellos

*X*3 = Su aspecto exterior y el área de comedor son visualmente atractivas

*X*4 = El restaurante luce descuidado o con poco mantenimiento

*X*5 = El restaurante está decorado de acuerdo a su imagen y escala de precios

*X*6 = El tipo de clientes no es el que usted esperaría encontrar en un lugar de este estilo

*X*7 = Parece que da apoyo a los empleados para que puedan hacer biensu trabajo

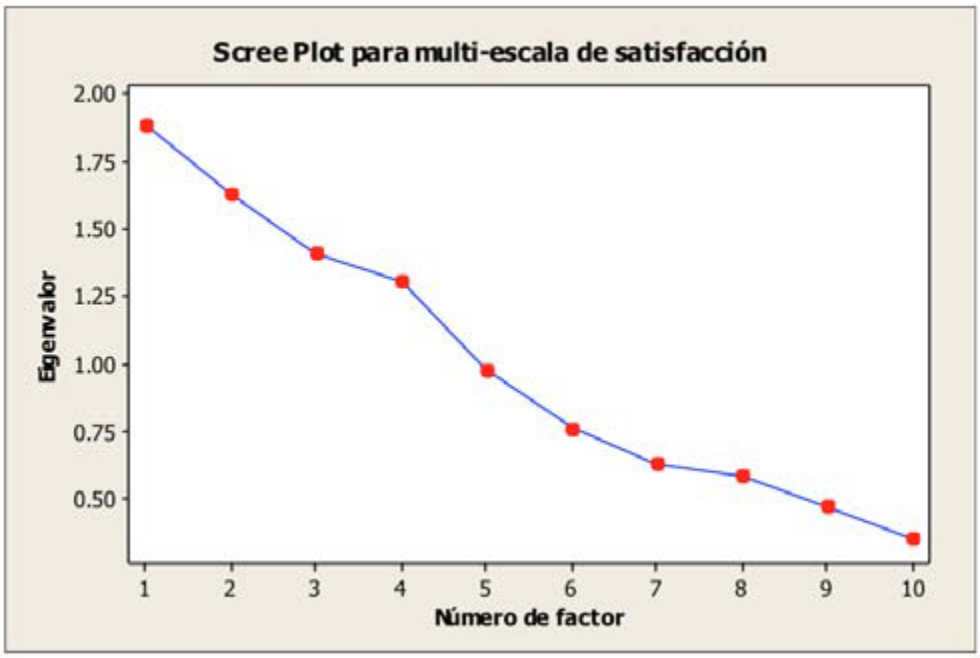
*X*8 = Los platillos tienen una presentación atractiva

*X*9 = La carta ofrece una variedad limitada de platillos

*X*10 = Los empleados entienden sus necesidades y tratan de atenderlas de inmediato

a) En cuanto al método de extracción para los factores, la elección en este ejemplo es Componentes Principales ya que la escala Likert en que están medidos los indicadores no es una escala para la cual sea razonable asumir una distribución normal puesto que la escala

es discreta y sus categorías, si bien representan distintos niveles de intensidad, no son estrictamente comparables.



b) Para determinar el número de factores, se procede a construir el scree-plot, que se muestra a continuación, en el que el punto de quiebre no es tan evidente como en otros ejemplos. Pero sí se puede apreciar un cambio importante en k = 4.

Con esta propuesta de k = 4 factores, se calcula la matriz factorial o de cargas no-rotada para analizar eigen-valores y porcentaje de varianza explicada y confirmar o proponer soluciones alternas en cuanto a número de factores. La matriz de cargas obtenida con componentes principales es la siguiente:

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Factor3	Factor4	Communality
X1	0.539	-0.164	-0.316	-0.372	0.556
X2	0.468	-0.417	-0.075	-0.158	0.423
X3	0.415	-0.249	0.671	-0.059	0.688
X4	-0.353	-0.057	-0.739	0.210	0.718
X5	0.183	0.378	0.016	0.605	0.542
X6	-0.203	0.182	-0.224	-0.746	0.681
X7	0.600	-0.428	-0.206	0.145	0.606
X8	0.381	0.795	0.099	-0.001	0.787
X9	-0.536	-0.607	0.203	0.222	0.747
X10	0.461	-0.081	-0.407	0.312	0.483
Variance	1.8868	1.6313	1.4114	1.3016	6.2311
% Var	0.189	0.163	0.141	0.130	0.623

c) En la matriz anterior no hay una distribución de cargas que permita la identificación; en consecuencia, es necesario aplicar una rotación. Se decide aplicar rotación Varimax debido a sus cualidades para simplificar la estructura de la matriz factorial, solo si esta rotación es insatisfactoria se consideran otras opciones. La matriz factorial rotada se da a continuación.

Rotated Factor Loadings and Communalities: Varimax Rotation

Variable	Factor1	Factor2	Factor3	Factor4	Communality
X1	0.648	0.214	0.006	-0.301	0.556
X2	0.607	-0.096	0.190	-0.096	0.423
X3	0.176	-0.074	0.801	0.096	0.688
X4	0.041	-0.172	-0.828	0.033	0.718
X5	-0.100	0.317	-0.081	0.652	0.542
X6	-0.101	0.195	-0.161	-0.779	0.681
X7	0.744	-0.073	0.066	0.207	0.606
X8	-0.146	0.853	0.127	0.147	0.787
X9	-0.204	-0.835	0.022	0.086	0.747
X10	0.534	0.152	-0.243	0.340	0.483
Variance	1.7429	1.6816	1.4768	1.3299	6.2311
% Var	0.174	0.168	0.148	0.133	0.623

En este punto es importante revisar las comunalidades y decidir si alguno de los indicadores debido a su baja comunalidad y a que tenga cargas altas con múltiples factores dificulta la interpretación. Según Costello y Osborne (2005) la eliminación de reactivos problemáticos - bajas cargas y por tanto comunalidad, cargas altas en múltiples factores o no agrupados con otros reactivos - ayuda a la interpretación más que probar otros tipos de rotación. En este caso X10 tiene el menor valor de comunalidad y cargas significantes en más de un factor. Al eliminar este reactivo y repetir todo el análisis se obtiene la siguiente matriz factorial rotada la cual tiene una estructura más simple que la anterior ya que las cargas están mejor distribuidas (cargas mayores en un único factor). Esta solución reconstruye el 66.7% de la varianza original lo que mejora un poco la solución anterior que fue de 62.3% además, es más interpretable por tanto se decide admitirla.

Rotated Factor Loadings and Communalities: Varimax Rotation

Variable	Factor1	Factor2	Factor3	Factor4	Communality
X1	-0.232	-0.687	0.037	0.249	0.589
X2	0.063	-0.728	-0.072	-0.040	0.541
X3	0.072	-0.185	-0.799	-0.099	0.687
X4	0.168	-0.049	0.851	-0.061	0.759
X5	-0.334	0.033	0.163	-0.744	0.694
X6	-0.194	0.076	0.172	0.770	0.666
X7	0.067	-0.702	-0.093	-0.184	0.539
X8	-0.849	0.168	-0.134	-0.138	0.786
X9	0.838	0.189	-0.019	-0.088	0.746
Variance	1.6682	1.6011	1.4519	1.2843	6.0055
% Var	0.185	0.178	0.161	0.143	0.667

d) La interpretación de la solución se hace en términos de los grupos de variables identificados: (X8 y X9), (X1,X2,X7) (X3,X4) y (X5,X6). Una vez revisado el contenido de los reactivos en cada grupo se hace la siguiente identificación para los factores:

F1: Satisfacción con el producto, este componente de la satisfacción se relaciona directamente con la calidad y variedad de los productos que sirve el restaurante.

¿Sabías qué?

XLSTAT es un software estadístico modular desarrollado a partir de 1993 para incrementar las opciones de análisis estadístico con Excel. El software se apoya en Excel para ofrecer una interfase amigable para la entrada de datos y el desplegado de resultados, pero es un software autónomo de análisis multivariable.

F2: Competencia de los empleados, esta dimensión latente considera la satisfacción derivada de las habilidades del personal que atiende al comensal.

F3: Dimensión tangible de la satisfacción ya que los reactivos asociados a este factor hacen referencia al atractivo y mantenimiento físicos del restaurante

F4: Ambientación, esta dimensión considera si la clientela y decoración del lugar son congruentes con el posicionamiento que tiene el restaurante.

2.7 Análisis factorial confirmatorio

Hay dos grandes tipos de análisis factorial definidos según los objetivos del análisis: **análisis factorial exploratorio** y **análisis factorial confirmatorio**. El primer tipo que es el descrito en este capítulo se aplica cuando hay conocimientos limitados sobre el tema que se estudia por tanto es una técnica particularmente apropiada para desarrollar teoría (Stewart, 1981). En contraste, el análisis factorial confirmatorio formula hipótesis específicas respecto a la estructura de la solución factorial de tal forma que el análisis sea una confirmación empírica respecto a lo apropiado del modelo propuesto. El diagrama devías de la Figura 2.8 ejemplifica un caso de aplicación del análisis factorial confirmatorio; cabe notar que a diferencia del diagrama en la Figura 2.1, el número de factores está fijo (k = 2) y los indicadores asignados a cada factor latente desde un principio.

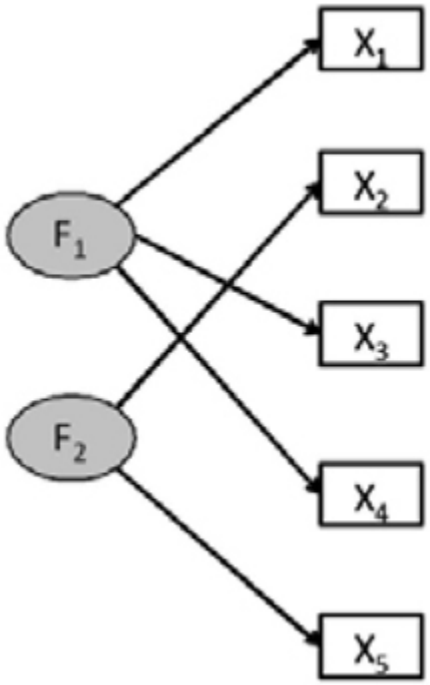
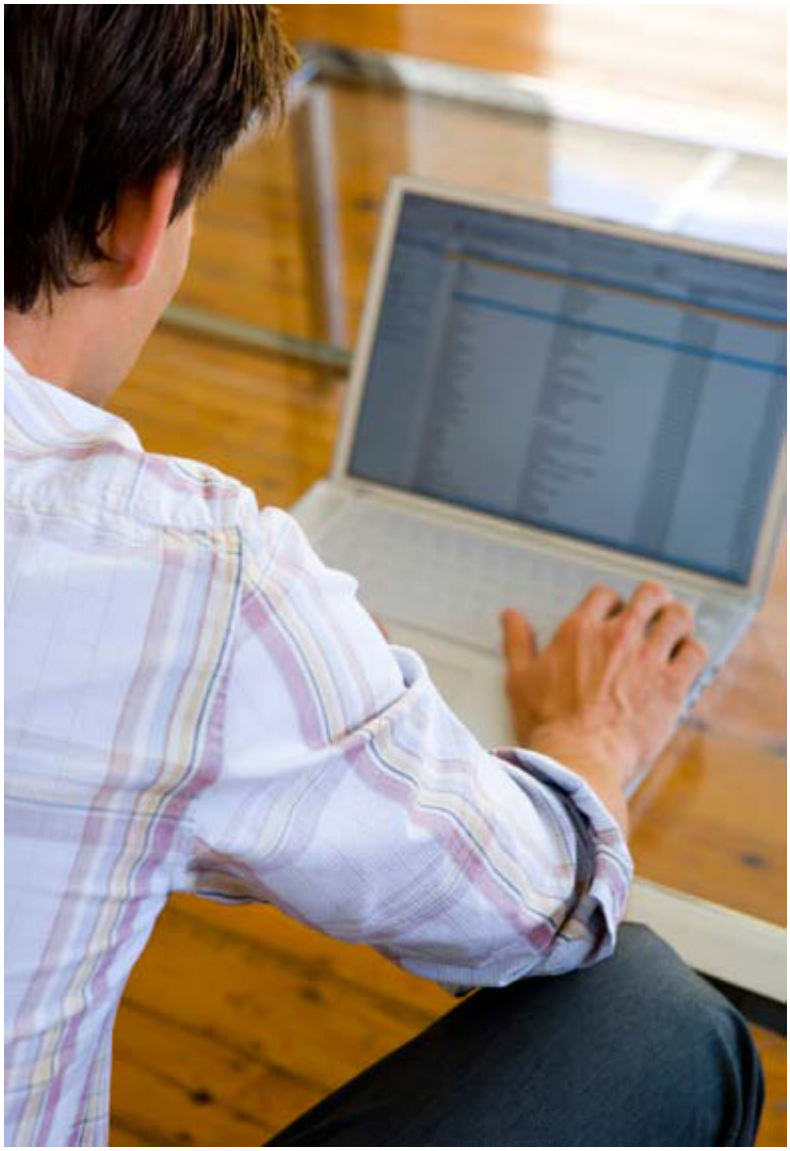
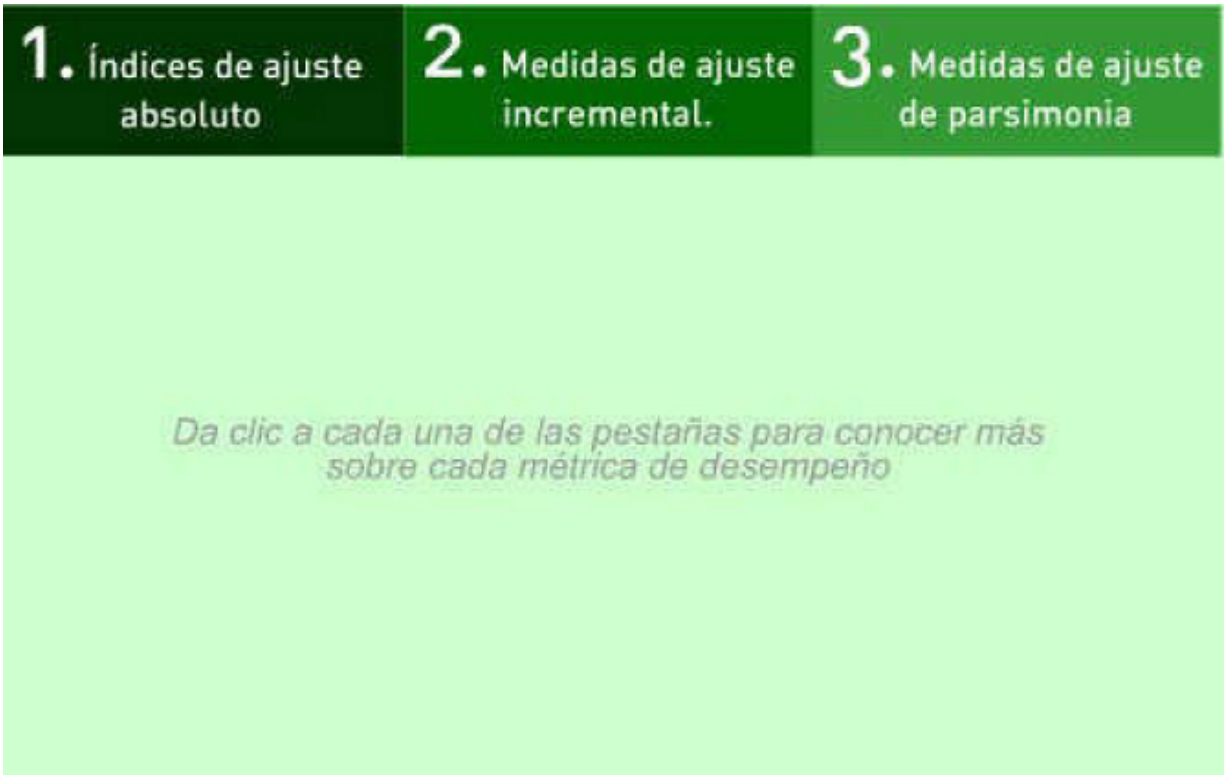


Figura 2.8 Diagrama de vías para el análisis factorial confirmatorio





El modelo de análisis factorial confirmatorio se estima utilizando software como LISREL o AMOS ya que el propósito del análisis es confirmar la estructura del modelo de medida propuesto en relación a: 1) las dimensiones subyacentes al concepto son las especificadas por el analista y 2) cada dimensión latente se infiere a través de indicadores manifiestos determinados. Para evaluar qué tan bien se ajustan los datos al modelo de medida propuesto se utilizan varias métricas de desempeño, los cuales se agrupan en las siguientes categorías (Hair, Anderson, Tatham y Black,1999):

Los valores recomendados para los indicadores GIF, TLI, NFI, CFI y AGIF son mayores o cercanos a 0.9 (Hair et al. 1999; Catena, Ramos y Trujillo, 2003), pero como estos son solo valores recomendados, algunos autores consideran conveniente hacer un análisis basado en la comparación de varios modelos y elegir el modelo que se adapta o ajusta mejor a los datos. Para el caso de RMSR se recomiendan valores pequeños ya que un menor valor de los errores implica un mejor ajuste del modelo (Catena et al., 2003). Respecto a este último índice cabe resaltar que entre mayor sea p = número de variables observadas en relación a k = número de

factores, mayor es el grado de confirmación empírica para el modelo ya que hay unmayor número de restricciones estructurales en la matriz de correlación **R**.

Ejemplo 6. En el contexto de administración del conocimiento uno de los conceptos relevantes es el de capacidades de absorción, que comprende a aquellas capacidades de la empresa para identificar, asimilar y explotar el conocimiento externo. Arroyo y Sánchez (2009) propusieron una multi-escala de 16 reactivos para evaluar las capacidades de absorción de los proveedores estratégicos de la industria automotriz. El propósito final de medir estas capacidades fue identificar aquellos proveedores que podrían aprovechar más los cursos y talleres de capacitación que ofrecen las armadoras para mejorar el desempeño de sus proveedores.

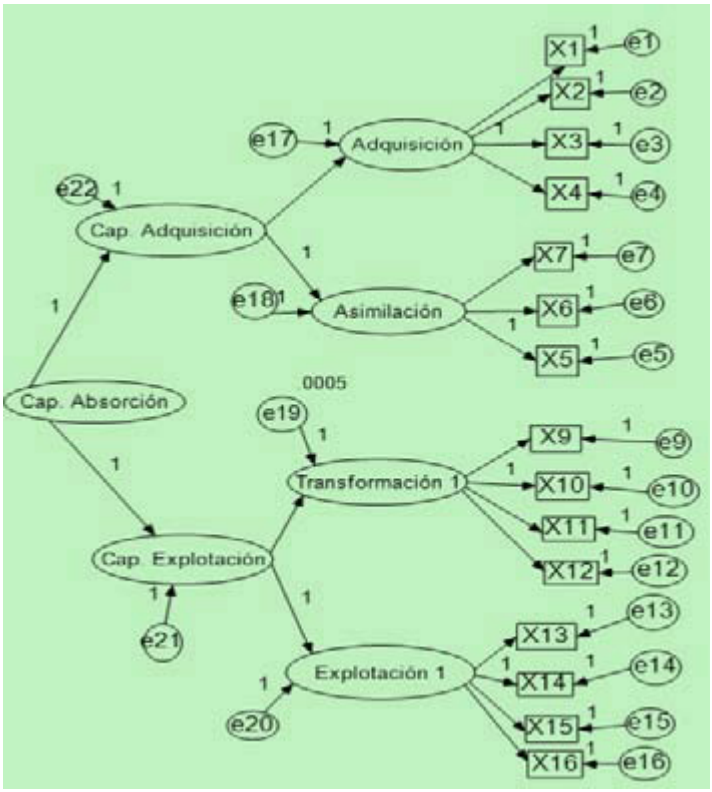


Figura 2.9 Modelo de medida para el concepto capacidades de absorción

La escala fue elaborada tomando como referencia el trabajo de Zahra y George (2002) quienes sugirieron cuatro dimensiones, procesos, para las capacidades de absorción, estas dimensiones se

agrupan en dos categorías: Capacidades de Adquisición y Capacidades de Explotación del conocimiento. El diagrama de vías que describe la estructura de la multi-escala elaborada se muestra en la Figura 2.9.

Significancia de la Chi-cuadrada del modelo	***GIF***	***RMSR***	***TLI***	***NFI***	***CFI***	***AGIF***
Chi-square = 232.452 Df = 101 Probability level = .000	.782	.054	.828	.746	.862	.686

En este caso, el objetivo del análisis factorial realizado con el apoyo del software AMOS fue confirmar la estructura de la multi-escala propuesta. Los **índices de ajuste** para el modelo en la Figura 2.7 son los siguientes

Si bien el estadístico ji-cuadrada es significante (P = 0.000) lo que indica que el modelo no ajusta bien a los datos, se ha observado que este estadístico es sensible tanto

Ejercicio Integrador

2. Análisis de factores

Instrucciones: Revisa y analiza la información sobre este estudio.

Las elecciones presidenciales en México incrementan el interés por definir estrategias de mercadotecnia enfocadas a satisfacer las necesidades políticas de los electores de manera rentable para los partidos. Entre las actividades de la mercadotecnia política está averiguar el posicionamiento que tiene un determinado candidato en la mente de los electores.

Para identificar aquellos componentes críticos que contribuyen al posicionamiento de un personaje público, se realizó un estudio exploratorio en donde se pidió a un grupo de votantes que eligieran a su candidato favorito para la presidencia de entre los doce elegibles reportados en la dirección de Internet http://www.eleccion2012mexico.com/ y procedieran a calificarlo. La evaluación consistió en expresar su nivel de acuerdo o desacuerdo con un conjunto de declaraciones respecto al candidato potencial. Los votantes expresaron sus juicios sobre las declaraciones formuladas sobre una escala de 7 categorías donde 1 = total acuerdo a 7 = total desacuerdo.

El contenido de las declaraciones fue el siguiente:

X1 = Su trayectoria política es la mejor evidencia de su capacidad para gobernar

X2 = Es capaz de organizar un plan de trabajo bien fundamentado

X3 = Expresa claramente sus ideas en público

X4 = Sus respuestas sobre cómo trabajará durante su gestión son concretas

X5 = Sabe rodearse de las personas adecuadas para formular su programa de trabajo

X6 = Es capaz de identificar a las personas que pueden hacerse cargo de los puestos críticos

X7 = Tiene una visión completa de los problemas que atenderá en su puesto

X8 = Es asertivo cuando responde a periodistas y grupos políticos

X9 = Tiene una personalidad atractiva y distintiva Los datos para una muestra de 45 entrevistados se reportan en el Anexo 1.

Responde las siguientes preguntas:

a) Utilice componentes principales para estimar la matriz factorial y aplique los criterios disponibles para proponer un número de factores relevan tes a extraer.

b) Analizar las comunalidades ¿hay algún reactivo cuya varianza compartida sea lo bastante baja como para proponer su eliminación?

c) Utilizar rotación Varimax para identificar las dimensiones subyacentes al “posicionamiento de un candidato a la presidencia”.

d) Representar las opiniones de los encuestados en el espacio bidimensional de los primeros dos factores. ¿Se identifican opiniones extremas? Puesto que cada dato representa la posición percibida del votante hacia algún candidato ¿se identifican grupos de candidatos con similar posicionamiento?

Conclusión capítulo 2

2. Análisis de factores

El objetivo principal del análisis factorial es explicar la estructura de las asociaciones entre indicadores manifiestos a través de la determinación del número y naturaleza de los factores latentes a quienes se atribuyen las correlaciones observadas. A diferencia del ACP, el análisis factorial asume un modelo de medida y logra una reducción en la dimensionalidad de los datos a través del cálculo de puntajes factoriales, por tanto el análisis factorial cubre tanto el objetivo de comprensión de la estructura de las asociaciones entre variables como el de simplificación de los datos.

El análisis factorial es un método de carácter exploratorio que, para su correcto uso e interpretación, requiere de comprender las ventajas, desventajas y resultados que aportan los varios procedimientos cuantitativos que lo integran, entre ellos los métodos de estimación disponibles para la matriz de factores, los procedimientos de rotación y de construcción para los puntajes factoriales. Debido a su naturaleza exploratoria no se apoya en inferencias estadísticas para tomar todas las decisiones relevantes, razón por la cual el método se ha calificado de subjetivo. Para facilitar su aplicación se sugiere la siguiente estrategia, adaptada de Johnson y Wichern (1998):

1) Utilizar componentes principales para estimar la matriz factorial y proponer el número de factores.

2) Utilizar rotación Varimax y proceder a identificar los factores.

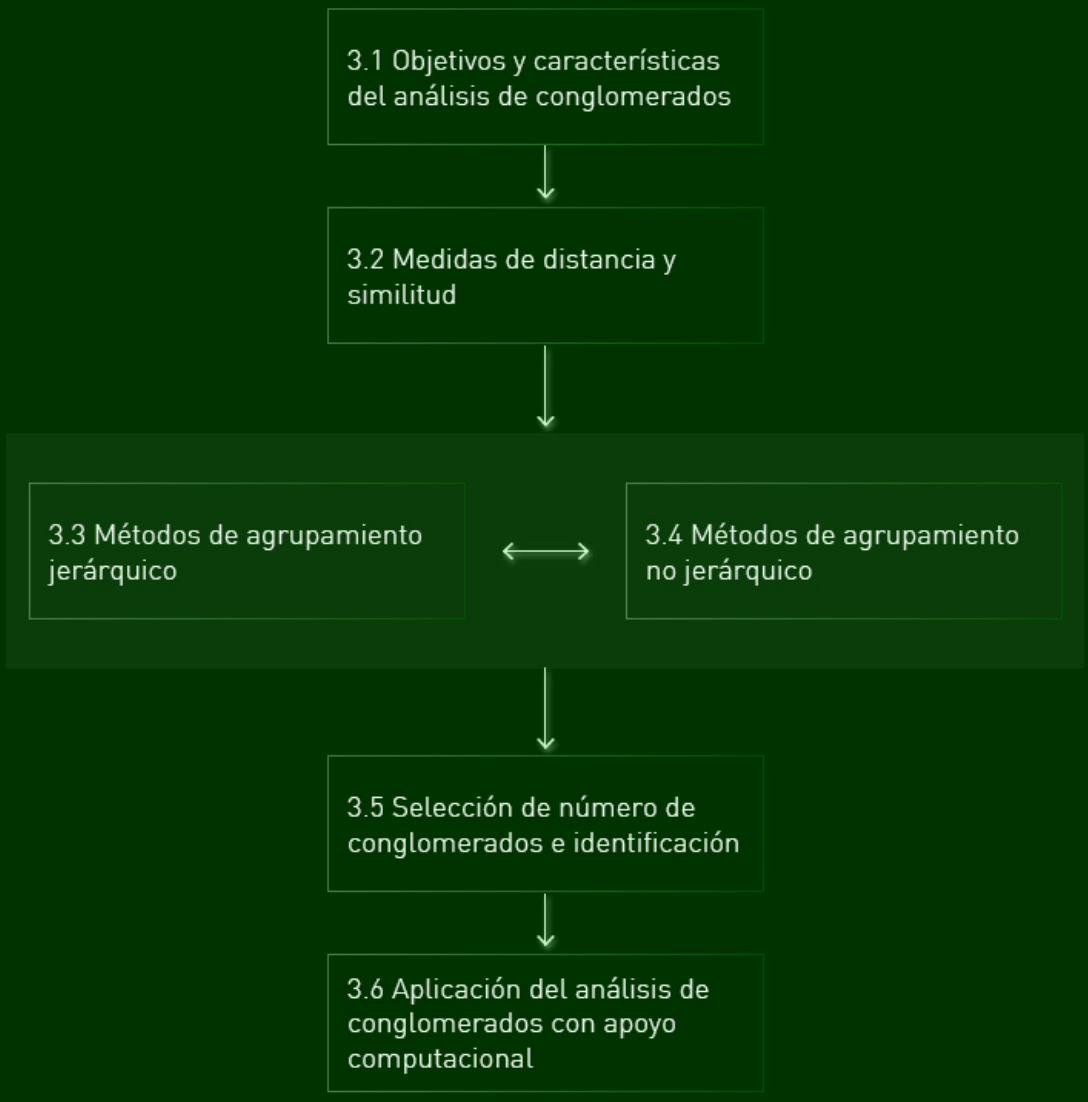
3) Repetir el análisis empleando máxima verosimilitud a menos que los datos exhiban claras desviaciones a la normalidad.

4) Si se cuenta con suficientes datos, probar la estabilidad de la solución mediante validación cruzada.

El análisis factorial ha demostrado ser una herramienta analítica útil para el diseño de instrumentos de medición en proyectos de inteligencia de mercados, que al igual que otras técnicas exploratorias, requiere no solo de competencias estadísticas sino de habilidad y conocimientos en el tópico que se estudia para así darle significado y utilidad a los resultados que proporciona.

Capítulo 3
Segmentación de consumidores a través del análisis de conglomerados

Organizador temático



3. Segmentación de consumidores a través del análisis de conglomerados

Introducción

Las empresas enfocadas al cliente no solo diseñan buenos productos y servicios sino que, además, consideran las necesidades específicas de grupos de consumidores. La segmentación se define como el proceso de subdivisión o partición del mercado en conjuntos de individuos con características, necesidades y patrones de compra similares. La segmentación permite desarrollar una estrategia de mercado basada en la diferenciación de productos y servicios enfocados a satisfacer las necesidades de ciertos nichos de mercado. Hay múltiples formas de realizar la segmentación, entre las más simples están el uso de atributos geográficos como zonas urbanas, semiurbanas y rurales; las características demográficas del consumidor como edad, género y estado civil; variables socioeconómicas como nivel de ingreso y tipo de vivienda.

Pero los segmentos de mercado también se pueden conformar en términos de variables menos evidentes como son las psicográficas y la búsqueda de ciertos beneficios cuando se hace una compra (Picón-Prado, Varela-Mallou y Lévy Mangín, 2004). Las variables

psicográficas son las relacionadas con el estilo de vida del consumidor tales como actitudes, sentimientos, conductas y personalidad. Al usar estas variables se asume que los individuos con intereses, preocupaciones y estilos de vida similares tienen patrones de compra y necesidades análogas. En cuanto los beneficios la idea es agrupar a los individuos en función de los criterios que utilizan para elegir un producto o servicios, y de los atributos que valoran más. La segmentación con variables psicográficas y por beneficios es una alternativa para la identificación de grupos de consumidores con necesidades y patrones de consumo afines que pueden ser atendidos o atraídos con una mezcla de mercadotecnia diferenciada. Este tipo de segmentación se ha utilizado en diversos contextos entre ellos el de protección al medio ambiente; la caracterización de individuos más proclives a apoyar el cuidado al ambiente se describe en el cuadro anexo.

A partir de los datos de una encuesta realizada en el 2004 en Portugal, Vicente y Reisa (2005) segmentaron a 2,000 jefes de familia de acuerdo con sus actitudes hacia el reciclaje, su participación en actividades de reciclaje, la importancia que le dan a los incentivos para reciclar y su perfil demográfico. La segmentación permitió identificar segmentos con disposiciones distintas hacia el reciclaje, esta información es importante para definir estrategias de comunicación y mercadotecnia específicas a cada segmento y encaminadas a promover el reciclaje entre los ciudadanos. La segmentación se hizo utilizando análisis de conglomerados jerárquico.

El análisis fue aplicado después de simplificar la estructura de los datos a través del uso de componentes principales. Esta simplificación fue útil para interpretar los resultados del análisis de conglomerados y caracterizarlos. Tres conglomerados fueron definidos: 1) individuos con actitudes favorables al reciclaje para quienes la conservación del ambiente es un tema crítico; 2) individuos con actitudes neutrales hacia el reciclaje pero que

apoyan la práctica porque son sensibles a la presión social y las normas personales, y 3) individuos escépticos hacia el reciclaje e indiferentes en cuanto a apoyar su práctica. Los primeros dos segmentos son el target para planes de mercadotecnia social que enfaticen los resultados positivos del reciclaje y se difundan en medios de comunicación masiva.

3.1 Objetivos y características del análisis de conglomerados

El análisis de conglomerados se refiere a la creación de grupos de objetos en términos de sus patrones de similitud en cuanto a un conjunto de variables. La idea clave es la representación de N objetos o entidades en K grupos homogéneos, esto es formados por objetos con características similares, pero con un alto grado de heterogeneidad entre ellos. Estos grupos de objetos o individuos se denominan conglomerados o clusters. Tryon (1939) fue el primero en utilizar el término **análisis de conglomerados** el cual comprende

una amplia variedad de métodos o algoritmos diseñados para realizar el agrupamiento de objetos. Muchas áreas del conocimiento relacionadas con los negocios requieren de agrupar individuos u objetos con el propósito de sugerir segmentos con características, hábitos, niveles de similares o que buscan beneficios específicos en los productos o servicios que adquieren. Una vez caracterizados los grupos, es posible diseñar productos y estrategias de mercadotecnia diferenciadas según el perfil de los segmentos identificados.

Muchas áreas del conocimiento relacionadas con los negocios requieren de agrupar individuos u objetos con el propósito de sugerir segmentos con características, hábitos, niveles de similares o que buscan beneficios específicos en los productos o servicios que adquieren. Una vez caracterizados los grupos, es posible diseñar productos y estrategias de mercadotecnia diferenciadas según el perfil de los segmentos identificados.

Existen varios objetivos que pueden ser cubiertos por el análisis de conglomerados:

1. Creación de taxonomías. Un problema que enfrentan muchos investigadores es el de organizar los datos observados en taxonomías que tienen las mismas características subyacentes. El análisis de conglomerados permite definir estas taxonomías en función de un conjunto de variables que caracterizan a los objetos sin que se busque explicar el porqué de los agrupamientos formados. La estructura subyacente que explica el grado de homogeneidad entre individuos que forman parte de un mismo grupo y la heterogeneidad entre los varios grupos formados está en la fundamentación teórica de las áreas de conocimiento de los objetos de estudio. En consecuencia, es esencial que se hayan identificado variables apropiadas que describen las particularidades de los objetos y por tanto que permiten distinguirlos de objetos en otras taxonomías.

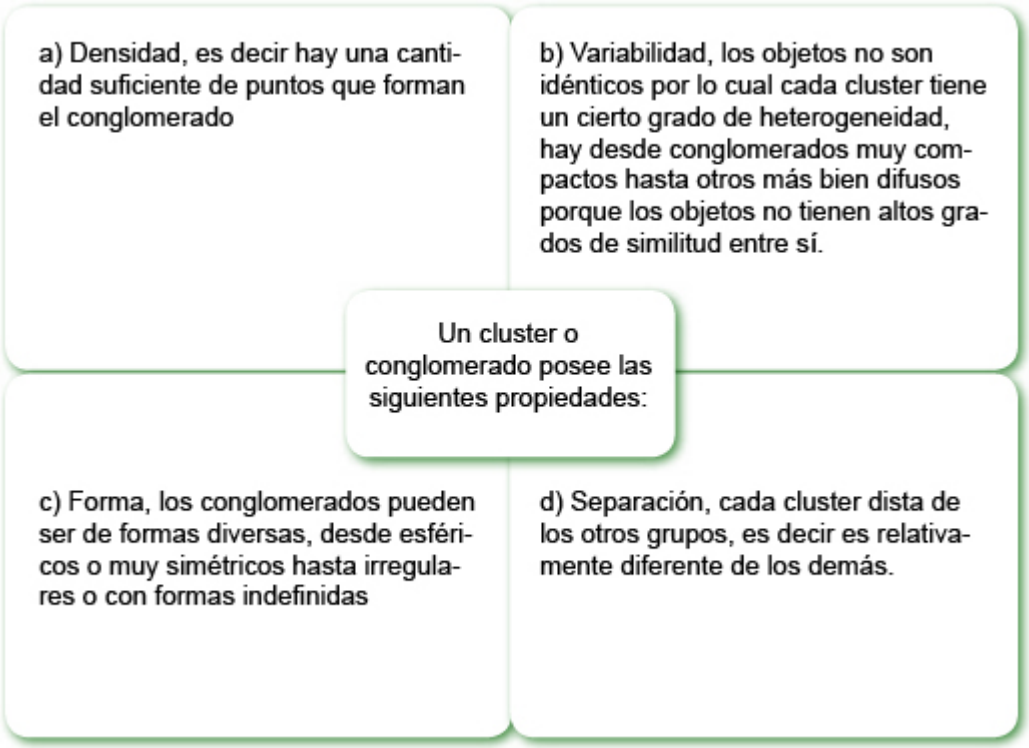
2. Definir hipótesis. El análisis de conglomerados permite formular hipótesis de investigación acerca de porqué los datos se agrupan cómo se agrupan y porqué difieren de otros grupos. Es decir que a partir del análisis, el investigador cuenta con los medios para establecer relaciones entre grupos de objetos y teorizar acerca de su cercanía y agrupación. Es importante aclarar que el análisis de conglomerados es una colección de algoritmos para el agrupamiento de objeto según criterios de similitud. Por tanto no hay pruebas estadísticas que permitan verificar si las sub-poblaciones o segmentos están bien constituidos o identificados.

3. Confirmar hipótesis. El análisis de conglomerados puede utilizarse también para propósitos confirmatorios acerca de la agrupación teórica de objetos, es decir a partir de sólidas bases conceptuales sugeridas por un área de conocimiento se establece una cierta categorización de los objetos la cual se puede verificar empíricamente con el análisis de conglomerados.

Es importante reconocer entre los problemas de agrupamiento y de clasificación. En el primer problema los grupos no están definidos a priori el propósito del análisis es utilizar la información de múltiples variables que describen a los individuos para dividirlos en grupos también conocidos como clusters o conglomerados En el caso del problema de clasificación, los grupos están definidos a priori, por lo que el problema es asignar o clasificar a un nuevo objeto en alguno de los grupos ya definidos con base en un conjunto de variables y normas (Nemery, 2006). El análisis de conglomerados se considera un tipo de clasificación no supervisada ya que se trata de asignar objetos a grupos formados por objetos que comparten características similares. Como se mencionó anteriormente, el análisis de conglomerados no involucra la construcción de inferencias estadísticas formales ni la formulación de hipótesis respecto a la estructura de los datos por lo que califica como un método exploratorio que permite identificar categorías o patrones encubiertos en los datos. En consecuencia, la interpretación de los resultados del análisis se fundamenta en los objetivos de la

investigación, las habilidades, conocimientos y experiencia del analista.

Un conglomerado se define como un conjunto de puntos compacto y asilado de otros grupos (Jain, 2009), no corresponde a una población de objetos o segmento de individuos con parámetros diferentes a los de otros grupos, es más bien una entidad subjetiva cuya interpretación se fundamenta en los conceptos y teorías del área en que será aplicado.



En el diagrama de la Figura 3.1 se describe gráficamente la estructura de los conglomerados en dos dimensiones: se identifican cuatro patrones de agrupamiento para los datos, uno de los conglomerados, el de mayor tamaño, es más difuso e irregular en forma que los otros tres; dos de estos conglomerados son esféricos y están poco separados entre sí mientras el tercero está bien distante de todos los clusters y es el de menor tamaño.

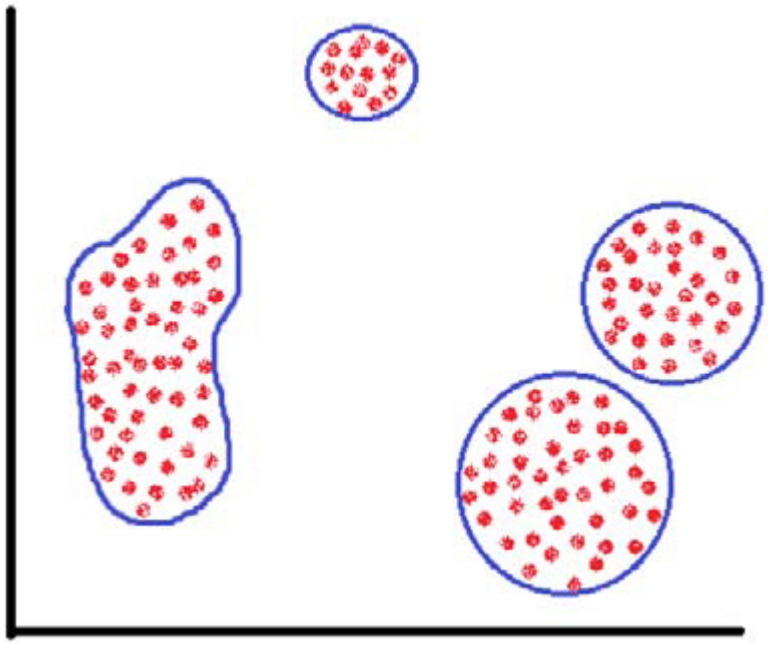


Figura 3.1 Conglomerados en el espacio bidimensional

¿Sabías qué?

La clasificación de los animales y las plantas ha jugado un rol muy importante en el ámbito de la biología y es el cimiento para la Teoría de la Evolución de Charles Darwin. Las técnicas cuantitativas para el agrupamiento de especies según su patrón de similitudes se formularon para liberar a la taxonomía de sus arraigadas interpretaciones subjetivas.

3.2 Medidas de distancia y similitud

El primer paso para realizar un análisis de conglomerados es definir una medida de similitud entre los individuos que se desea agrupar. Una medida de similitud apropiada debe permitir representar a los individuos u objetos como puntos en el espacio, en donde se les puede observar como distancias métricas entre ellos;

tal métrica de similitud y debe poseer las siguientes propiedades (Jhonson y Wichern, 1996):

a) Simetría. La similitud (distancia) entre el objeto “x” y el “y” es la misma que entre el objeto “y” y el “x”, es decir: d(x,y) = d(y,x)

b) Desigualdad del triángulo. Si se considera el plano bidimensional, la menor distancia entre dos objetos está definida por su distancia diagonal, la cual es menor que la suma de las distancias de los objetos respecto a un tercero, esto es: d(x,y) < d(x,z) + d(y,z)

c) Distinción. Si la medida de distancia, similitud, es diferente de cero, los objetos son diferentes: d(x,y) ≠ 0 , entonces x ≠ y

d) Identidad. Si la medida de distancia es de cero, los objetos son idénticos, lo que se expresa como si x y x’ idénticos, entonces d(x,x’) = 0

Si todas las variables del vector x que describe a un objeto cualquiera son métricas, esto es están en escalas de intervalo o razón, las medidas de similitud más simples que se pueden utilizar son medidas de distancia. La distancia euclideana o euclídea es la más utilizada ya que corresponde a la distancia geométrica entre dos objetos en el espacio multidimensional. El diagrama de la Figura 3.2 muestra la desigualdad del triángulo para el caso de la métrica de similitud de distancia euclideana.

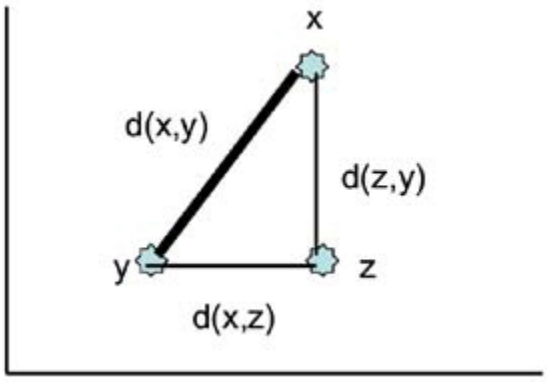


Figura 3.2 Distancia euclideana entre objetos: desigualdad del triángulo

En general para una pareja de objetos cualquiera (i, j), cada uno caracterizado cada uno por un vector de variables $\mathbf{x}_i$ e $\mathbf{x}_j$ respectivamente, la distancia euclideana se calcula como sigue:

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} = (x_i - x_j)'(x_i - x_j) \quad (1)$$

El último término en la expresión (1) indica que la distancia euclideana resulta de multiplicar el vector renglón o vector transpuesto de las diferencias entre los objetos i y j, lo cual se denota por el símbolo ‘, para el conjunto de variables por el mismo vector columna. Se puede también utilizar la distancia euclideana al cuadrado, se omite extraer raíz cuadrada, para describir la similitud entre objetos.

Ambos tipos de distancia son sensibles a las diferencias en la escala de las variables por ejemplo cuando una está expresada en miles de pesos y otra en una escala de preferencia de solo 5 categorías, razón por la cual las distancias o similitudes entre individuos son generalmente calculadas de los datos estandarizados, ver capítulo 1, de tal manera que todas las variables tengan el mismo promedio (0) y variabilidad (varianza de uno).

Esta estandarización evita que la medida de distancia realmente refleje las similitudes o diferencias para todo el conjunto de variables y que no esté dominada por aquella variable que tiene las unidades más grandes.

Otras medidas de distancia que también se utilizan en el análisis de conglomerados son las siguientes:

Distancia Manhattan. También conocida como distancia entre las cuadras de una ciudad (city-block), esta distancia es simplemente la suma de las diferencias en valor absoluto para todas las variables que caracterizan a cualquier pareja de individuos. Una ventaja de esta medida de distancia es que al no elevar al cuadrado las diferencias entre objetos la medida es menos sensible a la

presencia de datos atípicos. El cálculo para la distancia Manhattan se da en la expresión (2)

$$d_{ij} = \sum_{i=1}^{p} | x_{ik} - x_{jk} | \qquad (2)$$

Distancia de Minkowski. Las distancias euclidiana y Manhattan son casos particulares de esta medida de distancia generalizada cuando m =2 y m =1 respectivamente, la distancia Minkowski se calcula según la expresión (3).

$$d_{ij} = \left[\sum_{i=1}^{p} | x_{ik} - x_{jk} |^{m} \right]^{1/m} \qquad (3)$$

Distancia de Mahalanobis. Esta es una medida de distancia estadística que ya se describió en el primer capítulo de este eBook y que toma en cuenta, además de las diferencias netas entre los dos vectores de variables $\mathbf{x}_i$ e $\mathbf{x}_j$ que caracterizan a dos objetos, los patrones de asociación entre las variables medidas. La distancia de Mahalanobis está dada en la expresión (4) en la cual ***S*** es la matriz de varianza covarianza del vector multivariable **x**.

$$d_{ij} = (x_i - x_j)' S^{-1} (x_i - x_j) \qquad (4)$$

Las distancias entre todos los objetos o individuos que se desean agrupar se resumen en un arreglo matricial o matriz de distancias D.

Ejemplo 1. México ha sido un destino turístico atractivo para el turismo internacional pero desafortunadamente ha perdido ventaja sobre otros países latinoamericanos. La Secretaría de Turismo desea mostrar que el atractivo de México es comparable al de otros destinos turísticos, para ello se dio a la tarea de recopilar información para tres países latinoamericanos populares entre los turistas norteamericanos y europeos en relación a dos variables X1 = cientos de kilómetros de playas con infraestructura de primer nivel y X2 = total de sitios declarados Patrimonio de la Humanidad.

	X1	X2
México	27	38
Costa Rica	11	29
Colombia	19	22

El cálculo de las distancias euclideanas entre todas las parejas de países es como sigue:

$$d(Mex, CR) = \sqrt{(27-11)^2 + (38-29)^2} = 18.3576$$
$$d(Mex, Col) = \sqrt{(27-19)^2 + (38-22)^2} = 17.8885$$
$$d(CR, Col) = \sqrt{(11-19)^2 + (29-22)^2} = 10.6301$$

La información anterior se resume en la siguiente matriz de distancias, note que los elementos en la diagonal son cero puesto que miden la similitud entre elementos idénticos (propiedad d).

$$D = \begin{bmatrix} 0.0000 & 18.3576 & 17.8885 \\ 18.3576 & 0.0000 & 10.6301 \\ 17.8885 & 10.6301 & 0.0000 \end{bmatrix}$$

Si bien las medidas de similitud basadas en distancias son fáciles de calcular y comprender, tienen la fuerte limitación de que únicamente pueden utilizarse cuando las variables elegidas para el agrupamiento son métricas. Otra clase de medidas de similitud, también fáciles de calcular e interpretar, son los llamados coeficientes de concordancia que permiten determinar la similitud entre individuos en términos de variables nominales que únicamente tienen dos categorías. Estas medidas se basan en el principio de que dos objetos son más similares entre sí cuando tienen más características en común. Para calcular estas medidas es necesario definir variables binarias que indican si cierto atributo está presente o ausente en un objeto. Sea x_{ik}=1 si el atributo k está presente en el objeto i; si x_{ik} = 0, el objeto i no posee el atributo k.

Si dos objetos cualesquiera, *i* y*j* se comparan entre sí, hay tres posibles casos: ambos objetos poseen el atributo k lo que resulta en (1,1); ninguno de los objetos cuenta con el atributo *k*, es decir ambas variables indicador son cero (0,0); y solo alguno de los dos objetos posee el atributo *k* lo que resulta en discordancias que se escriben como (1,0) ó (0,1).

Si se suma sobre todos los atributos, se obtiene el total de las concordancias entre los objetos, es decir del total de variables en que los objetos coinciden. El problema es que los casos en que ambos objetos poseen el atributo en cuestión (1,1) y aquellos en que el atributo está ausente en ambos objetos (0,0) se ponderan en forma equivalente. Aldenderfer y Blashfield (1948) argumentan que coincidir en un atributo representa un mayor grado de similitud entre dos individuos que cuando ambos carecen del atributo. En consecuencia se han sugerido dos coeficientes de concordancia:

a) S = total de concordancias/total de atributos.

b) S’ = total de concordancias positivas (ambos objetos poseen el atributo)/total de atributos.

Para calcular estos coeficientes que cuantifican la similitud entre objetos conviene construir una tabla cruzada en la cual se muestren las concordancias y discordancias entre los individuos que se comparan. A partir de la tabla se procede al cálculo de los coeficientes de concordancia.

Ejemplo 2. Los aspirantes a la gerencia de innovación de una empresa van a ser comparados entre sí respecto a los siguientes atributos:

X1 = el aspirante tiene dominio del idioma inglés

X2 = el aspirante cuenta con una maestría

X3 = el aspirante tiene experiencia de al menos dos años en un puesto gerencial

X4 = el aspirante está certificado en el manejo de software estadístico y de minería de datos.

El perfil de dos aspirantes al puesto está descrito por los siguientes vectores:

$$x_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}, x_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Ambos aspirantes coinciden en su dominio del idioma inglés y en su nivel de experiencia en puesto gerencial, esto es hay dos concordancias positivas (1,1). Ambos coinciden también en no tener certificación en el uso de software estadístico y de minería de datos, esto es hay una concordancia negativa (0,0) y también una discordancia ya que el sujeto 1 cuenta con maestría pero no el aspirante 2. En la siguiente tabla cruzada se reportan las concordancias y discordancias.

Sujeto 1 / sujeto 2	**Posee atributo(1)**	**Posee atributo(2)**
Posee atributo (1)	2	1
No posee atributo(0)	0	1

Por tanto el coeficiente S = (2+1)/4 = 0.75, en tanto el coeficiente S’ = 2/4 = 0.5. Notar que entre mayor el coeficiente, mayor la similitud entre los objetos. Si S=1 ó S’=1 los objetos serían idénticos en el sentido de que los atributos coinciden en tener o no los mismos atributos.

El cálculo de los coeficientes de concordancia anteriores permite apreciar que la similitud entre objetos no únicamente se puede cuantificar en términos de medidas de distancia y que también es viable comparar objetos empleando características no-métricas.

Cuando las variables que son la base del agrupamiento tienen más de dos categorías resulta más complicado no solo el definir medidas de similitud sino también algoritmos para el agrupamiento de objetos. Buchta, Dolnicar y Reutterer (2002) discuten métodos de agrupamiento para la segmentación en mercadotecnia cuando las variables que tipifican a los objetos están en escalas nominales.

3.3 Métodos de agrupamiento jerárquico

Una vez definida la medida de distancia o similitud entre objetos, el siguiente paso en el análisis de conglomerados es elegir un procedimiento que establezca reglas sobre cómo agrupar los objetos. Pueden utilizarse varios algoritmos para formar los clusters, los algoritmos disponibles difieren en términos del procedimiento que usan para agrupar los objetos. Los llamados algoritmos jerárquicos van uniendo, o separando, objetos sucesivamente formando una estructura jerárquica que gráficamente se representa en un gráfico denominado dendograma. Este diagrama se puede

crear de dos formas: de abajo hacia arriba o de arriba hacia abajo. El primer caso es el más utilizado, se comienza con cada objeto formando un grupo separado y se busca combinar sucesivamente a los demás objetos hasta que todos se hayan unido en un único grupo. Los algoritmos jerárquicos que siguen esta secuencia son del tipo aglomerativo o de amalgamiento. Los algoritmos de división proceden de abajo hacia arriba, todos los objetos forman parte de un mismo cluster; en cada paso o iteración se busca formar subconjuntos más pequeños de objetos hasta que cada uno de los objetos esté en un cluster individual. Para ambos tipos de algoritmos el dendograma muestra cómo se van eslabonando, o dividiendo los objetos, se indica también el valor de la medida de distancia o similitud a la cual ocurrió la combinación o división de los clusters.

Cuando se utilizan algoritmos jerárquicos de amalgamiento, el tamaño y variabilidad de los clusters se incrementa en cada etapa ya que se van agrupando cada vez más objetos los cuales son menos similares entre sí. Con estos algoritmos, una posible asignación “incorrecta” de objetos a clusters, hecha en las primeras iteraciones, no se puede modificar posteriormente, puesto que una vez que un objeto es agrupado en un cluster, su asignación no cambia en iteraciones subsecuentes. Los algoritmos de aglomeración resultan útiles cuando los clusters son de forma irregular, sin embargo algunos de estos algoritmos, como el de eslabonamiento simple, tienden a formar grandes cadenas de objetos que resultan en conglomerados muy grandes y difusos (Pérez, 2004).

El procedimiento general que utiliza un algoritmo jerárquico de amalgamiento es el siguiente:

1) Iniciar con N conglomerados con un único objeto cada uno y una matriz D, de orden NxN que describe las distancias o similitudes de los N objetos que se van a agrupar.

2) En D identificar la pareja más cercana de objetos o conglomerados, sea d(U,V) la distancia entre ellos.

3) Combinar U con V en un nuevo conglomerados (UV). Actualizar D eliminando los renglones y columnas asignados a los objetos (clusters) U y V y modificar o actualizar las distancias del nuevo conglomerado (UV) respecto a los otros objetos (clusters).

4) Repetir N-1 veces los pasos 2 y 3 hasta que todos los objetos se agrupen en un único cluster.

Los algoritmos de amalgamiento o aglomeración más usados se describen a continuación:

1. Eslabonamiento simple, también conocido como el método del vecino más próximo (*nearest neighbor* o *NN-method*). En este método una vez que dos objetos o clusters se han unido para formar un nuevo conglomerado, la distancia entre el nuevo grupo y los otros elementos se re-define como d(UV,i) = min {d(U,i), d(V,i)}. La distancia entre dos conglomerados queda por tanto determinada por la distancia entre los dos objetos más cercanos vecinos más próximos que están en distintos conglomerados; por esa razón el método tiende a encadenar muchos objetos en clusters de forma irregular.

2. Eslabonamiento completo, en este algoritmo el criterio de distancia es d(UV,i) = max {d(U,i), d(V,i)}. Esta distancia representa la esfera de menor diámetro que cubre a dos posibles conglomerados candidatos a combinarse. Para que un elemento sea incluido en un determinado conglomerado es necesario que tenga cierto grado de similitud con todos los elementos que forman el conglomerado.

3. Eslabonamiento promedio, en este método la distancia actualizada entre dos clusters es el promedio de las distancias entre todas las parejas de objetos que están en grupos diferentes. Cuando se calcula el promedio de todas las distancias entre dos objetos recién agrupados en un nuevo conglomerado (UV) respecto a cada uno de los objetos en otros clusters se tienden a formar conglomerados con varianzas pequeñas y que contienen objetos bastante similares entre sí. Este método es el mejor para producir conglomerados esféricos.

4. **Método de centroides**, en este caso la distancia actualizada entre conglomerados se define como la distancia euclideana entre centroides, o vectores de promedios en el espacio multidimensional, definido por el conjunto de variables que describen a los datos. Cada vez que un individuo se agrupa a un cluster, se calcula un nuevo **centroide** y su distancia respecto a otros grupos queda determinada por la diferencia entre sus centroices. La principal fortaleza de este algoritmo es que se ve poco afectado por la presencia de observaciones extremas ya que el cálculo del centroide compensa las disimilitudes notables que pueda haber entre los objetos de un cluster.

5. Eslabonamiento de Ward. Este algoritmo es también conocido como el de mínima varianza ya que su objetivo es unir aquellos conglomerados que resulten en el menor incremento en la varianza de los grupos. El uso del **método de Ward** está bastante difundido porque suele discriminar mejor entre grupos en relación a otros métodos como el de amalgamiento simple, completo y promedio, produce conglomerados de similar tamaño y variabilidad lo que facilita su identificación. En el primer paso del proceso de agrupamiento, donde cada observación equivale a un conglomerado, la variabilidad de cada cluster es de cero. En cada paso subsecuente, el algoritmo evalúa todas las combinaciones posibles de dos conglomerados y se procede a calcular la suma de cuadrados entre los objetos del nuevo conglomerado, esta suma corresponde a las desviaciones al cuadrado de cada observación respecto al centroide del cluster al que está asignado, esto es SC = $\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$ donde i =1,2,…,*n* describe a cada uno de los objetos en el conglomerado en el *j*-ésimo conglomerado. La suma de cuadrados también se puede interpretar como la disimilitud o pérdida en información en que se incurre al combinar dos conglomerados. El método de Ward es eficiente para formar cluster elipsoidales, con tamaños mejor balanceados y de baja variabilidad, pero tiende a formar muchos conglomerados de tamaño pequeño. La definición del número de conglomerados a utilizar se facilita a través del análisis de los cambios en las sumas de cuadrados de los clusters (Jhonson y Wichern, 1996). En cada iteración se puede apreciar en

qué momento la combinación de dos clusters incrementa considerablemente la variabilidad, lo que resulta en un conglomerado heterogéneo que puede ya no ser aceptable para fines de estrategias de servicio o productos enfocados al perfil específico del cluster.

Dadas las propiedades de las medidas de distancia y de las ventajas y desventajas de los métodos de eslabonamiento antes descritos, la combinación que se recomienda es el uso de distancias euclideanas o de Mahalanobis con el método de eslabonamiento promedio o el de Ward.

Ejemplo 3. Las tiendas departamentales que operan en México se perciben diferencias en cuanto al posicionamiento.

La tienda Liverpool es la más reconocida porque siempre hay una tienda cerca y ofrece buenas baratas de temporada. El Palacio de Hierro se percibe como la más aspiracional y con mejores marcas mientras Suburbia y Sears se perciben como las que ofrecen los mejores precios y créditos. Para definir el nivel de diferenciación entre las seis principales tiendas departamentales de México se realizó una encuesta entre 50 consumidores quienes expresaron su percepción sobre cinco variables que describen características importantes de las tiendas. Los consumidores valoraron cada atributo de la tienda a través de una escala semántica diferencial de siete categorías. La mediana de las evaluaciones hechas por los consumidores se muestra en la Tabla 3.1.

Tienda departamental	Variedad	Calidad	Arreglo de la tienda	Ambiente	Servicio
Liverpool	7	6	6	6	6
Palacio de Hierro	7	7	7	7	7
Sears	6	6	5	6	5
Suburbia	4	4	4	4	4
C&A	4	4	5	2	5

Tabla 3.1 Evaluación mediana de tiendas departamentales mexicanas

El primer paso del análisis de conglomerados consiste en calcular las similitudes entre objetos, en este ejemplo las cinco tiendas. Dado que la escala de diferencial semántico, en el que se expresaron los distintos atributos de las tiendas, califica como una escala de intervalo (Malhotra, 2008 p. 276), la distancia euclideana entre diferentes tiendas es una medida de similitud apropiada. Para la pareja Liverpool, Palacio de Hierro esta medida se calcula como sigue:

$$d(L,PH) = \sqrt{(7-7)^2 + (6-7)^2 + (6-7)^2 + (6-7)^2 + (6-7)^2} = 2$$

De igual manera se calculan el resto de las distancias entre parejas de tiendas, el resultado de estos cálculos se presentan en la matriz de distancias de la Tabla 3.2.

	Liverpool	Palacio de Hierro	Sears	Suburbia	C&A
Liverpool	0.00				
Palacio de Hierro	2.00	0.00			
Sears	1.73	3.32	0.00		
Suburbia	5.00	6.71	3.74	0.00	
C&A	5.57	7.14	4.90	2.45	0.00

Tabla 3.2 Matriz de distancias entre tiendas departamentales que operan en México

Paso 1. En la matriz anterior, la distancia entre Liverpool y Sears es la menor de todas, por lo que se elige agrupar a dichas tiendas en un nuevo conglomerado o cluster A = {Sears, Liverpool}. Empleando el método de los centroides, el siguiente paso es recalcular la matriz de distancias entre los restantes objetos (tiendas) y el centroide del nuevo cluster A. En el primer renglón de la siguiente tabla se reporta el centroide o promedio de las dos tiendas recién agrupadas para cada una de las cinco variables usadas para describir a las tiendas.

Tienda departamental	Variedad	Calidad	Layout	Ambiente	Servicio
Cluster A (Sears - Liverpool	6.5	6	5.5	6	5.5
Palacio de Hierro	7	7	7	7	7
Suburbia	4	4	4	4	4
C&A	4	4	5	2	5

Con los datos de la tabla anterior se recalcula la matriz de distancias, la cual ahora contiene la información de las similitudes

de únicamente cuatro objetos: cluster A, Palacio de Hierro, Suburbia y C&A.

	Cluster A	Sears	Suburbia	C&A
Cluster A (Sears - Liverpool	0.00			
Palacio de Hierro	2.60	0.00		
Suburbia	4.33	6.71	0.00	
C&A	5.17	7.14	2.45	0.00

Si en lugar del método de los centroides se aplica otro algoritmo, la matriz de distancias se actualiza de manera diferente. Los métodos de amalgamiento simple, completo y promedio no requieren de calcular centroides, la actualización de las distancias se realiza al emplear las entradas de la matriz de distancias reportada en la Tabla 3.2.

Con amalgamiento simple, una vez que las tiendas Liverpool y Sears se han agrupado en un nuevo cluster A, la distancia de este cluster respecto a las otras tiendas se calcula como sigue:

$$d(A, PH) = \text{mín}\{d(Sears, PH), d(Liverpool, PH)\} = \text{mín}(3.32, 2) = 2$$

$$d(A, Suburbia) = \text{mín}\{d(Sears, Sub), d(Liv, Suburbia)\} = \text{mín}(3.74, 5.00) = 3.74$$

$$d(A, C\&A) = \text{mín}\{d(Sears, C\&A), d(Liv, C\&A)\} = \text{mín}(4.90, 5.57) = 4.90$$

De donde la matriz de distancias revisada queda como sigue:

	Cluster A	Palacio de Hierro	Suburbia	C&A
Cluster A (Sears - Liverpool	0.00			
Palacio de Hierro	2.00	0.00		
Suburbia	3.74	6.71	0.00	
C&A	4.90	7.14	2.45	0.00

En la matriz anterior, la pareja de objetos más similares o que guardan la menor distancia entre sí (2.00) son el cluster A y el Palacio de Hierro. Es importante hacer notar que la secuencia de agrupamiento es diferente a la del método de los centroides para el cual los objetos que se unieron fueron Suburbia y C&A. En general, los diferentes algoritmos de agrupamiento proporcionan distintas estructuras por lo que la aplicación de otros métodos no es una opción para validar la solución obtenida.

Paso 2. Continuando con el método de los centroides, una vez que las tiendas Suburbia y C&A quedan agrupadas en un nuevo cluster, es necesario calcular el centroide del grupo recién formado el cual se reporta en el primer renglón de la tabla siguiente.

Tienda departamental	Variedad	Calidad	Layout	Ambiente	Servicio
Cluster A (Sears - Liverpool	6.5	6	5.5	6	5.5
Cluster B (Suburbia-C&A)	4	4	4.5	3	4.5
Palacio de Hierro	7	7	7	7	7

La nueva matriz de distancias después de esta iteración está dada por:

	Cluster A (Sears - Liverpool	Cluster B (Suburbia-C&A)	Palacio de Hierro
Cluster A (Sears - Liverpool	0		
Cluster B (Suburbia-C&A)	4.61	0	
Palacio de Hierro	3.28	6.82	0

Paso 3. En esta matriz distancias se observa que la distancia entre el Palacio de Hierro y el cluster A es la menor. Esto lleva a incorporar al Palacio de Hierro con las tiendas que ya componen el cluster A (Sears y Liverpool), formando un nuevo cluster que se identificará como C. En este tercer paso del algoritmo, todas las tiendas departamentales quedan agrupadas en dos grandes conglomerados C (Sears, Liverpool y Palacio de Hierro) y B (Suburbia y C&A). La última tabla que se construye muestra los centroides de estos dos conglomerados. De las entradas de la tabla se aprecia que los promedios para las cinco variables son menores para el cluster B que para el C lo que permite concluir que las tiendas de este segundo cluster–Sears, Liverpool y Palacio de Hierro- tienen mejores niveles de variedad y calidad de su mercancía, así como mejor layout, ambiente y servicio con respecto a las otras dos tiendas.

Tienda departamental	Variedad	Calidad	Layout	Ambiente	Servicio
Cluster C	6.75	6.5	6.25	6.5	6.25
Cluster B	6.5	6	5.5	6	5.5

El proceso completo de fusiones o formación de conglomerados termina cuando todos los objetos se combinan en un único cluster. Los resultados obtenidos en cada etapa de aplicación del algoritmo jerárquico se resumen gráficamente en el dendograma que se

muestra en la Figura 3.3 En el eje horizontal del diagrama se representan a las tiendas departamentales y en el eje vertical están las distancias a las cuales se agruparon estos elementos. Las tiendas departamentales Sears y Liverpool se fusionaron en el cluster A con un valor de distancia de 1.73; a este cluster se agregó el Palacio de Hierro para formar el cluster C a un nivel de distancia mayor e igual a 3.28. En tanto las tiendas Suburbia y C&A las cuales conforman el cluster B fueron agrupadas a una distancia de 2.45.

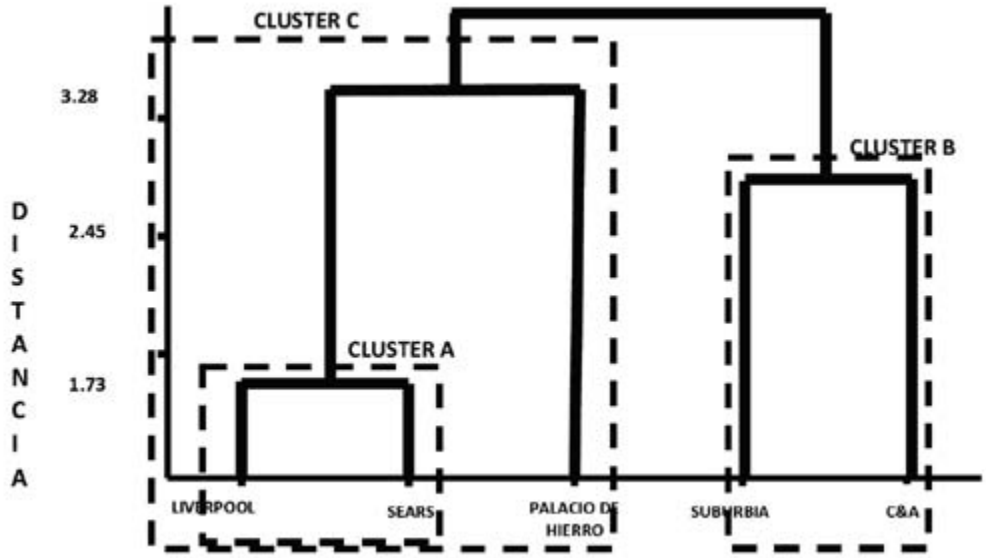


Figura 3.3 Dendograma para el agrupamiento de tiendas departamentales que operan en México.

Actividad de repaso (1)

3. Segmentación de consumidores a través del análisis de conglomerados

RELACIONAR COLUMNAS

1. Medida de similitud propuesta para variables binarias que consiste en cuantificar los casos de coincidencia entre dos individuos.	○ ○	Método de Ward
2. Las entradas de este arreglo representan las distancias entre los objetos que se desea agrupar.	○ ○	Coeficiente de concordancia
3. Son aquellos algoritmos que, en etapas sucesivas, unen elementos hasta que todos quedan asignados en un único grupo.	○ ○	Dendograma
4. Algoritmos que no permiten la reasignación de objetos, una vez que un elemento queda asignado a un conglomerado su pertenencia no se modifica.	○ ○	Métodos jerárquicos
5. Descripción gráfica de los resultados de un algoritmo jerárquico.	○ ○	Distancia euclideana
6. Medida de distancia estadística que toma en consideración los patrones de asociación entre las variables que describen los objetos.	○ ○	Matriz de distancias
7. Medida de distancia geométrica más utilizada para cuantificar la disimilitud entre objetos.	○ ○	Métodos de división
8. Algoritmo de agrupamiento jerárquico que asigna objetos a grupos bajo el criterio de minimizar la suma de cuadrados dentro de los grupos.	○ ○	Métodos de amalgamiento
9. Son algoritmos que dividen los conglomerados en subconjuntos más pequeños y homogéneos hasta que todos los objetos están en un cluster individual.	○ ○	Eslabonamiento simple
10. Método de amalgamiento que tiende a formar largas cadenas de objetos que conforman clusters irregulares.	○ ○	Distancia de Mahalanobis

Actividad de repaso (2)

3. Segmentación de consumidores a través del análisis de conglomerados

3.4 Métodos de agrupamiento no-jerárquico

Los algoritmos de agrupamiento no generan una estructura jerárquica como la descrita en un dendograma, lo que realizan es una división inicial de los datos en k clusters a los que se reasignan los objetos de tal manera que se minimice la desviación total de cada objeto respecto al centroide del grupo en donde fue clasificado. Los métodos no jerárquicos o de partición ofrecen ventajas sobre los jerárquicos cuando se desea agrupar una gran cantidad de datos, pues con el uso de un agrupamiento jerárquico el resultado sería un dendograma muy complicado. Los algoritmos involucran múltiples iteraciones en las que los objetos son reasignados hasta obtener la mejor configuración posible en el sentido de obtener clusters cada vez más homogéneos.

Los métodos no jerárquicos inician ya sea con una partición inicial de los objetos en k clusters o bien con un conjunto inicial de

“**semillas**” a las cuales se agregan objetos para formar los clusters (Johnson y Wichern, 1998). El hecho de tener que especificar el número de clusters a formar es el problema principal con estos algoritmos de partición, razón por lo cual se recomienda utilizarlos cuando ya se cuenta con una segmentación o agrupamiento inicial que requiera ser refinado, de otra manera se podría terminar con una estructura de agrupamiento inapropiada. El conformar los clusters a partir de un conjunto de semillas también representa un problema; si las semillas no son representativas de los grupos existentes, la solución será insatisfactoria. Los diagramas de la Figura 3.5 ilustran los inconvenientes de los algoritmos de partición. En el primer diagrama, los dos clusters circulares están muy cercanos entre sí, un mejor arreglo sería que todos los objetos en la parte superior del diagrama formaran un cluster y los ubicados en la parte inferior se combinaran en un segundo cluster. Bajo esta estructura los grupos serían más homogéneos en tamaño y estarían mejor separados, la solución que se representa en el diagrama es consecuencia de una incorrecta partición del número de conglomerados, tres en lugar de dos. A diferencia de los algoritmos jerárquicos que ofrecen al analista flexibilidad para especificar estructuras con diferente número de conglomerados, los algoritmos de partición dividen los datos en el número de conglomerados que el analista haya especificado. En el segundo diagrama, los objetos se han agrupado en dos conglomerados que mezclan objetos de diferente tipo, otro arreglo alternativo sería agrupar los objetos triangulares en un cluster y los estrella en otro. En este diagrama el problema fue la especificación de semillas a partir de las cuales se conformaron los clusters, las dos semillas elegidas fueron puntos estrella, lo que resultó en el arreglo espacial de objetos que describe la segunda parte de la figura.

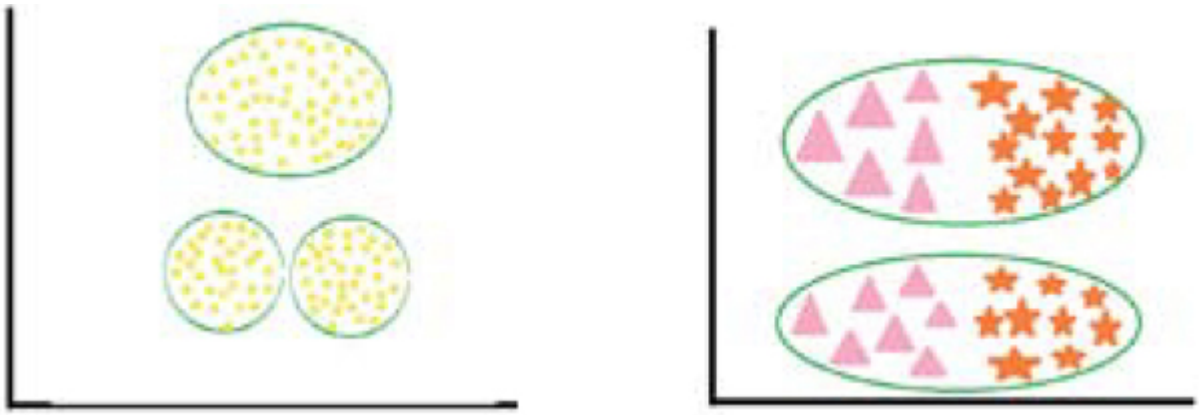


Figura 3.4 Problemas de aplicación de los algoritmos no jerárquicos o de partición.

Entre los algoritmos de partición más usados están los siguientes:

1. Valor crítico (*thereshold*) secuencial. Este algoritmo realiza el agrupamiento a partir de un conjunto de semillas. Se selecciona una de las semillas y todas las observaciones que estén dentro de un valor crítico respecto al dato-semilla se agrupan con éste. Después se continúa con otra semilla hasta obtener k conglomerados, cada uno conformado alrededor de un dato-semilla.

2. Procedimiento de valor crítico paralelo. En este algoritmo las semillas no se eligen sucesiva sino simultáneamente, todos los objetos dentro de un valor crítico se agrupan con la semilla más cercana.

3. Métodos óptimos. En esta categoría se incluyen métodos que inician con una partición inicial de objetos en k conglomerados y buscan maximizar una función de similitud que usualmente está definida como la suma de las distancias euclideanas de todos los objetos respecto al cluster de los conglomerados a donde están asignados. El algoritmo más popular y simple dentro de esta categoría es k-medias el cual fue propuesto desde 1955 y para el que se han propuesto múltiples variantes para mejorarlo (Jain, 2009), entre estas variantes figura una basada en un enfoque multicriterio que no requiere de especificar el número de clusters, lo que corrige la desventaja principal del algoritmo (Fernández, 2010).

Los pasos del algoritmo k-medias son los siguientes:

1. Los objetos se agrupan en k conglomerados iniciales.
2. Se calculan las distancias de cada objeto respecto a los centroides de los k clusters.
3. Los objetos se asignan a los conglomerados con el centroide más próximo al objeto.
4. Se vuelven a calcular los centroides de los conglomerados con los objetos re-asignados.
5. Se repiten los paso 2, 3 y 4 hasta que ya no haya más re-asignaciones de objetos a conglomerados.

El análisis de agrupamiento ya sea a través del uso de algoritmos jerárquicos o de k-medias se puede realizar empleando software estadístico comercial como SPSS o MINITAB.

Ejemplo 4. Considere el problema de agrupar las tiendas departamentales que operan en México, esta vez el algoritmo usado para formar los clusters es k-medias. Se especifican k=2 clusters. Las tiendas Liverpool y C&A se agrupan en un cluster A en tanto las otras tiendas -Sears, Suburbia y Palacio de Hierro- se agrupan en un segundo cluster identificado como B.

Se calculan los centroides de ambos conglomerados, los cuales se muestran en la Tabla 3.3.

Tienda departamental	Variedad	Calidad	Layout	Ambiente	Servicio
Cluster A (Liver-pool - C&A)	5.5	5	5.5	4	5.5
Cluster B (Sears-Suburbia-Palacio de Hierro)	5.7	5.7	5.3	5.7	5.3

Tabla 3.3 Centroides para una partición inicial en k = 2 clusters.

De acuerdo con el paso 2 descrito para el algoritmo k-means, se procede a calcular las distancias entre cada tienda departamental y

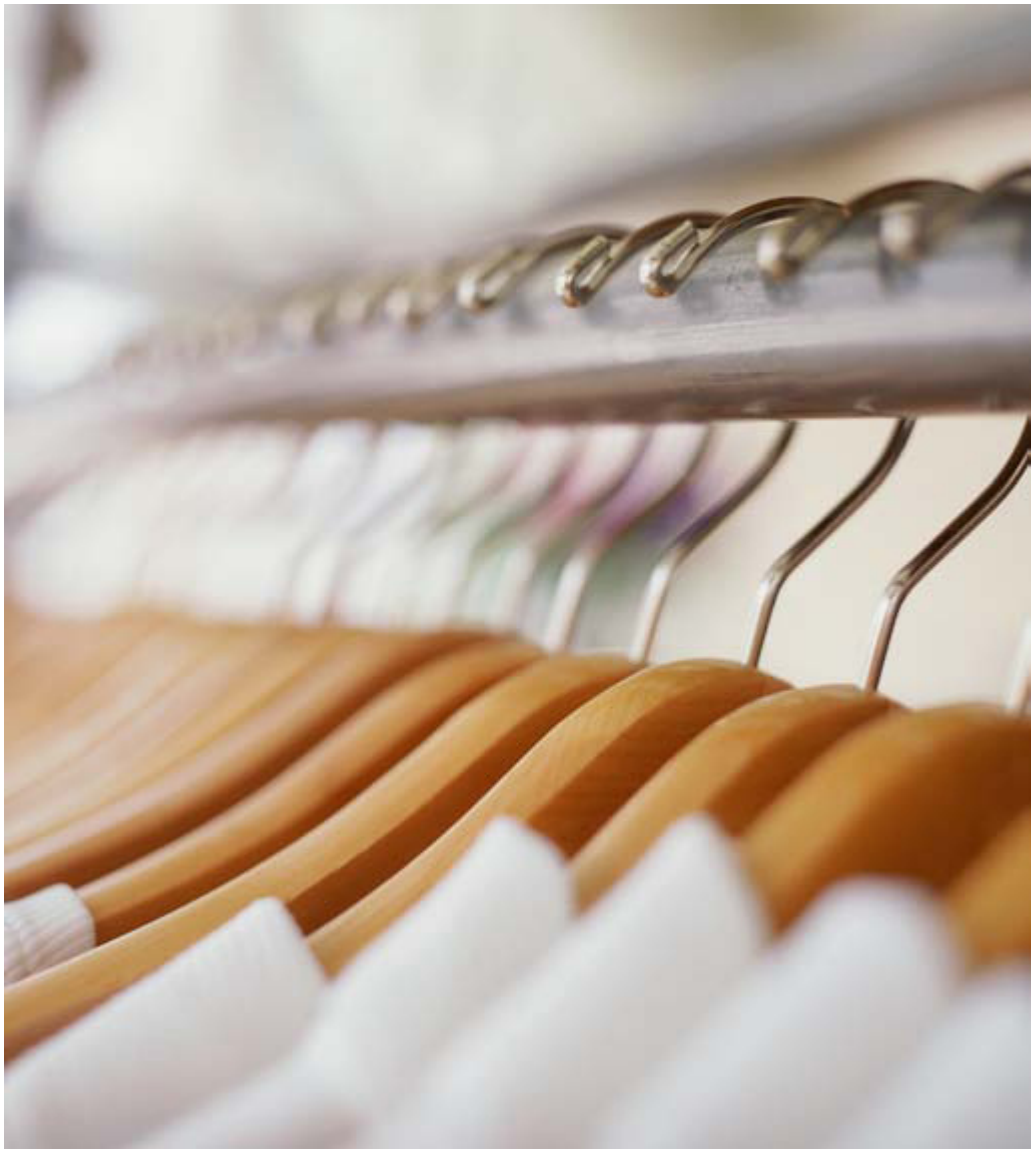
los centroides de los clusters iniciales A y B, las distancias se muestran en la Tabla 3.4.

Tienda departamental	Cluster A	Cluster B
Liverpool	5.15	0.50
Sears	4.18	1.63
C&A	1.22	5.77
Suburbia	1.22	5.10
Palacio de hierro	6.82	1.73

Puesto que no se produce ninguna reasignación de las tiendas departamentales, se declara una solución estable y termina el proceso de agrupamiento. Los clusters finales son A: Suburbia y C&A y B: Liverpool, Sears y Palacio de Hierro.

¿Sabías qué?

El agrupamiento y la clasificación han desempeñado una tarea esencial en el desarrollo de las teorías acerca de la comprensión de la naturaleza de la materia y de la estructura del átomo. Un ejemplo de esta labor es el trabajo de Dmitri Mendeléyev en 1860 para agrupar los elementos químicos de la tabla periódica.



3.5 Selección del número de conglomerados e identificación

Los algoritmos de análisis de conglomerados forman grupos de individuos aún cuando no exista una estructura de agrupamiento natural. La partición del total de objetos en grupos bien separados y uniformes no siempre indica que los datos provienen de segmentos bien definidos, razón por la cual resulta difícil establecer si una solución en exactamente k conglomerados es adecuada. La forma más usual de validar la solución es demostrar, a través de un

análisis de varianza, que las medias de los conglomerados difieren significativamente entre sí. Sin embargo esta comparación de medias es cuestionable puesto que los métodos de agrupamiento buscan minimizar la variabilidad dentro de los clusters y maximizar la variabilidad entre ellos, lo que es justamente el cociente en que está basada la prueba estadística de comparación de medias razón por lo cual la prueba es en general significante (Arnold, 1979). Dado que al análisis de conglomerados es una técnica exploratoria, la mejor recomendación es definir el número de clusters en función de su estructura, grado de homogeneidad, buena separación y densidad, así como la interpretación y aplicabilidad de la solución.

Cuando los clusters se pueden representar gráficamente como en la Figura 3.1, la calidad de la solución se aprecia visualmente. Pero cuando se emplean más de dos variables para realizar el análisis de agrupamiento, la estructura de la solución, en el espacio multi-dimensional, se establece a partir de la información numérica que proporciona el listado de MINITAB. Esta información incluye: la variabilidad de los clusters, medida a través de la suma de cuadrados de los objetos que integran el conglomerado, la distancia promedio de los objetos al centroide, la mayor distancia que un objeto tiene respecto al centroide y la distancia entre conglomerados. Esta información es útil para modificar el número de clusters. Por ejemplo cuando la variabilidad de un conglomerado es muy grande, significa que los objetos dentro de éste son poco semejantes, en este caso el analista puede tomar la decisión de dividir el conglomerado para formar subconjuntos de objetos más similares. El caso contrario ocurre cuando la distancia entre dos conglomerados es muy pequeña, en esta situación se puede optar por combinar los conglomerados más próximos; si bien esto incrementa la variabilidad, los nuevos grupos estarán más distantes entre sí. La forma de un conglomerado se puede conjeturar si se compara la distancia promedio de los objetos al centroide con la distancia máxima. Si la diferencia es pequeña el conglomerado es de forma esférica mientras una gran diferencia revela conglomerados de forma irregular que agrupan objetos con perfiles diversos que no pueden ser caracterizados por el centroide.

El **perfilado** o caracterización de los conglomerados requiere de identificar aquellas variables que, en promedio, asumen valores muy grandes o pequeños en cada cluster. Este examen se hace buscando nombrar y entender el agrupamiento generado a partir del perfil de los centroides. Puesto que, en el análisis de conglomerados, la práctica usual es estandarizar las variables, los centroides de cada conglomerado representan desviaciones respecto al gran centroide o vector promedio de todos los objetos el cual es igual al vector **0** ya que al estandarizar las variables su promedio global es de cero. Considere la gráfica de la Figura 3.5, en la que se han representado los promedios de tres conjuntos de botanas para tres diferentes variables: calorías por porción, contenido de grasa y contenido de sodio. En el gráfico es fácil apreciar que el cluster 1 incluye botanas cuyo contenido calórico es inferior al de los otros grupos (-2) mientras que las botanas del cluster 3 tienen una cantidad mayor de grasa (1.85). En cuanto a la cantidad de sodio, los tres grupos poseen cantidades similares.

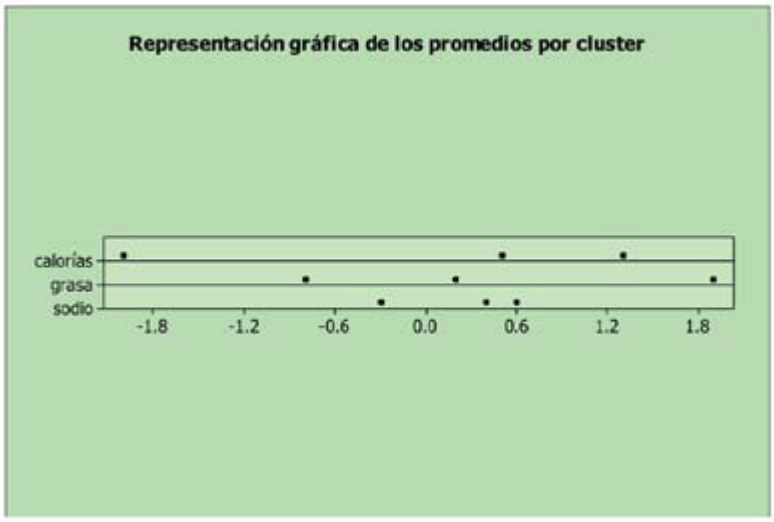


Figura 3.5 Comparación de centroides para una solución en k = 5 clusters

El listado de MINITAB para el análisis de conglomerado incluye una tabla con los centroides para cada uno de los clusters formados, a partir de ella se perfilan los grupos.

Ejemplo 5. Con el propósito de incrementar de enfocar mejor la publicidad que se hace en televisión, una agencia de investigación de mercados se dio a la tarea de obtener información sobre el perfil de la audiencia para varias series. Se recopiló información sobre las siguientes variables:

X1 = Edad promedio de la audiencia determinada a partir de 100 llamadas hechas a personas que veían el programa.

X2 = Nivel socioeconómico de los televidentes de la serie clasificado como 1 = nivel socioeconómico A+, 2 = nivel socioeconómico A, 3 = B, 4 = C, 5 = D

X3 = Promedio de escolaridad de la audiencia medido en años y a partir de 100 llamadas hechas a personas que veían el programa.

X4 = Género del individuo, 1 = femenino y 0 = masculino

X5 = Nivel de concordancia de los personajes de la serie con personas de la vida real, determinado como un índice de 0-10, donde 0 significa no concordancia de los personajes con la realidad y 10 total concordancia de los personajes con la realidad.

Veinticinco series televisivas se agruparon empleando el algoritmo k-medidas y como medida de similitud a la distancia euclideana. La información general con que contaba la agencia permitió especificar k = 3 conglomerados. El listado numérico de MINITAB es el siguiente:

Number of Within cluster Average distance Maximum distance

	observations	sum of squares	from centroid	from centroid
Cluster1	11	31.191	1.621	2.743
Cluster2	10	42.139	1.980	2.140
Cluster3	8	23.457	1.964	2.348

A partir del listado anterior se concluye que el conglomerado más homogéneo es el conglomerado tres el cual incluye 8 observaciones (series). El cluster más heterogéneo es el segundo con 11 observaciones.

El cluster 1 es el más irregular en su forma. Para este cluster 1, la diferencia (d promedio–d máxima = 1.122), esta diferencia es menor para los otros dos clusters lo que indica que tienen forma más bien esférica.

Distance between Cluster Centroids

	Cluster 1	Cluster2	Cluster3
Cluster1	0.0000	1.5554	3.2455
Cluster2	1.5554	0.0000	2.1877
Cluster3	3.2455	2.1877	0.0000

La matriz de distancias entre clusters que se reporta en esta porción del listado revela que los clusters más cercanos entre sí son el 1 y el 2 (distancia = 1.5554) mientras los clusters más alejados son el 1 y el 3 (distancia = 3.2455). Las distancias más cortas indican perfiles más cercanos y las más grandes indican conglomerados con características muy diferentes.

Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Gran centroide
edad	-1.0335	0.0907	0.4460	-0.0000
NSE	0.7796	-0.0584	-0.9179	0.0000
educación	-0.9270	0.3431	0.5558	0.0000
género	-0.6506	0.2947	1.0551	0.0000
concord	-0.6330	-0.1965	0.0730	-0.0000

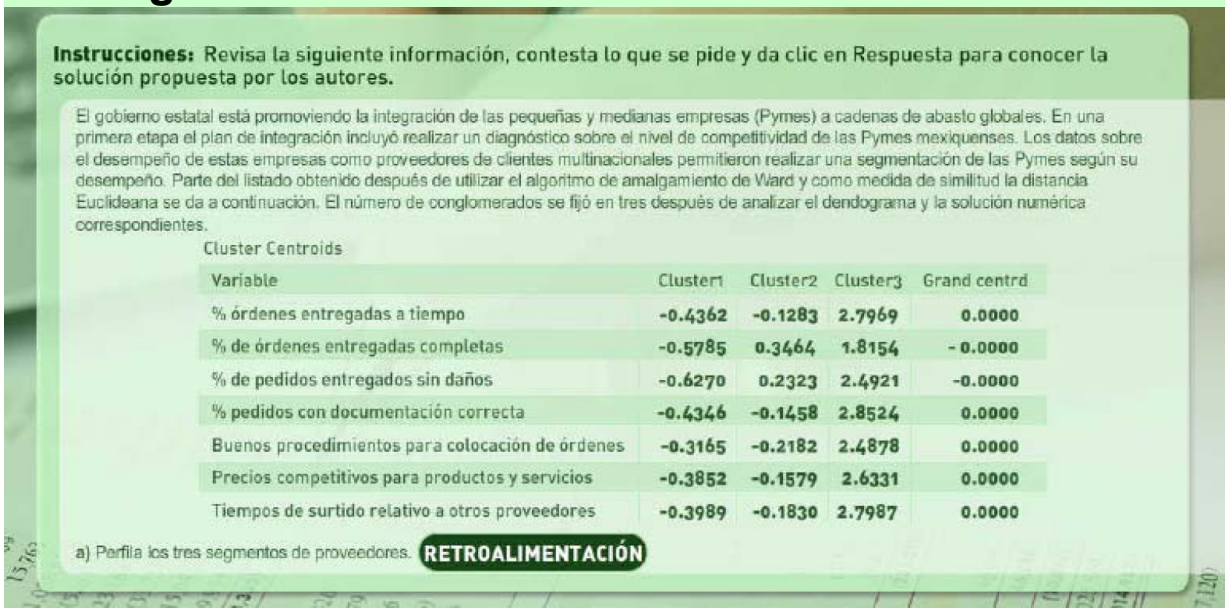
Esta última parte del listado permite identificar los clusters. La audiencia de las series del conglomerado 1 son individuos jóvenes de edad promedio menor a la de la audiencia total, con bajos niveles socioeconómicos, nivel educativo relativamente bajo y que gustan de series con personajes ficticios. Notar que el promedio de la variable género, que es nominal carece de un significado válido. Sin embargo el valor negativo del centroide indica que la audiencia es en su mayoría hombres (0 = género masculino). Se sugiere que los mensajes publicitarios para esta audiencia sean los relacionados a comida y ropa.

La audiencia de las series del cluster 2 se distingue por incluir individuos con un nivel educativo solo un poco arriba del promedio

global de los televidentes; no hay otras variables que particularicen este cluster para el que se puede hacer publicidad genérica. Finalmente la audiencia de las series que integran el conglomerado 3 se distingue por tener mayor edad que el promedio de televidentes, su nivel socioeconómico es el más alto y hay mayoría de mujeres. El perfil de esta audiencia sugiere una publicidad de viajes en cruceros, perfumes y ropa de marca.

Actividad de repaso

3. Segmentación de consumidores a través del análisis de conglomerados



Instrucciones: Revisa la siguiente información, contesta lo que se pide y da clic en Respuesta para conocer la solución propuesta por los autores.

El gobierno estatal está promoviendo la integración de las pequeñas y medianas empresas (Pymes) a cadenas de abasto globales. En una primera etapa el plan de integración incluyó realizar un diagnóstico sobre el nivel de competitividad de las Pymes mexiquenses. Los datos sobre el desempeño de estas empresas como proveedores de clientes multinacionales permitieron realizar una segmentación de las Pymes según su desempeño. Parte del listado obtenido después de utilizar el algoritmo de amalgamiento de Ward y como medida de similitud la distancia Euclideana se da a continuación. El número de conglomerados se fijó en tres después de analizar el dendograma y la solución numérica correspondientes.

Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Grand centrd
% órdenes entregadas a tiempo	-0.4362	-0.1283	2.7969	0.0000
% de órdenes entregadas completas	-0.5785	0.3464	1.8154	- 0.0000
% de pedidos entregados sin daños	-0.6270	0.2323	2.4921	-0.0000
% pedidos con documentación correcta	-0.4346	-0.1458	2.8524	0.0000
Buenos procedimientos para colocación de órdenes	-0.3165	-0.2182	2.4878	0.0000
Precios competitivos para productos y servicios	-0.3852	-0.1579	2.6331	0.0000
Tiempos de surtido relativo a otros proveedores	-0.3989	-0.1830	2.7987	0.0000

a) Perfila los tres segmentos de proveedores. RETROALIMENTACIÓN

3.6 Aplicación del análisis de conglomerados con apoyo computacional

El agrupamiento de objetos es un procedimiento que involucra varias etapas en cada una de las cuales es necesario considerar las cualidades de las medidas de similitud además de las ventajas y desventajas de los algoritmos disponibles para realizar el proceso de agrupamiento. El siguiente ejemplo de aplicación integra todas las etapas del proceso y utiliza software estadístico para realizar los cálculos los que, si bien son sencillos, resultan largos y tediosos como para efectuarse únicamente con el apoyo de una calculadora.

Ejemplo 6. Se reconoce que en México que los distintos estados de la república tienen diferentes perfiles de desarrollo, así los estados en la zona centro tienden a tener mayor nivel de población y competitividad que los estados del sureste, que tienen menores recursos y un porcentaje mayor de la población en niveles de pobreza o marginación. Se tiene el interés de agrupar objetivamente a los estados de la República Mexicana en términos de su crecimiento económico y de los servicios públicos con que cuentan sus ciudadanos.

El agrupamiento resultante permitiría al gobierno federal identificar perfiles de desarrollo regional relevante para implementar políticas y programas diferenciados según el perfil del cluster de estados. A través del sitio Web del Instituto Nacional de Geografía, Estadística e Informática (INEGI) se recolectaron datos para once variables que describen la situación de los estados. Las variables medidas fueron:

X_1: Porcentaje de desocupación

X_2: Inversión pública (miles de pesos)

X_3: Egresos en servicios públicos (miles de pesos)

X_4: Densidad de población

X_5: Número de hospitales con quirófano

X_6: Porcentaje de muertes violentas

X_7: Porcentaje de viviendas con agua potable

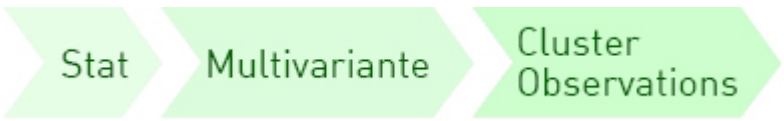
X_8: Número de árboles plantados

X_9: Crecimiento industrial

X_{10}: Producto interno bruto (miles de millones de pesos)

X_{11}: Inversión extranjera (miles de pesos)

En primer lugar se debe seleccionar una medida de distancia, puesto que todos los indicadores elegidos para describir la situación de los estados están en una escala de razón, las medidas de distancia resultan adecuadas para expresar la similitud entre entidades federativas. La distancia euclideana o de Mahalanobis son las opciones más apropiadas, en este caso se propone la distancia euclideana. Siguiendo la recomendación establecida en la sección 3.3, se usará el método de Ward que resulta en conglomerados altamente homogéneos y mejor diferenciados que los formados al utilizar otros métodos de amalgamiento. Los cálculos requeridos para realizar el análisis de agrupamiento se facilitan notablemente con el uso de software estadístico, en este caso MINITAB. La secuencia de comandos es la siguiente:



El programa despliega el cuadro de diálogo que se muestra en la Figura 3.6.

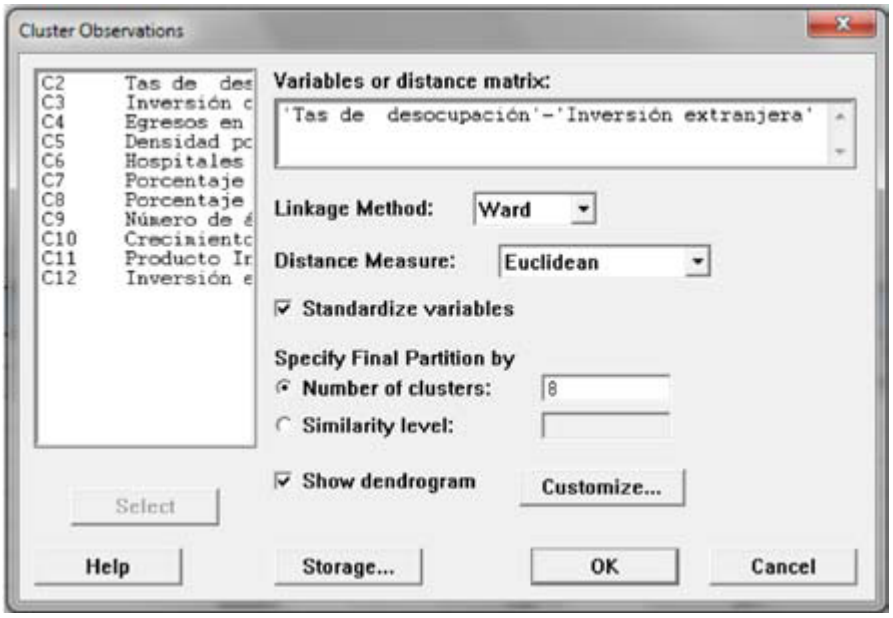


Figura 3.6 Ventana de diálogo principal para el agrupamiento de objetos con MINITAB

Para efectuar el análisis es necesario declarar en el recuadro correspondiente el nombre o columna que contiene las variables sobre las que se hará el agrupamiento. En este mismo cuadro de diálogo se define también la distancia euclídea como medida de distancia y al método de Ward como **método de aglomeración**. Dado que las variables están expresadas en distintas unidades, es conveniente estandarizarlas, lo que también se realiza marcando la opción correspondiente en el mismo cuadro de diálogo. Por último es necesario especificar ya sea un posible número de conglomerados o bien un límite en la medida de similitud a partir del cual resula inaceptable agrupar objetos. Se recomienda proponer un número de conglomerados que asegure que cada uno de ellos agrupe a una cantidad apropiada de objetos, ya que conglomerados muy pequeños, con solo uno a dos objetos, resultan poco relevantes para fines de una segmentación. Para definir el número de conglomerados en los que ha de dividirse la muestra de objetos es importante analizar el dendograma correspondiente, que también se elije en el cuadro de diálogo.

Es deseable, además, identificar cuáles entidades federativas fueron agrupadas en un mismo conglomerado, para ello es necesario solicitar que se almacene el grupo al que cada

observación fue asignada. La opción storage de MINITAB permite generar una nueva columna en la hoja de trabajo en la que se indica el cluster en que cada objeto fue agrupado; la columna que se declare debe estar vacía pues de otra forma la nueva columna sustituiría a la que ya contenía datos. Los detalles para guardar no solo el cluster de pertenencia sino también la distancia de cada dato al centroide de su cluster además de la matriz inicial con las distancias entre los objetos, se muestran en la ventana de diálogo de la Figura 3.7.

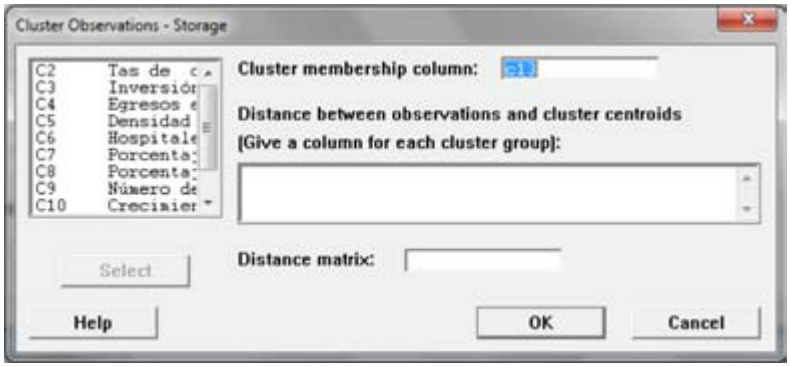


Figura 3.7 Cuadro de diálogo para almacenar información de la asignación de objetos

El resultado básico del análisis de conglomerados jerárquico es el dendograma que se muestra en la Figura 3.8.

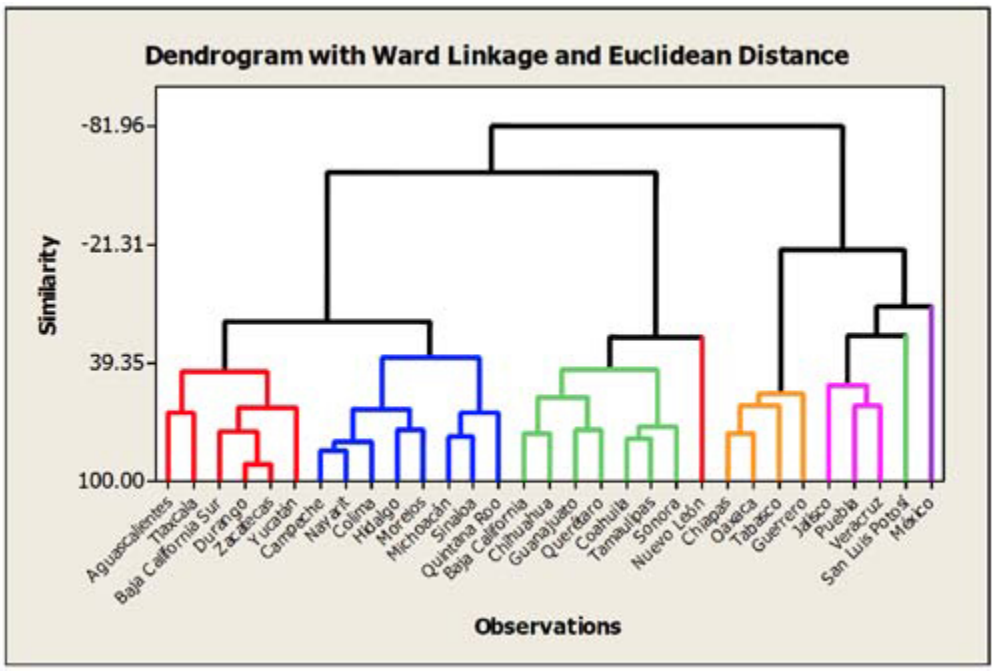


Figura 3.8 Dendograma para el agrupamiento de entidades federativas en México

En el dendograma anterior MINITAB muestra los conglomerados de estados que se formaron, los cuales están marcados con colores diferentes. Aparte de la solución en k = 8 conglomerados que generó el dendograma, otras soluciones como k = 3 ó k = 6 son evidentes del diagrama. El dendograma también permite identificar un cluster con un único elemento, el estado de México. Esta entidad no fue agrupada con otras porque su perfil es muy distintivo, el analista tendrá que decidir si elimina esta entidad del agrupamiento y lo considera como un elemento único.

Un agrupamiento en k = 3 clusters resulta poco diferenciador ya que en México se reconocen al menos cinco regiones (Centro, Norte, Pacífico, Sureste, Golfo) con particularidades económicas y de consumo. Mientras que la solución en k = 8 resulta en una división excesiva con tres clusters individuales, identificados con los colores rojo, verde bandera y azul en el dendograma, por tanto la decisión es analizar a detalle la solución en k = 6 conglomerados.

Para perfilar los conglomerados se utiliza la porción del listado de MINITAB que reporta las distancias de los centroides de conglomerado respecto al gran centroide o vector promedio de todos los datos. La Tabla 3.6 muestra las distancias correspondientes, en la tabla se han resaltado las diferencias más grandes entre los centroides de los conglomerados y el gran centroide. Estas distancias permiten identificar las características distintivas de los seis segmentos formados.

	Baja inversión pública y extranjera	Estados con maquila	Bajo crecimiento	Bajo crecimiento y alta inseguridad	Estados inversionistas	Motor desarrollado
1) Desocupación	0.64	0.75	-0.65	-0.94	-0.48	1.06
2) Inversión obra pública	-0.76	0.57	-0.42	0.03	0.63	0.75
3) Egresos en servicios públicos	-0.88	0.19	-0.57	0.86	0.98	1.01
4) Densidad de población	-0.08	-0.30	-0.02	-0.31	0.00	4.26
5) Hospitales con quirófano	-0.77	0.18	-0.52	-0.11	1.27	-2.71
6) Porcentaje de muertes violentas	-0.23	-0.12	0.40	0.86	-0.94	-0.55
7) Porcentaje de viviendas con agua potable	0.55	0.62	0.29	-2.05	-0.68	0.30
8) Árboles plantados	-0.48	-0.33	-0.21	-0.17	2.01	-0.16
9) Crecimiento industrial	-0.76	0.84	-0.85	-0.85	-0.05	-0.85
10) PIB	-0.73	0.59	-0.57	-0.48	0.70	3.37
11) Inversión extranjera	-0.47	1.29	-0.53	-0.44	-0.37	-0.06

Tabla 3.6 Distancias de los centroides de los clusters al gran centroide

El uso de gráficos descriptivos complementa la información de la Tabla 3.6 y facilita la identificación y perfilado de los conglomerados generados. Las Figuras 3.9 y 3.10 muestran diagramas de bloques y líneas que ayudan a la identificación visual de las diferencias entre clusters en cuanto a dos variables que son consideradas críticas: Producto Interno Bruto estatal e Inversión Extranjera. Los diagramas de la Figura 3.9 permiten apreciar que los estados en los conglomerados 1 y 3 tienen una economía deprimida puesto que la mediana de su PIB es considerablemente menor a la de los otros conglomerados de estados.

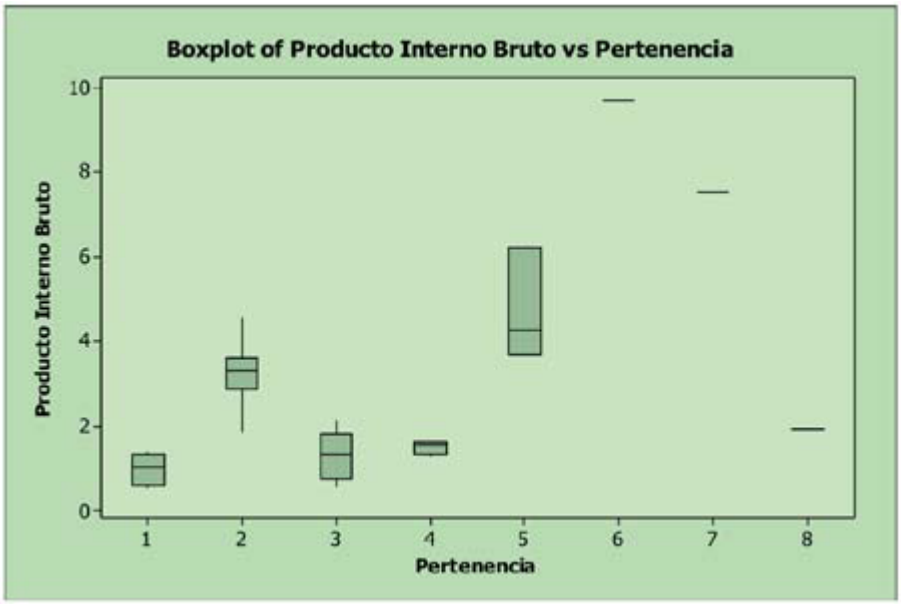


Figura 3.9 Comparación de conglomerados estatales en términos del PIB

Por otra parte, en los diagramas de la Figura 3.10 destaca el conglomerado 2, que incluye a aquellos estados en los que se concentra la mayor inversión extranjera.

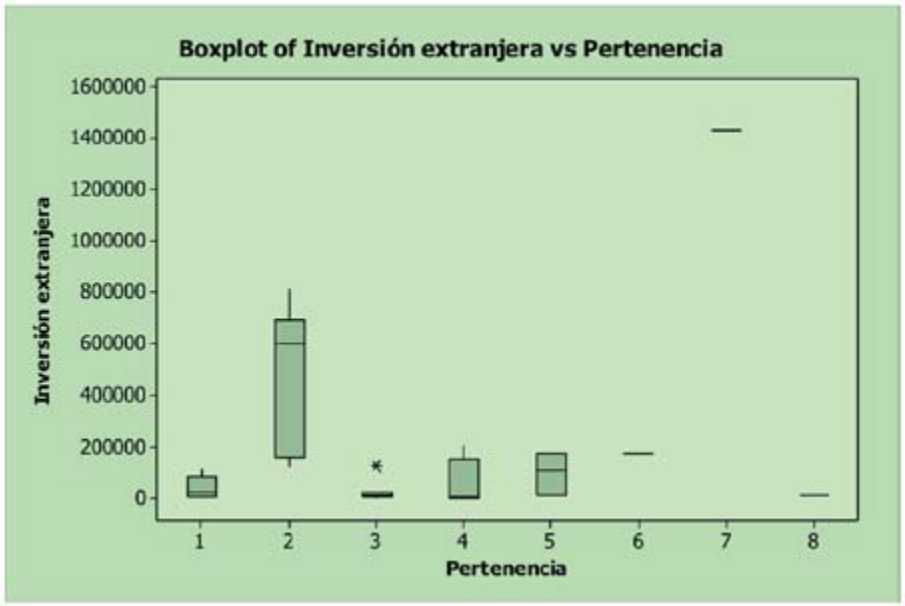


Figura 3.10 Comparación de conglomerados estatales según la inversión extranjera

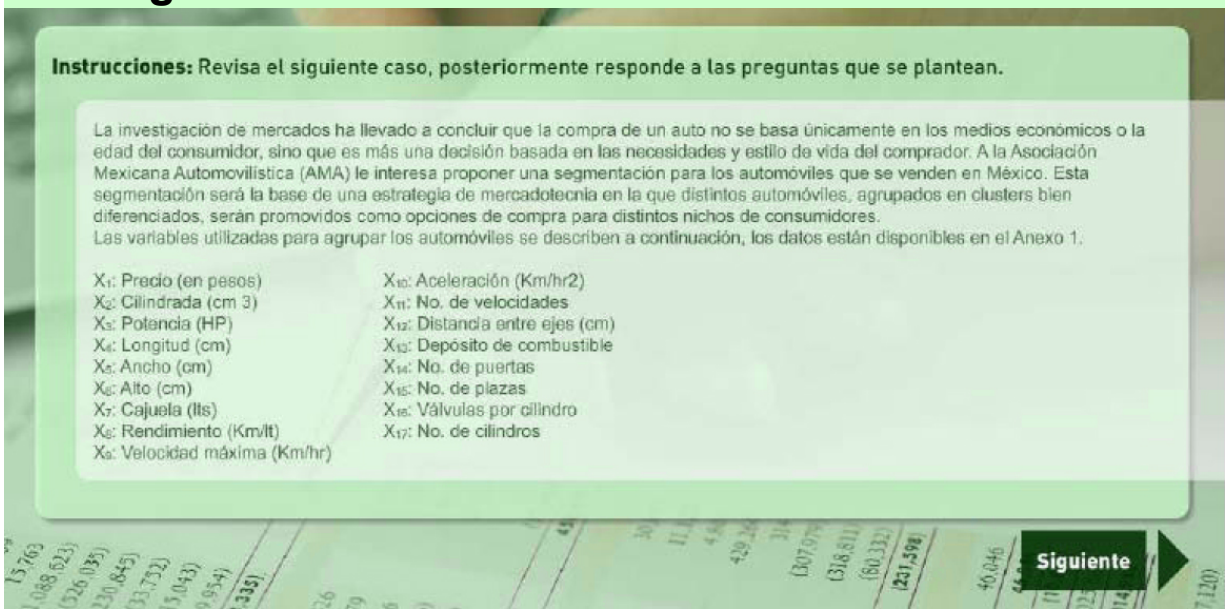
Los seis conglomerados estatales se identifican como sigue:

Baja inversión pública y extranjera	Estados con maquila	Bajo crecimiento	Bajo crecimiento y alta inseguridad	Estados inversionistas	Motor desarrollado

A partir de la segmentación post-hoc, es posible sugerir políticas y programas de desarrollo regional que fortalezcan aquellos factores rezagados (baja inversión pública y extranjera) o débiles (bajo crecimiento) de cada segmento, así como aprovechar o promover los aspectos favorables (inversión y desarrollo) que tienen algunos conglomerados para mantener o aumentar el potencial económico de los estados.

Ejercicio Integrador

3. Segmentación de consumidores a través del análisis de conglomerados



Conclusión capítulo 3

3. Segmentación de consumidores a través del análisis de conglomerados

La segmentación es importante en inteligencia de mercados para:

a. Definir estrategias de mercadotecnia dirigidas y para diseñar o modificar productos en términos del perfil y los beneficios que valoran distintos segmentos de clientes

b. Identificar segmentos meta cuyas características y necesidades pueden ser atendidas convenientemente por los productos o servicios que ofrece la empresa.

El análisis de conglomerados permite una segmentación flexible –número de segmentos y variables en las que se basa la segmentación– del mercado y facilita el perfilar los clusters resultantes. Este análisis se puede realizar empleando una amplia variedad de algoritmos; las características deseables para un algoritmo de agrupamiento son (Cruz-Hernández, 2010):

1. *Escalabilidad.* Refiere a la cantidad de datos que se pueden agrupar; los algoritmos presentados en este capítulo permiten agrupar cantidades pequeñas a moderadas de observaciones, hasta 200 aproximadamente.

2. *Habilidad para trabajar con distintos tipos de atributos,* es decir datos medidos en cualquier tipo de escala no únicamente de intervalo y razón como los algoritmos básicos que se han discutido.

3. *Manejo de conglomerados con formas arbitrarias,* esto es si el algoritmo puede identificar clusters que tienen formas difusas.

4. *Habilidad para trabajar con datos imprecisos,* se relaciona con la capacidad del algoritmo para utilizar datos con errores de medición o atípicos.

5. *Alta dimensionalidad,* esta propiedad requiere que un algoritmo pueda manejar una gran cantidad de variables por cada observación.

6. *Uso de restricciones.* Se refiere a que el algoritmo no solo construya los clusters sino que estos además satisfagan algunas restricciones en tamaño u orden.

Para satisfacer las propiedades anteriores se han diseñado algoritmos más flexibles, la mayoría de los cuales siguen utilizando el concepto de distancia para comparar individuos, sin embargo también hay métodos que siguen un enfoque multi-criterio, bajo el que se busca establecer una función de distancia basada en una estructura de preferencias entre los atributos que caracterizan a los elementos que se desea agrupar (Nemery de Bellaveux, 2006). En la literatura se reporta una gran diversidad de métodos para la formación de conglomerados que varían en términos de los principios que utilizan (Jain, 2009).

Los algoritmos de agrupamiento se pueden clasificar en cuatro tipos:

1. Basados en conceptos de distancia y similitud, los que a su vez se dividen en jerárquicos y no jerárquicos.

2. Técnicas de agrupamiento basadas en la densidad de los clusters.

3. Métodos de celdas o cuadrículas (grid).

El diagrama de la Figura 3.11 describe la variedad de métodos disponibles para realizar el agrupamiento de objetos. La selección de un método en particular requiere de considerar los objetivos del problema, las ventajas y desventajas de los varios algoritmos, los conocimientos del analista y la disponibilidad de software para su utilización.

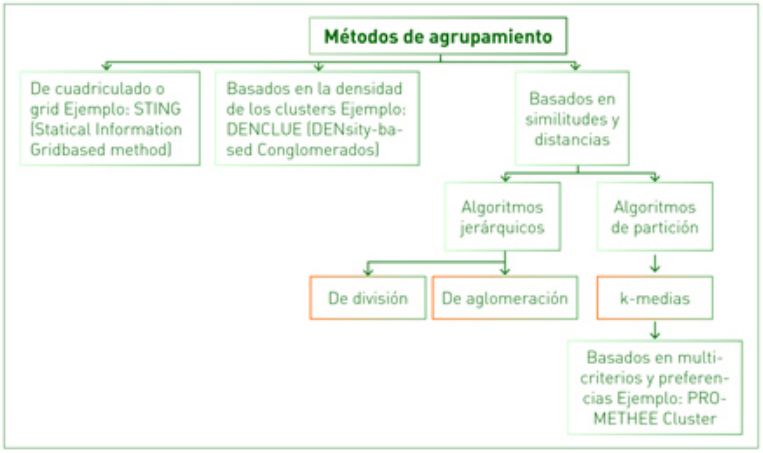


Figura 3.11 Clasificación de los algoritmos de agrupamiento

¿Sabías qué?

En muchas disciplinas del conocimiento aparte de negocios y mercadotecnia, el problema de agrupar objetos resulta de interés. En el resumen de Hartigan (1975) podrás consultar aplicaciones del análisis de conglomerados en áreas diversas como en psicología, agrupamiento de síntomas de padecimientos como paranoia y esquizofrenia antes de dar terapia, en ingeniería, agrupamiento de productos en familias que pueden fabricarse con un mismo proceso, en medicina, agrupamiento de enfermedades con síntomas similares o en arqueología, creación de taxonomías para herramientas de piedra u objetos funerarios.

Capítulo 4
Regresión lineal múltiple

Organizador temático

4. Regresión lineal múltiple

Introducción

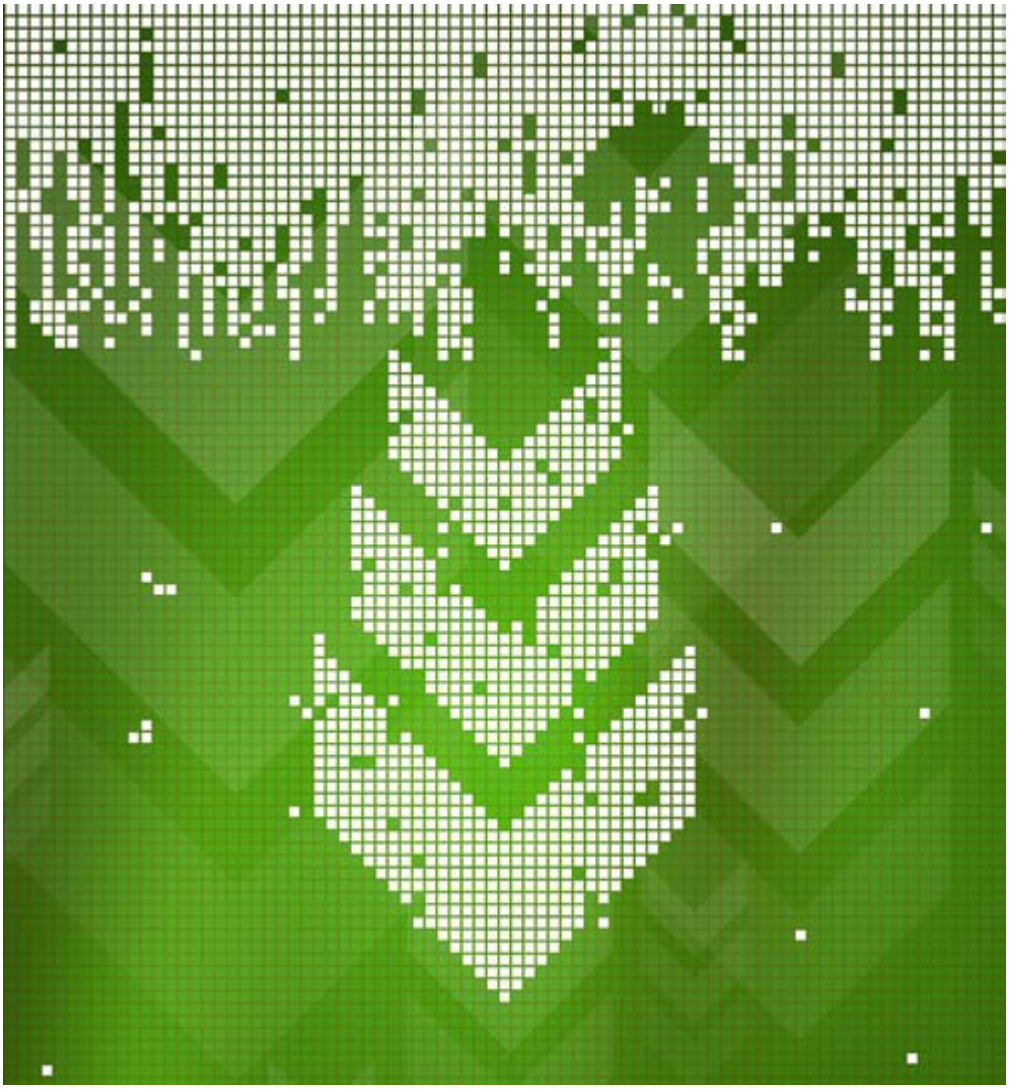
En los modelos de regresión lineal simple se analizan las relaciones que existen entre una variable dependiente y una variable

independiente, sin embargo, es común que la variable que se intenta predecir esté relacionada con más de una variable, en este caso se pueden obtener mejores resultados al predecir la variable dependiente considerando múltiples variables predictoras. El modelo de regresión lineal más sencillo es una combinación lineal de las variables predictoras, que se muestra enseguida:

$$Y_i = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1} + \varepsilon_i \qquad (1)$$

En este modelo hay una variable dependiente Y_i, $p-1$ variables predictoras X_1, X_2, …, X_{p-1}, los parámetros del modelo son β_0, β_1 , β_2 , …, β_{p-1} y ε_i es el término de error.

En dos dimensiones, es decir cuando únicamente se incluyen dos variables independientes en el modelo, el modelo anterior es un plano en el espacio tridimensional. Cuando el plano de la regresión se corta por otro plano paralelo al eje $X_1 - Y$ (ver Figura 4.1), lo que significa fijar el valor de X_2, la recta que corresponde a la intersección de los dos planos tiene una pendiente igual al coeficiente de la primera variable del modelo, esto es. Dada esta representación geométrica para el modelo de regresión lineal múltiple los coeficientes de regresión resultan ser tasas de cambio marginales o parciales, esto es, miden o representan el cambio en la respuesta cuando una de las variables independientes se incrementa en una unidad mientras que la otra variable permanece constante. Es importante tomar en cuenta esta interpretación de los coeficientes de regresión en aquellos casos en los que las variables independientes están asociadas entre ellas ya que en esta situación no es posible mantener constante una variable cuando otra se incrementa.



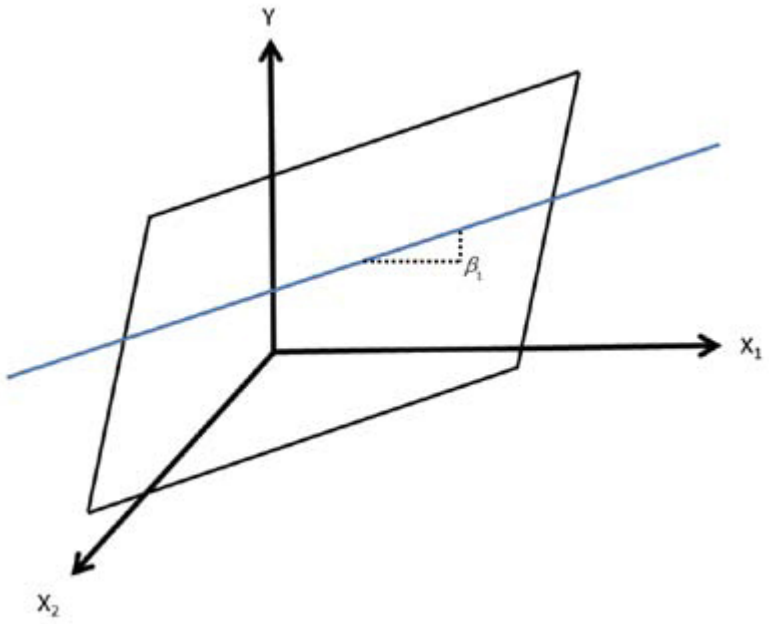


Figura 4.1 Representación geométrica del modelo de regresión múltiple.

En este capítulo de presenta primero el modelo de regresión lineal múltiple en términos matriciales, enseguida se realiza la estimación del modelo mediante el uso del método de **mmínimos cuadrados** ordinarios (OLS por sus siglas en inglés Ordinary Least Squares). Sin embargo, cabe acotar que el uso del modelo de regresión lineal múltiple no se reduce a la estimación de la ecuación de regresión o de estimados puntuales o de intervalo para la variable respuesta sino que es también relevante evaluar la calidad del modelo así como los supuestos estadísticos que soportan las inferencias estadísticas que pueden construirse para éste. Así pues, se plantea como indispensable proyectar un proceso para la construcción de un modelo de regresión lineal múltiple. En el análisis de regresión básico todas las variables tanto la dependiente como las independientes se asumen cuantitativas, esto es representadas en escalas ya sea de intervalo o razón, sin embargo es posible incluir también variables en escala nominal, en las últimas dos secciones de este capítulo se discute en detalle cómo incluir este tipo de variables en un análisis de regresión lineal múltiple.

El análisis de regresión es el método de estadística multivariante más utilizado en problemas de aplicación en distintas áreas. Kimes y

Fitzmmons (1990) utilizaron esta técnica multivariante para apoyar a los gerentes de La Quinta Motor Inn en la selección de sitios para ubicar hoteles de la cadena, este caso de aplicación se describe en el recuadro siguiente.

Uno de los elementos críticos para la rentabilidad de un hotel es el sitio donde está ubicado. La elección del lugar donde localizar un hotel requiere de considerar aquellos factores que son los generadores de la demanda; estos factores son de tipo demográfico, físico, de reconocimiento del mercado y descriptores del nivel de competencia. Con el propósito de seleccionar sitios en los cuales ubicar hoteles de la cadena La Quinta Motor Inn, se construyó un modelo de regresión lineal múltiple. La variable dependiente fue el margen de operación del hotel, el cual está directamente vinculado con la rentabilidad de la instalación hotelera. Un total de 35 indicadores que definen el atractivo del sitio y la demanda esperada fueron utilizados como variables predictoras. Por ejemplo el generador de demanda "características físicas" incluyó a los siguientes indicadores: accesibilidad del sitio, dterminado por el número de rutas de acceso, número de avenidas grandes cercanas al sitio, distancia del sitio al centro de la ciudad y visibilidad de las señales de acceso.

Para estimar el modelo de regresión se utilizaron los datos de 57 hoteles de la cadena que ya tenían cinco años en operación. Con base en las inferencias estadísticas se identificaron a las siguientes variables como los mejores predictores del margen de operación para un hotel: población en la zona donde se ubica el hotel, precio de la habitación, raíz cuadrada del ingreso medio en la zona y número de estudiantes universitarios que residen dentro de un área de cuatro millas alrededor del hotel. Para verificar que se había hecho una correcta identificación de las variables que representan a los generadores de demanda críticos, el modelo de regresión que incluyó a las variables importantes se utilizó para pronosticar si el hotel era rentable (margen de operación superior

al 35%) o no rentable (margen de operación menor de 35%). Para una muestra independiente de 94 hoteles, el modelo de regresión pronosticó correctamente el margen de operación de 93 de ellos lo que valida su calidad predictiva. La Quinta Inn adoptó este modelo y sus extensiones para predecir la rentabilidad de sitios potenciales y elegir el más conveniente para ubicar una nueva unidad de sus hoteles.

4.1 El modelo de regresión lineal múltiple

Para trabajar con el modelo de regresión lineal múltiple representado en (1) es conveniente expresarlo en forma matricial de la siguiente manera:

$$Y_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} \quad (2a) \qquad X = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{p-1,1} \\ 1 & X_{12} & X_{22} & \cdots & X_{p-1,2} \\ 1 & X_{13} & X_{23} & \cdots & X_{p-1,3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{p-1,n} \end{bmatrix} = \left[1 \mid X_1 \mid X_2 \mid \ldots X_{p-1}\right] \quad (2b)$$

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_{p-1} \end{bmatrix} \quad (2c) \qquad \underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix} \quad (2d)$$

X es una matriz de orden nxp, cada una de las columnas de esta matriz, excepto la primera, incluyen las observaciones de las p-1 variables predictoras que se utilizaran en el modelo. La primera columna esta formada por “1s” para tomar en cuenta que el modelo (1) incluye una constante β0.

Y es el vector con los datos de las variables independientes o de respuesta y es de orden nx1

β es el vector de parámetros de orden px1

$\boldsymbol{\varepsilon}$ es el vector de errores aleatorios de orden nx1

En términos matriciales el modelo de regresión lineal múltiple está dado por:

$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\ \underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon} \qquad (3)$$

El vector **ε** es un vector de n variables aleatorias que se asume tienen una distribucion normal multivariada; el vector tiene media de cero, esto es ***E(ε) = 0*** y su matriz de varianza-covarianza esta dada por:

$$\underset{n\times n}{\sigma^2(\varepsilon)} = \begin{bmatrix} \sigma^2 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \sigma^2 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I} \qquad (4)$$

Los términos en la diagonal de la matriz (4) son las varianzas constantes u homocedásticas σ^2 para cada uno de los términos del error. Los términos por arriba y por debajo de la diagonal corresponden a las covarianzas entre todas las parejas de errores. Dado que se asume normalidad para los errores y siendo todas las covarianzas de cero, significa que los términos de error son independientes entre sí. Estos son los supuestos estadísticos para el modelo de regresión lineal básico: normalidad, independencia y varianza constantes; estos supuestos también se aplican a otros métodos multivariantes de dependencia como es el Análisis de Varianza que se discute en el capítulo 6 de este eBook.

¿Sabías qué?

El trabajo de Sir Francis Galton (famoso explorador y científico inglés, 1822-1911) sobre las características que las semillas de chícharos dulces heredan de los padres derivó en la conceptualización inicial para el análisis de regresión. A pesar de

haber elegido estadísticos descriptivos de pobre calidad para describir la relación entre dos variables, Galton fue capaz de generalizar sus resultados a varios problemas sobre características genéticas hereditarias tales como habilidad artística, personalidad e incidencia de enfermedades. Lo más interesante de su trabajo fue establecer que si el grado de asociación entre dos variables se mantiene constante, la pendiente de la recta de regresión que las relaciona puede describirse cuando la variabilidad de cada variable es conocida.

4.2 Estimación del modelo de RLM con mínimos cuadrados ordinarios

Para encontrar los estimadores de los coeficientes de regresion β_0, β_1 a β_p se emplea el método de mínimos cuadrados ordinarios que considera una función Q que representa la sumatoria de las desviaciones de los valores observados Y_i de su valor esperado $\mu_{Y|X}$ en donde el subíndice Y|X hace explícito que los valores de la media de Y dependen del vector de variables X.

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1} X_{i,p-1})^2 \qquad (5)$$

Dada la función anterior, los estimadores de mínimos cuadrados se definen como aquellas cantidades que reducen la función Q. Esto mismo expresado en forma matricial se formula como sigue:

$$\min(\varepsilon'\varepsilon) = \min(\mathrm{Y} - \mathrm{X}\beta)'(\mathrm{Y} - \mathrm{X}\beta) \qquad (6)$$

Diferenciando la función (6) respecto del vector de parámetros que se desea estimar e igualando a cero para encontrar el minimo de la función se obtiene:

$$\frac{\partial(\varepsilon'\varepsilon)}{\partial\beta} = -2\mathrm{X}'\mathrm{Y} + \mathrm{X}'\mathrm{X}\,\mathrm{b} = 0 \qquad (7)$$

Expresada de otra forma:

$$X'Xb = X'Y \qquad (8)$$

Cabe notar que el vector ***β*** ha sido sustituido por el vector **b** ya que para resolver la ecuación (8) es necesario contar con datos muestrales que permitan calcular a las matrices X'X y X'Y, que tienen la siguiente estructura:

$$\underset{p\times p}{X'X} = \begin{bmatrix} n & \sum X_{1i} & \sum X_{2i} & \cdots & \sum X_{ki} \\ \sum X_{1i} & \sum X_{1i}^2 & \sum X_{1i}X_{2i} & \cdots & \sum X_{1i}X_{ki} \\ \sum X_{2i} & \sum X_{1i}X_{2i} & \sum X_{2i}^2 & \cdots & \sum X_{2i}X_{ki} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum X_{ki} & \sum X_{ki} & \sum X_{ki} & \cdots & \sum X_{ki}^2 \end{bmatrix} \qquad (9)$$

El sistema de ecuaciones dado en (8) tiene una única solución siempre que el número de datos n sea menor que el número de parámetros p en el modelo. Para obtener los estimadores de mínimos cuadrados ordinarios dados en b basta con invertir la matriz X'X para obtener (10)

$$\boldsymbol{b} = (\mathbf{X'X})\text{-}1\ \mathbf{X'Y} \qquad (10)$$

Los valores pronosticados para la respuesta Y de acuerdo con el modelo de regresión lineal múltiple quedarían representados por $\hat{\mu}_{Y|X} = E(Y \mid X) = \underset{n\times 1}{\hat{Y}} = X\,b$ donde Ŷ,Y “gorro”, se conoce como el valor ajustado de Y ya que corresponde al valor que se calcula para la respuesta a sustituir en el modelo de regresión ajustado.

Así mismo, los residuos del modelo están dados por:

$\underset{n\times 1}{\varepsilon} = Y - \hat{Y} = Y - Xb$ o en forma equivalente como $\underset{n\times 1}{\varepsilon} = (I - H)Y$ **(11)**

donde $H = X(X'X)^{-1}X'$

Finalmente la llamada matriz de varianza-covarianza para el conjunto de n errores está dada por:

$$\underset{nxn}{\sigma^2(\varepsilon)} = \sigma^2(I - H)Y \qquad \textbf{(12a)}$$

donde σ^2 es la varianza de las poblaciones de Y que están definidas para cada vector x de variables independientes, como este parámetro es usualmente desconocido, la matriz anterior se estima como sigue:

$$\underset{nxn}{s^2(\varepsilon)} = MSE\,(I - H)Y \qquad \textbf{(12b)}$$

MSE se denomina el cuadrado medio del error y es un estimado de la varianza de la variable independiente Y, este estimador se calcula como:

$$MSE = s_Y^2 = \frac{\sum_{i=1}^{n}(Y - \hat{\mu}_{Y|X})}{n-p} = \frac{\sum_{i=1}^{n}(Y - \hat{Y}_i)}{n-p} \qquad \textbf{(13e)}$$

Notar que en la expresión anterior, el estimado para la media o valor esperado de Y no es simplemente el promedio de los datos $\hat{\mu} = \overline{Y}$ como se establece cuando se calcula a la varianza muestral s^2 para una muestra univariada. En este caso la media de Y debe estimarse a partir del modelo de regresión. Dado que el modelo incluye p coeficientes–los cuales se estiman a partir de los datos disponibles- el divisor de MSE es n-p en lugar de n-1 que es el divisor utilizado en el cálculo de s^2 para el caso univariado. Esta

cantidad n-p se denominan los grados de libertad del error (gl_e), de donde MSE = SSE/gl_e.

Es importante observar que bajo los supuestos básicos impuestos al modelo de regresión lineal múltiple dado en (3) el uso de mínimos cuadrados como método de estimación para los parámetros del modelo garantiza las buenas propiedades estadísticas de los estimadores, puede demostrarse que éstos son estimadores lineales insesgados, suficientes, consistente y eficientes.

¿Sabías qué?

La forma más temprana de regresión fue el método de los mínimos cuadrados que fue publicado por Legendre en 1805 y Gauss en 1809 quienes lo aplicaron en la astronomía. Si se utilizan métodos estadísticos formales de estimación–por ejemplo máxima verosimilitud para obtener los estimadores de un modelo de regresión lineal, estos estimadores coinciden con aquellos derivados por el simple método algebraico-geométrico de mínimos cuadrados ordinarios cuando los supuestos de normalidad, independencia y variables independientes medidas sin error son válidos.

Ejemplo 1. Un analista desea elaborar un modelo que ayude a los pequeños negocios en la Ciudad de México a comprender mejor los factores que afectan sus ventas, expresadas en miles de pesos. El analista consideró que la situación de los pequeños negocios depende fundamentalmente del porcentaje del personal contratado, sin considerar a los profesionistas que trabajan por su cuenta, y del índice de remuneración obtenida por los empleados con respecto a la base del promedio de pago del año.

Para obtener el vector **b** de los estimadores de mínimos cuadrados se calcula primero la matriz **X'X** y se calcula su matriz

inversa, luego se procede a calcular el vector **X'Y** y se realiza la multiplicación como se indica:

$$\begin{bmatrix} b_o \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 263.778 & -2.755 & 0.022 \\ -2.755 & 0.029 & -0.001 \\ 0.022 & -0.001 & 0.001 \end{bmatrix} \begin{bmatrix} 1385 \\ 133534.6 \\ 115036.4 \end{bmatrix} = \begin{bmatrix} -121.389 \\ 1.326 \\ 0.917 \end{bmatrix}$$

El modelo de regresión lineal múltiple estimado por el analista es:

$$\hat{Y} = -121.389 + 1.326X_1 + 0.917X_2$$

El analista realiza un pronóstico puntual para una pequeña empresa o negocio bajo las condiciones de X1 = 90 % del personal contratado y un índice de remuneraciones con un crecimiento respecto al año anterior de 6% (X2 = 106% en total). Si estos valores se sustituyen en el modelo de regresión estimado, se encuentra que las ventas serían de 95.2 millones de pesos.

4.3 Inferencias para el modelo de regresión lineal múltiple

4.3.1 Análisis de varianza para la regresión y coeficiente de determinación

La hipótesis de mayor interés en el análisis de regresión es la asociada con la significancia del modelo de regresión múltiple propuesto. Esto implica probar la hipótesis nula de que ninguno de los predictores provoca un cambio significativo en la respuesta

contra la alternativa de que hay al menos una variable que induce cambios en la respuesta. Esta alternativa significa que conque una única variable independiente dé lugar a un cambio, es suficiente para rechazar la hipótesis nula y concluir que el modelo de regresión es útil o relevante para explicar los cambios observados en la respuesta de interés. De acuerdo a lo anterior se plantean las siguientes hipótesis:

$$H_O : \beta_1 = \beta_2 = \ldots \beta_{p-1} = 0$$
$$H_a : Al\ menos\ un\ \beta_k \left(k = 1 \ldots p-1\right)\ es\ diferente\ de\ cero$$

Estas hipótesis se prueban mediante un análisis de varianza. La tabla ANOVA (por sus siglas en inglés ANalysis Of VAriance) correspondiente a este análisis se calcula como se muestra a continuación:

Fuente de Variación	SS	gl	MS
Regresión	$SSR = \mathbf{b'X'Y} - \left(\frac{1}{n}\right)\mathbf{Y'JY}$	$p-1$	$MSR = \frac{SSR}{p-1}$
Error	$SSE = \mathbf{Y'Y} - \mathbf{b'X'Y}$	$n-p$	$MSE = \frac{SSE}{n-p}$
Total	$SSTO = \mathbf{Y'Y} - \left(\frac{1}{n}\right)\mathbf{Y'JY}$	$n-1$	

Tabla 4.1 Tabla ANOVA para el Modelo de Regresión Lineal Múltiple

La hipótesis nula formulada previamente se prueba a través del estadístico de prueba dado en (14) el cual, bajo los supuestos estadísticos impuestos al modelo, sigue la distribución F con (p-1, n-p) grados de libertad.

$$F = \frac{MSR}{MSE} \qquad (14)$$

¿Sabías qué?

La presencia de observaciones extremas puede afectar considerablemente a los estimados de los coeficientes de regresión cuando se utilizan mínimos cuadrados ordinarios para estimarlos. Para ajustar modelos de regresión a partir de datos que están contaminados con observaciones extremas se cuenta con la regresión robusta. Hay muchos métodos disponibles para realizar una regresión robusta, el más popular es la estimación M de Huber. El software estadístico SAS/STAT tiene disponible la opción para realizar regresión robusta, aprende al respecto consultando el artículo de Chen (s.f.).

La regla de decisión asociada con la prueba de hipótesis que se formuló previamente es:

$$Si\ F \leq F\left(1-\alpha; p-1, n-p\right)\ se\ concluye\ H_o$$

$$Si\ F > F\left(1-\alpha; p-1, n-p\right)\ se\ concluye\ H_a$$

Si H_0 es no es rechazada, se debe de revisar y modificar el modelo propuesto, ya sea definiendo nuevas variables de acuerdo a la teoría del área que motivó el problema o bien estimando un modelo más complicado que la simple combinación lineal de las variables independientes.

Si el modelo de regresión lineal múltiple es significante, lo que indica que se rechazó la hipótesis nula, una pregunta adicional de interés es cuantificar hasta dónde el modelo explica los cambios o variabilidad observada en la respuesta. Para responder a esta pregunta se define al coeficiente de determinación múltiple R^2

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \qquad (15)$$

Este coeficiente se interpreta como la variabilidad de Y explicada a través del modelo de regresión múltiple. Los valores extremos de este coeficiente se alcanzan cuando *SSE* = 0, lo que implica que R^2 = 1 y que toda variación de la respuesta puede atribuirse a la relación funcional entre ésta y las variables independientes consideradas. En el caso que *SSE* = *SSTO* el total de la variabilidad observada de la respuesta sería atribuida a causas aleatorias y el valor de R^2 sería igual a cero. Es importante indicar que cuando la regresión es múltiple, un incremento en el número de variables independientes siempre incrementará *SSR* y por lo tanto el valor de R^2, aun cuando la contribución de algunas de estas variables no sea estadísticamente significante. Para corregir esta situación se define al coeficiente de determinación ajustado como:

$$R_a^2 = 1 - \frac{[(n-1)SSE]}{[(n-p)SSTO]} \qquad (16)$$

Este coeficiente penaliza el valor de R^2 por el número de variables incluidas en el modelo. Si la introducción de una variable extra no reduce en una cantidad considerable la variación no explicada de *Y*, entonces R^2_a puede llegar incluso ser menor a R^2.

Ejemplo 2. Una tienda de departamentos desea mejorar el servicio que da sus clientes en relación a los tiempos totales, de espera y de servicio, para pago en cajas. La empresa se ha propuesto mejorar los tiempos de servicio en cajas a través del manejo de cajas rápidas, capacitación a cajeros y redefinición de códigos para productos a granel. El gerente de servicios al cliente decidió recopilar información respecto a las variables que considera influyen sobre los tiempos de espera al hacer el pago a fin de decidir en qué forma se podrían operar mejor las cajas rápidas. Una muestra aleatoria de 50 clientes dio información sobre las siguientes variables:

Y = Tiempo que tarda el cliente, en fila y en servicio, en completar su pago en caja

X1= Artículos que requieren pesarse porque se venden a granel

X2 = Número total de artículos comprados

X3 = Pago con tarjeta de crédito (0 = no, 1 = sí)

Se ajustó un modelo de regresión lineal múltiple con los datos y parte del listado se da a continuación. En la matriz de correlación no se incluye a la variable X3, ya que la variable está en una escala nominal y por lo tanto el **coeficiente de correlación** de Pearson es una medida inválida para describir la asociación que guarda con otras variables.

	tiempo	*artículos a granel*	*total de artículos*
Tiempo	1		
artículos a granel	0.576109252	1	
total de artículos	0.640106566	0.663200154	1

Tabla 4.2 Matriz de Correlación Ejemplo 2

	gl	*Suma de cuadrados*	*Cuadrados medios*	*F*
Regresión	3	120.9188137	40.3063	
Error	46			
Total	49	140.0177778		

Variable	*Coeficiente*	*Error estándar*
Intercepción	-1.58401117	0.726408882
Artículos a granel	0.413060294	0.082583473
Total de artículos en la compra	0.025011116	0.029885751
Pago con tarjeta	1.801809935	0.593500563

Tabla 4.3 Tabla ANOVA Ejemplo 2

Completa la tabla ANOVA y prueba la hipótesis asociada al análisis. Calcula el coeficiente de determinación ajustado. Establece conclusiones

4.3.2 Inferencias individuales para los parámetros del modelo RLM

Para construir inferencias individuales en los parámetros del modelo de regresión lineal múltiple es necesario conocer las distribuciones de muestreo de los estimadores de mínimos cuadrados. Anteriormente se estableció que los estimadores de mínimos cuadrados son insesgados, esto significa que el valor esperado de dichos estimadores es igual a los verdaderos coeficientes del modelo de regresión (3) es decir **E(b) = *β***.

La matriz de varianza-covarianza para el vector de estimadores b está dada por:

$$\underset{pxp}{\boldsymbol{\sigma}^2(\mathbf{b})} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \qquad (17)$$

puesto que MSE = $\hat{s}_Y^2$, la matriz anterior se estima como

$$\underset{pxp}{\mathbf{s}^2(\mathbf{b})} = MSE(\mathbf{X}'\mathbf{X})^{-1} \qquad (18)$$

Los elementos en la diagonal de la matriz (18) son las varianzas estimadas para cada uno de los p coeficientes de regresión, esto es $s^2[b_o]$, $s^2[b_1]$, etcétera. Estas cantidades son necesarias para probar hipótesis individuales para los coeficientes del modelo de regresión así como para construir estimados por intervalos de confianza.

El problema de inferencia estadística que involucra la prueba de hipótesis de que el coeficiente de regresión parcial de la k-ésima

variable del modelo de regresión es cero se formula como sigue:

$$H_O : \beta_k = 0$$
$$H_a : \beta_k \neq 0$$

El estadístico apropiado bajo los supuestos de normalidad, independencia y **homocedasticidad**, resulta ser una t-Student que se calcula como en (19)

$$\frac{b_k - \beta_k}{s[b_k]} \approx t(n-p) \quad para\ k = 0, 1, \ldots, p-1 \tag{19}$$

Para el caso particular de la prueba de hipótesis de que el coeficiente es igual a cero, el estadístico de prueba y la región de rechazo correspondiente resultan ser:

$$t = \frac{b_k}{s[b_k]}$$

$$RR = \{|t| > t(\alpha/2, n-p)\}$$

Los resultados de las pruebas de hipótesis individuales para los coeficientes de regresión permiten responder a la pregunta ¿cuáles predictores tienen un efecto significativo en la respuesta? Es decir, que las hipótesis individuales complementan el ANOVA al permitir identificar aquellas variables que son críticas para explicar los cambios observados en la variable independiente Y.

Además de pruebas de hipótesis individuales, también es posible construir estimados por intervalo de confianza para los coeficientes

de regresión. El intervalo de confianza individual al (1-α) x100 para el k-ésimo coeficiente parcial se calcula como:

$$b_k \pm t(1-\alpha/2; n-p)s\left[b_k\right] \quad (20)$$

4.3.3 Intervalos de confianza para la respuesta media e intervalos de pronóstico

Una vez que se ha seleccionado un modelo de regresión apropiado para describir el comportamiento de una variable independiente de interés, el modelo puede utilizarse con fines de predicción. Hay dos inferencias estadísticas que resultan de interés: 1) la construcción de intervalos de confianza para la respuesta media o valor esperado de Y para valores dados de las variables independientes y 2) el cálculo de intervalos de pronóstico para un valor único de Y.

Es importante destacar la ventaja de la estimación por intervalos de confianza sobre la estimación puntual; para la primera se tiene asociada una declaración de probabilidad que garantiza la calidad de la inferencia además de establecerse un rango posible de valores para el verdadero valor del parámetro o del pronóstico.

Para construir las inferencias anteriores, se definen valores específicos para cada una de las variables independientes, estos valores integran el vector X_0 expresado en (21), notar que la primera entrada del vector es “1” ya que el modelo de regresión incluye al término (constante) b_0.

$$\underset{p\times 1}{\mathbf{X}_0} = \begin{bmatrix} 1 \\ X_{h1} \\ \cdot \\ \cdot \\ \cdot \\ X_{h,p-1} \end{bmatrix} \qquad (21)$$

El estimador puntual para el valor esperado o la media de la variable dependiente para cualquier conjunto de valores específicos de las variables independientes se determina a partir del modelo de regresión estimado como se indica en seguida:

$$\hat{Y}_0 = \hat{\mu}_{Y.X_0} = \mathbf{b}'\mathbf{X}_0 = b_0(1) + b_1 X_{10} + \cdots + b_k X_{k0} \qquad (22)$$

Puede probarse (Searle, 1997) que la varianza de este estimador es igual a:

$$\mathrm{Var}\left(\hat{Y}\right) = \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\sigma^2 \qquad (23)$$

En la expresión (23), σ^2 es la varianza de Y la cual, según se indicó previamente, se estima a través del cuadrado medio del error MSE. De donde la varianza muestral para el valor estimado de la respuesta $\hat{Y} = \hat{\mu}_{Y|X}$ es igual a

$$s^2\left(\hat{Y}\right) = \left[\mathrm{MSE}\,\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\right] = \mathbf{X}_0'\mathbf{H}\mathbf{X}_0 \qquad (24)$$

Bajo los supuestos de la regresión, $\hat{Y}$ tiene una distribución de muestreo normal y es un estimador insesgado para la respuesta media o valor esperado de Y cuando $X = X_0$. De estos resultados se deriva el siguiente intervalo de confianza para $\mu_{Y\text{-}X_0}$ esto es la media o valor esperado de Y cuando el vector $X = X_0$ dado en (21):

$$\hat{Y} - t\left(\alpha/2, n-k-1\right)s\left(\hat{Y}\right) < \mu_{Y.X_0} < \hat{Y} + t\left(\alpha/2, n-k-1\right)s\left(\hat{Y}\right) \quad (25)$$

Si se desea establecer cuál será no el valor medio de la respuesta sino un valor individual dados valores específicos de los predictores, debe considerarse que la varianza de un pronóstico está formada por dos componentes: la variabilidad asociada con la estimación de la respuesta media y la variabilidad de Y alrededor de su media, de donde la varianza de un pronóstico Y_p está dada por:

$$\mathrm{Var}\left(Y_p\right) = \mathrm{Var}\left(\hat{Y}\right) + \mathrm{Var}(Y) \quad (26)$$

Puesto que la verdadera varianza de la media no es usualmente conocida sino estimada según (24), se obtiene la siguiente expresión para estimar la varianza de un pronóstico

$$s^2\left(Y_p\right) = \left[\mathrm{MSE}\,\mathbf{X}_0'(\mathbf{X}\mathbf{X})^{-1}\mathbf{X}_0 + 1\right] \quad (27)$$

El intervalo de pronóstico siempre será más amplio que el intervalo de confianza ya que hay mayor incertidumbre en establecer que ocurrirá en único ensayo o para un valor individual de Y que en establecer cuál será el valor medio o esperado de esta variable independiente. El ancho de los intervalos de confianza y pronóstico también se incrementará entre más alejado esté el vector x_0 del centroide de las variables independientes, esto es del punto $\left(\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots, \bar{X}_k\right)$.

Ejemplo 3. El dueño de una distribuidora de automóviles realizó un estudio con el propósito de pronosticar su volumen de ventas en términos de las siguientes variables predictoras:

Y = Número de automóviles vendidos en el mes

X1 = Número de modelos en exhibición durante el mes

X2 = Número de vendedores en la distribuidora

X3 = Tasa de interés aplicable en la compras a crédito (en porcentaje)

Empleando los datos de los últimos 25 meses de ventas se ajustó un modelo de regresión lineal múltiple. El listado parcial de computadora se da a continuación:

ANOVA

	SS	df	MS	F
Regression	19.972	3	6.657	11.76
Residual	11.888	21	0.566	
TOTAL	31.860	24		

R-squared: 0.6269

Table of estimates

Variable	Coefficient	Stdev	t-test
Constatnt	14.1411	0.3286	43.03
X1	0.15106	0.0614	2.46
X2	1.17396	0.5023	2.34
X3	-1.3971	0.3191	-4.38

Es de interés para la distribuidora determinar un intervalo de confianza y uno de pronósticos, al 95% ambos, para el Número de automóviles vendidos en un mes en que se tuvieron 5 modelos en exhibición, 10 vendedores y se aplicó una tasa de interés del 15% para compras a crédito ya que estas son las condiciones usuales bajo las cuales opera la agencia.

Sustituyendo los valores X1 =5, X2 =10 y X3 = 15 se calcula un estimado puntual para el valor esperado de Y el cual también corresponde al mejor pronóstico puntual, y que en este caso es igual a:

$$\hat{Y}_0 = \hat{\mu}_{Y|X_0} = 14.1411 + 0.1511(5) + 1.1740(10) - 1.3971(15) = 5.6801$$

El estimado para la varianza de la respuesta media $\hat{\mu}_{Y|X}$ se calcularía como sigue:

$$s^2\left(\hat{Y}\right) = \begin{bmatrix}1 & 5 & 10 & 15\end{bmatrix}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0 \mathrm{MSE}$$

Es necesario contar con los datos originales para determinar la matriz $(X'X)^{-1}$ para realizar el cálculo. Afortunadamente hay software

estadístico disponible, incluso se tiene la opción de realizar análisis de regresión en Excel, para calcular todas estas cantidades y facilitar al analista el trabajo numérico. Empleando estos recursos computacionales se calcula la varianza muestral para el valor ajustado como:

$$s^2\left(\hat{Y}\right) = \begin{bmatrix} -0.04 & 0.046 & 0.04 & 0.0375 \end{bmatrix} \mathbf{X}_0 \, (0.566) = 0.6523$$

De donde los intervalos requeridos, ambos al 95 % de nivel de confianza son:

$$\hat{Y} \pm t(0.025, 21)s\left(\hat{Y}\right) = 5.6801 + 2.080(0.6523)^{1/2} = 4.00 < \mu_{Y.X_0} < 7.36 \text{ autos/mes}$$

$$\hat{Y} \pm t(0.025, 21)s(Y_p) = 5.6801 \pm 2.080(0.6523 + MSE)^{1/2} = 3.38 < Y < 7.98 \; autos / mes$$

El primer intervalo significa que de mantenerse por un largo periodo las condiciones de operación de 5 modelos en exhibición, 10 vendedores y una tasa de interés del 15% el promedio de ventas mensuales estaría entre 4 y 7.36 autos. Mientras que para el intervalo de pronóstico, significa que si en un mes en particular la agencia operara con estas condiciones, sus ventas estarían entre tres (3.38) y ocho (7.98) autos.

Actividad de repaso

4. Regresión lineal múltiple

Instrucciones: con referencia al ejemplo 2 con Y = tiempo total para pagar en caja en una tienda de autoservicio, realizar lo siguiente:

a) Calcular un intervalo de confianza para el incremento promedio en el tiempo de atención del cliente si compra un artículo adicional

b) ¿Hay un incremento significativo en el tiempo de atención para el cliente cuando este paga con tarjeta de crédito respecto a cuándo lo hace en efectivo o cheque?

c) Calcular un intervalo de confianza para el tiempo promedio de atención de los clientes en las cajas rápidas que piensan instalarse. Para estas cajas se manejará únicamente el pago en efectivo, un máximo de 10 artículos y cuando mucho 5 de ellos artículos de venta a granel. Asuma que la varianza de la media estimada para estas condiciones es de 0.524.

Respuestas

a) El coeficiente de la variable “número de artículos adquiridos” estima el cambio, incremento en este caso ya que el coeficiente es positivo, en el tiempo de atención cuando esta variable se incrementa en una unidad y las otras dos permanecen fijas. Por lo que el IC solicitado está dado por la expresión siguiente. El valor de t hay que determinarlo de tablas y es igual a 1.96 para un nivel de confianza del 95% que es el usual para este tipo de inferencia:

bi± t(1-α/2; gle) s(bi): 0.025011 ± 1.96 (0.029885751): (0.033565; 0.083587)

b) El coeficiente de la variable “pago con tarjeta” es positivo (1.80181) lo que indica que si esta variable X3 va de X3 = 0 (pago

contado) a X3 = 1 (pago tarjeta), hay un incremento de 1.80181 en el tiempo de servicio. Sin embargo, como esta cantidad es sólo un estimado del verdadero cambio que registra la respuesta cuando la variable se modifica es necesario probar formalmente la siguiente hipótesis:

Ho: $\beta 3 = 0$ versus H1: $\beta 3 > 0$

La región de rechazo para un nivel de significancia del 5 % y los 46 gl asociados a MSE es la siguiente:

$RR = \{t > t(\alpha; gle) = 1.645\}$

Usando como estadístico de prueba t = bi / s(bi) = 1.80181 / 0.5935 = 3.036 se rechaza la hipótesis nula y se concluye que cuando se paga con tarjeta (estado 1 de la variable X3) respecto a cuándo se paga en efectivo (estado 0 de la variable X3) hay un incremento significativo en los tiempos de atención al cliente, estimado en 1.8 minutos.

c) Se requiere un intervalo de confianza para E(Y | X1 = 5, X2 = 10, X3 = 0), el estimador puntual es igual a: -1.5840117 + 0.413060294 (5) + 0.025011116 (10) + 1.80181 (0)

De donde el intervalo de confianza al 95% dada la varianza proporcionada para el anterior estimado puntual es igual a:

$0.731401 \pm t(1-\alpha/2; gle)\ (0.524)$ ½ : de cero a máximo 2.1502 min

El límite inferior del intervalo es negativo, pero dada la naturaleza de la variable de respuesta (tiempo) su menor valor posible es cero, en consecuencia lo que puede establecerse es que aquellos clientes que pagan 10 artículos, 5 de ellos de compra a granel y pagan en efectivo, tardarán no más de 2.15 minutos en promedio para completar su pago en las cajas.

4.4 Construcción del modelo de regresión lineal múltiple

El análisis de regresión no se limita a la estimación de la ecuación de regresión y al cálculo de estimados puntuales para la respuesta de interés al sustituir en el modelo. Un modelo apropiado debe ser significante, esto es contribuir a explicar una porción relevante de la variabilidad de *Y*, incluir únicamente a aquellas variables que son buenos predictores y satisfacer los supuestos estadísticos que garantizan la calidad de las inferencias que de él se derivan. En la Figura 4.2, se ha construido un diagrama de flujo que muestra las etapas de construcción y evaluación del modelo de regresión múltiple antes de que éste sea considerado apropiado para aplicaciones administrativas.

La primera etapa consiste en la cuidadosa selección de las variables a utilizar como predictores, la definición de medidas convenientes para cada una de ellas y la recopilación de datos suficientes para estimar el modelo propuesto. En esta primera etapa, es importante recordar que la elección de variables independientes ha de estar sustentada en el cuerpo de conocimientos del área específica a la que pertenece el problema de pronósticos que se atiende. Por ejemplo, el problema de definir aquellas variables que son determinantes críticos para la satisfacción en el trabajo tiene que considerar los aspectos de la tarea, el ambiente de trabajo y las retribuciones económicas y sociales; mientras que un análisis de regresión para predecir la satisfacción con un servicio debe considerar como variables independientes a la capacidad técnica de los empleados, las garantías ofrecidas por la empresa y la disposición para atender las necesidades del cliente, entre otras.

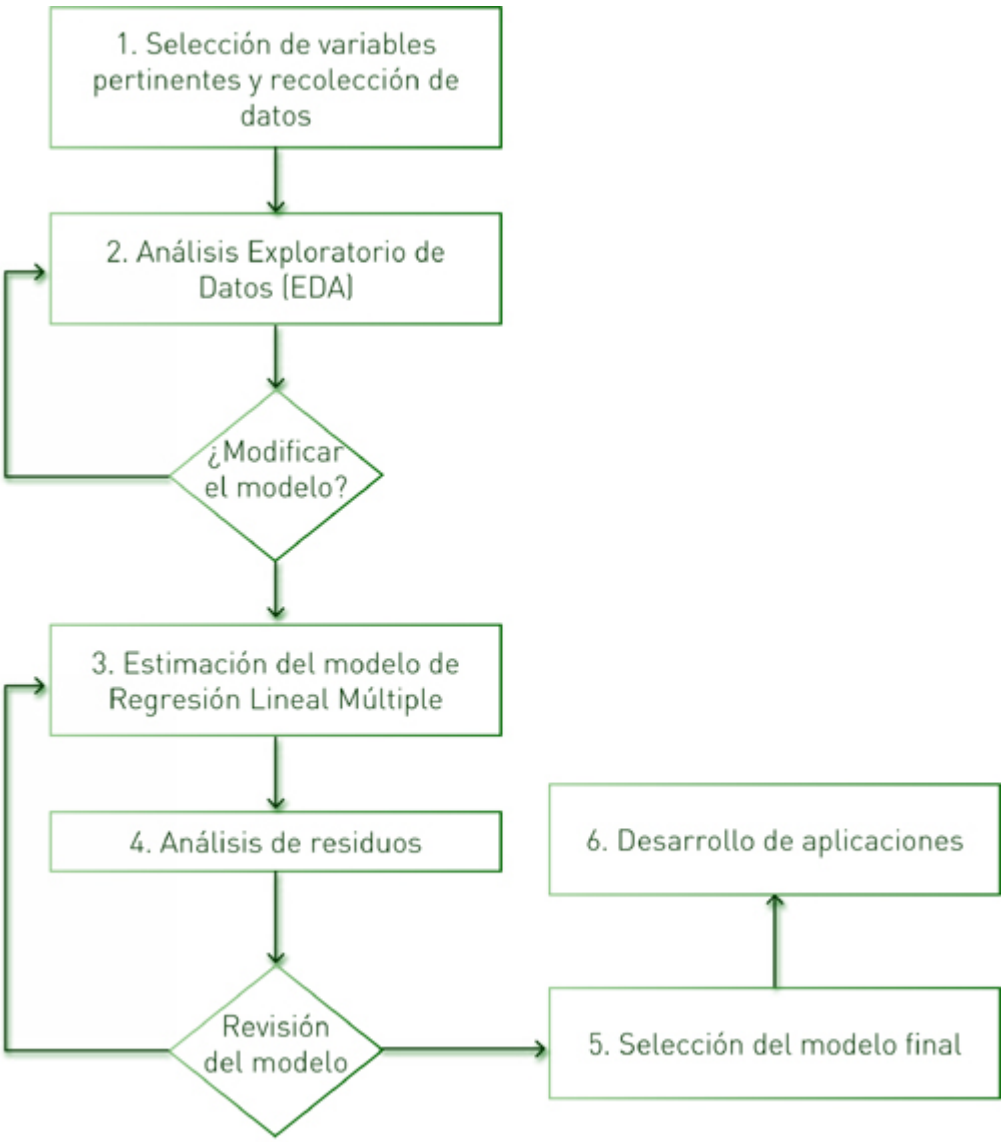


Figura 4.2 Proceso para la construcción del Modelo de regresión Lineal múltiple.

La segunda etapa consiste en realizar un Análisis Exploratorio a los Datos (EDA), término introducido por Tukey (1977) para denominar aquellos métodos de análisis, principalmente de tipo gráfico, que se utilizan para explorar las características de los datos y sugerir patrones de comportamiento que posteriormente puedan examinarse más a detalle con métodos de inferencia estadística.

En el sentido de análisis de regresión múltiple, este análisis se enfoca a explorar los siguientes aspectos:

a. Los supuestos básicos del modelo de regresión dado en (3), esto es independencia, normalidad y heterocedasticidad,

b. La linealidad de las relaciones entre las variables independientes y la dependiente,

c. La identificación de datos extremos, y

d. El "cernido" inicial de variables independientes consideradas con el fin de descartar aquellas que no están estadísticamente asociadas con la respuesta. Gráficos de cuartiles, histogramas, diagramas de correlación y el cálculo de coeficientes de correlación simples son los recursos para la exploración en esta etapa.

El resultado de este paso es la propuesta de uno o varios modelos de regresión múltiple tentativos. Estos modelos pueden incorporar el uso de transformaciones con los objetivos de (1) satisfacer los supuestos del modelo y (2) inducir linealidad en las relaciones entre Y y las variables del vector $\boldsymbol{x}$.

La tercera etapa del proceso es la estimación de los modelos propuestos mediante mínimos cuadrados ordinarios. Cada modelo es, entonces, revisado en una cuarta etapa para identificar desviaciones a los supuestos y posibles problemas debido a la existencia de patrones de asociación entre el conjunto de variables independientes (ver siguiente sección para la descripción y detección de este problema). Los modelos revisados son comparados entre sí para seleccionar finalmente aquel que satisface tanto los criterios de calidad estadística como de habilidad predictiva. Este modelo se usa para hacer aplicaciones de interés.

Según describe el diagrama de la Figura 4.2, una vez que los datos están disponibles, es necesario especificar la función f (.) que relaciona el valor medio de Y con los predictores, esto es $Y = f(\boldsymbol{x}) + \varepsilon$. Es importante reconocer que el modelo de regresión lineal múltiple (1) asume que f (.) es una función lineal, si éste no es el caso, el análisis de regresión no proporcionará resultados apropiados. Cuando el analista prefiere no especificar a la función f

(.) sino estimarla directamente de los datos, lo que procede es utilizar regresión no-paramétrica, este tipo de análisis es más complicado pero también menos restrictivo ya que no impone una relación funcional entre Y y el conjunto de variables independientes. Entre los varios métodos para estimar modelos de regresión no-paramétricos están los polinomios locales ponderados, la suavización con splines y los métodos de núcleos, kernel, (Fox, 2002).

El análisis de correlación es parte de la exploración de los datos y sus objetivos son: identificar aquellas variables estadísticamente relacionadas con la respuesta y que por lo tanto conviene incorporar al modelo de regresión; detectar observaciones extremas; identificar la presencia de relaciones no lineales entre el conjunto de *Xs* y *Y*, y posibles problemas de interpretación en los coeficientes de regresión debido a la fuerte asociación entre las variables predictoras. Pare realizar este análisis exploratorio se recurre a la construcción de la llamada matriz de correlación, *R*.

$$R = \begin{bmatrix} 1 & r_{y1} & r_{y2} \\ r_{y1} & 1 & r_{12} \\ r_{y2} & r_{12} & 1 \end{bmatrix} \qquad (28)$$

Para analizar la matriz (28), conviene considerar dos bloques: 1) el primer renglón o columna que incluye todos los coeficientes de correlación simples entre respuesta y predictores; y 2) la submatriz de coeficientes de correlación simples entre parejas de predictores.

En relación al bloque (1) los coeficientes de correlación r_{ij} miden la asociación entre la i-ésima variable en el modelo y la respuesta e ignoran todas las otras variables independientes en el modelo. El cuadrado de este coeficiente, es igual al coeficiente de determinación de la regresión lineal simple de Y con la i-ésima variable independiente. Por lo que, si el coeficiente de correlación simple se declara igual a cero, la conclusión será que la i-ésima

variable independiente no guarda una relación lineal significante con la respuesta ó explica una porción insignificante de la variación de Y, en consecuencia es apropiado eliminar esta variable del modelo de regresión. Por otra parte, cuando el coeficiente de correlación es significante la variable independiente en cuestión debería ser incluida en el modelo de regresión. Si el proyecto de investigación es de tipo de descriptivo, el análisis de los coeficientes de correlación es suficiente para proporcionar evidencia empírica a favor de la existencia de relaciones entre las variables bajo estudio. Se entiende por estudio descriptivo aquel proyecto de investigación que se limita a describir las asociaciones entre variables, sin que se asuma que estas asociaciones son del tipo causa-efecto como para proponer un modelo de dependencia que además permita controlar o predecir la respuesta de interés.

El segundo bloque de la matriz R, que es importante analizar, incluye a todos los coeficientes de correlacion identificados como r_{ij}. Estos coeficientes miden la asociacion que hay entre las variables independientes X_i y X_j , para toda $i \neq j$.Cuando dos variables independientes estan correlacionadas entre ellas (su coeficiente de correlacion se declara significante) se dice que hay multicolinealidad. Cuando la multicolinealidad es muy alta surgen problemas importantes en la interpretacion y calidad de los coeficientes del modelo de regresion. Los problemas asociados a una severa multicolinealidad son los siguientes (Hair et al., 1999, Montgomery, 2001, Neter et al., 1996, y Weisberg, 1985):

1. Cambios sustanciales en los estimados de los coeficientes de regresión dependiendo de cuáles variables se consideran en el modelo. Esta situación dificulta la descripción de la influencia individual que cada variable tiene sobre la respuesta.

2. Las desviaciones estándar de los coeficientes se incrementan considerablemente. Esto implica que los estimadores de mínimos cuadrados ordinarios varían notablemente de muestra a muestra, llegando incluso a cambiar de signo, lo que deriva nuevamente en la imposibilidad de interpretar la influencia relativa de cada predictor. Cabe aclarar que la severa multicolinealidad incrementa o “infla” la

varianza de los estimadores OLS, aún cuando no anula sus propiedades estadísticas de ser insesgados.

3. Los coeficientes de regresión parciales de aquellas variables independientes correlacionadas entre sí pueden declararse estadísticamente iguales a cero aun cuando los predictores estén significativamente correlacionados con la respuesta. además de la reiterada pérdida de interpretación, esta problemática puede derivar en que variables importantes que explican o predicen la respuesta sean eliminadas del modelo.

4. Los pronósticos y la media estimada para la respuesta no resultan afectados por la multicolinealidad, siempre y cuando las inferencias se hagan dentro de la región de estudio donde se tienen datos para la regresión. Si se extrapola fuera de este rango de estudio, el riesgo de que los pronósticos se desvíen de forma considerable de los valores reales se incrementa cuando hay multicolinealidad.

La presencia de coeficientes de correlación simples r_{ij} de gran magnitud (cercanos a ± 1) permite detectar severa multicolinealidad antes de ajustar el modelo de regresión. Si una vez ajustado el modelo, los coeficientes de regresión parciales de las variables con gran r_{ij} resultan no-significantes, estadísticamente cero, se tiene evidencia adicional de problemas de interpretación asociados con la multicolinealidad.

Un método formal para detectar la presencia de multicolinealidad, no sólo entre parejas de variables sino entre grupos de variables, es el cálculo de los llamados factores inflacionarios de la varianza, VIF por sus siglas en inglés Variance Inflationary Factors. El VIF asociado al coeficiente de regresión de la j-ésima variable independiente se define como

$$VIF = \left(\frac{1}{1-R^2_j}\right) \qquad j = 1,2,\ldots,k \tag{29}$$

Donde R^2_j es el coeficiente de determinación de la regresión cuando la variable X_j se utiliza como respuesta y el resto de las k-1 variables como predictores. Este coeficiente mide el porcentaje de la variación de X_j que es común o compartida con las otras k-1 variables independientes. El VIF =1 cuando este coeficiente es cero, esto es, cuando la j-ésima variable no está relacionada con los otros predictores, a medida que R^2_j se incrementa, es decir, su máximo es uno, mayor es el efecto inflacionario sobre la varianza del coeficiente de regresión de X_j, debido a la fuerte asociación lineal que esta variable guarda con las otras variables en el modelo. En general, entre mayor el valor del VIF, mayores los problemas de interpretación para los coeficientes de regresión debido a su inestabilidad ante cambios en la estructura del modelo de regresión o de la muestra que se utilice. Se sugiere que se tome un valor de VIF > 5 como indicativo de una multicolinealidad que deriva en los problemas antes mencionados (Willan y Watts, 1978 y Belsley et al., 1980).

¿Sabías qué?

Ronald Aylmer Fisher fue un matemático, biólogo y genetista británico. Entre otras cosas, Fisher es conocido por sus contribuciones a las estadísticas mediante la creación de la prueba exacta de Fisher y la ecuación de Fisher. Muchos científicos contemporáneos de Fisher lo consideraron un genio que creó las bases de la moderna ciencia estadística prácticamente solo.

4.5 Métodos de selección del mejor modelo de regresión

Un buen modelo de regresión debe contribuir a explicar una porción significativa de la variabilidad de la respuesta y ser además

parsimonioso. Esta última característica implica que únicamente aquellas variables que contribuyen a explicar la respuesta deben ser consideradas en la regresión y que la función de asociación *f* (.) que relaciona a la respuesta con las variables independientes es simple. Para que una variable independiente se incluya en el modelo de regresión, la ganancia adicional en la variabilidad explicada de la respuesta, que está medida por SSR, tiene que ser grande. Es decir que únicamente cuando la suma de cuadrados extra asociada a un predictor es significante, su introducción al modelo representa una ganancia global neta. Un problema para la selección de variables a introducir en el modelo de regresión es que si los predictores están correlacionados entre sí, cuando son introducidos adicionalmente o en conjunto no se registra un incremento substancial en la variación explicada de *Y*, esto es en SSR.

La ganancia adicional sobre SSR al introducir una nueva variable en el modelo de regresión múltiple puede cuantificarse a través de los coeficientes de determinación parcial, que miden el porcentaje adicional en la variación de la respuesta, ésta se explica cuando otra variable independiente se introduce al modelo de regresión. Si se considera un modelo de regresión con únicamente dos variables, el coeficiente de determinación parcial de *X2* dado *X1* se define como en (30):

$$R^2_{x_1|x_2} = \frac{SSR\left(X_2|X_1\right)}{SSE\left(X_1\right)} \qquad (30)$$

donde

$SSR\left(X_2|X_1\right) = SSR\left(X_1,X_2\right) - SSR\left(X_1\right)$ es la suma de cuadrados "extra" de *X2* dado *X1*, la cual cuantifica la contribución adicional que hace *X2* a la suma de cuadrados de la regresión del modelo en dos variables, respecto al modelo que tiene como única variable a *X1*. El coeficiente parcial (30) se interpreta entonces como la porción de la variabilidad no explicada por el modelo con *X1* -medida como SSE(*X1*)- y que puede explicarse cuando se introduce *X2*. Así, el

coeficiente de determinación parcial mide el porcentaje adicional de la variabilidad de *Y* explicada por *X2*, después de que la asociación lineal entre ambas variables con *X1* se ha tomado en cuenta. Cuando las variables predictoras están altamente correlacionadas entre sí, aún si algunas de ellas están estadísticamente asociadas con la respuesta, el coeficiente de determinación parcial tenderá a cero. Esto indica que una vez que otras variables asociadas con *Xk* se han usado para explicar a *Y*, el agregar a X_k no contribuye ya a explicar la variabilidad de *Y*.

La definición anterior para el coeficiente de determinación parcial puede extenderse fácilmente al caso de más de dos variables. Por Ejemplo, si se tiene un modelo de tres variables (*X1* , *X2* y *X3*) y se agregan dos más (*X4* y *X5*), el coeficiente de determinación parcial correspondiente se calcularía

$$R^2_{x_1|x_2} = \frac{SSR\left(X_5, X_4 \middle| X_1, X_2, X_3\right)}{SSE\left(X_1, X_2, X_3\right)} \qquad (31)$$

Observe que para calcular la suma de cuadrados extra es necesario ajustar dos modelos de regresión, uno reducido que incluye a las variables (*X1* ,*X2* y *X3*) y otro completo con las cinco variables. Esta suma de cuadrados extra tiene asociados 5-3 = 2 gl, equivalente al número de nuevas variables a incorporadas al modelo de regresión.

Para determinar si el porcentaje de la variación explicada adicional es estadísticamente significante, se utiliza como estadístico de prueba una Fparcial. Este estadístico, calculado según la expresión (32) establece si las variables adicionales contribuyen significativamente a explicar la variabilidad de Y que no era explicada por el conjunto de variables X1, X2 y X3. Los cuadrados medios requeridos para computar la F-parcial se calculan de la forma usual, es decir, dividiendo cada suma de cuadrados entre sus grados de libertad asociados. En general una suma de

cuadrados extra tiene asociados p-q grados de libertad (p>q) mientras que los grados de libertad del error son igual a n-p.

$$F^{*} = \frac{MSR\left(X_5, X_4 \middle| X_1, X_2, X_3\right)}{MSE\left(X_1, X_2, X_3, X_4, X_5\right)} \qquad (32)$$

Cuando las F-parciales se calculan para cada una de las variables independientes, dado que las otras variables ya fueron incluidas en el modelo, se tiene que la suma de cuadrados extra es de la siguiente forma:

$$SSR\left(X_k \middle| X_1, X_2, \ldots, X_{k-1}, X_{k+1}, X_p\right) \qquad k = 1, 2, \ldots, p \qquad (33)$$

Si la SSR extra en (33) se usa como numerador para la F-parcial, el estadístico resultante es equivalente a la t-Student, esto es $F(1, n-p) = t^2(n-p)$, que se definió en (19) y que permite probar la hipótesis individual de que el k-ésimo coeficiente de regresión parcial es igual a cero. Debido a esta equivalencia es que, bajo severa multicolinealidad, los coeficientes de regresión parciales se declaran cero aun cuando la variable independiente si esté significativamente asociada con la respuesta.

Cuando las variables independientes se van agregando una por una a la ecuación de regresión, como es el caso de las rutinas automáticas para seleccionar el modelo de regresión, el estadístico de prueba F se denomina F-secuencial ya que las sumas de cuadrados extras, cada una con un grado de libertad asociado, están dadas por la siguiente secuencia:

$$SSR\left(X_2 \middle| X_1\right),\ SSR\left(X_3 \middle| X_1, X_2\right), \ldots, SSR\left(X_k \middle| X_1, X_2, \ldots, X_{k-1}, X_{k-1}\right) \qquad (34)$$

Con el objetivo de seleccionar el mejor modelo de regresión se han diseñado algoritmos específicos que se enfocan en la identificación de las variables críticas a incluir en los modelos. Para

usar estos métodos automáticos de selección de variables, el usuario sólo necesita elegir un valor crítico de la F para la entrada, o salida, de una variable en la ecuación de regresión (F = 4 es el valor predeterminado en software estadístico como MINITAB) o bien, especificar el nivel de significancia deseado para la prueba con la F-parcial (0.15 es el valor sugerido por MINITAB). Usualmente este nivel de significancia se propone superior a los usuales de 0.05 y 0.01 ya que la distribución muestral de las F-secuenciales no es exactamente una distribución F lo que hace que el verdadero nivel de significancia sea menor al especificado. Draper y Smith (1998) sugieren que estos niveles de significancia sean sopesados con precaución y se manejen más como una referencia para la decisión, que como una auténtica probabilidad de cometer el error tipo I.

En general todos los métodos automáticos de selección de variables presentan dificultades cuando hay multicolinealidad, con riesgo de eliminarse variables críticas a favor de predictores menos asociados con otras variables independientes. Los métodos para la selección del modelo de regresión más populares se describen a continuación:

1. Backward

Como primer paso, en este método, se ajusta un modelo de regresión que contiene a todas las variables independientes consideradas por el analista. La estrategia seguida por este método es ir eliminando variables una por una en cada iteración. Una variable “sale” del modelo de regresión cuando su contribución marginal o adicional a SSR dada las otras variables no es considerable, esto es cuando su F-secuencial no es significante. El procedimiento continúa iterativamente hasta que todas las variables han sido eliminadas del modelo de regresión o cuando no hay más variables con F-secuenciales no significantes.

2. Forward

El primer paso, al usar este método, es ajustar un modelo de regresión lineal simple, donde la única variable es aquella con el mayor coeficiente de correlación significante con la respuesta. En las siguientes iteraciones, se agregan variables una por una. En este procedimiento una variable “entra” al modelo de regresión sólo si su contribución marginal a SSR determinada a través de la F-parcial es significante.

3. Stepwise

Este procedimiento inicia como Forward, estimando un modelo de regresión lineal simple con el “mejor” predictor y en un paso posterior se selecciona una segunda variable para tener un modelo bivariado. En el tercer paso se da la diferencia básica con Forward, ya que una vez que en el modelo hay dos variables, se evalúa la conveniencia de eliminar a la variable que se introdujo inicialmente lo cual se hace si su F-parcial resulta no-significante. En consecuencia, en este método pueden eliminarse variables que entraron en etapas anteriores, situación que no se presenta en el procedimiento Forward. Este método es el que mejor considera el problema de multicolinealidad entre predictores. El procedimiento continúa agregando nuevas variables y verificando por eliminación variables que se incluyeron en etapas previas hasta que todas éstas se han incorporado al modelo o bien cuando las F-parciales o F-secuenciales indiquen que ya no es apropiado agregar o eliminar más variables.

4. Todas las regresiones

Este procedimiento es altamente costoso y menos sistematizado que los anteriores, ya que propone ajustar todos los modelos de regresión en una, dos y hasta p variables, y examinarlos para elegir el más conveniente. El criterio de elección del “mejor” modelo se basa en el coeficiente de determinación, el modelo seleccionado es aquel que tenga el mayor coeficiente de determinación.

Todos estos métodos están disponibles en el software comercial pero se recomienda usarlos con precaución, siempre se debe tomar en cuenta el objetivo y la naturaleza del problema con el propósito de evitar la selección de modelos ilógicos desde el punto de vista de las relaciones teóricamente válidas entre las variables elegidas para el análisis de regresión.

¿Sabías qué?

El análisis de regresión es uno de los métodos de estadística multivariante de mayor difusión. Incluso en áreas que tradicionalmente no se asocian con métodos cuantitativos, como las de abogacía y política, el análisis de regresión ha demostrado su aplicabilidad. Una aplicación en el área legal es mostrar que hay discriminación de género en el área de trabajo y determinar sus consecuencias para un posible proceso legal. Otra aplicación interesante, dentro del área de consejería académica es elegir la universidad para la cual un aspirante está mejor calificado. Todas las disciplinas pueden beneficiarse de los métodos estadísticos multivariantes, pero se requiere de profesionistas que los comprendan y apliquen correctamente.

Ejemplo 4. Considera el problema planteado en el Ejemplo 2 de este capítulo. Dado que las cajeras tienen diferentes niveles de experiencia, la cual influye en el tiempo que tardan en hacer el cobro a un cliente, se decidió recopilar información adicional respecto a los meses en el trabajo que tienen las cajeras. Se ajustó entonces un modelo de regresión múltiple con 4 variables (las tres utilizadas en el Ejemplo 1 más *X4* = meses de experiencia de la cajera), *SSR* (*X1* ,*X2*,*X3*,*X4*) = 132.007. ¿Es apropiado incluir a *X4* = meses de experiencia de la cajera en el modelo de regresión?

Para responder a la pregunta planteada se requiere calcular la suma de cuadrados "extra" asociada a la introducción de esta nueva variable y calcular la prueba F-parcial correspondiente. La suma de cuadrados extra se calcula como:

SSR(*X4*| *X1* , *X2*, *X3*) = *SSR* (*X1* , *X2*, *X3*, *X4*) - *SSR* (*X1* , *X2*, *X3*) = 132.007 - 120.92 = 11.087

donde *SSR*(*X1* , *X2*, *X3*) se extrajo de la Tabla ANOVA reportada en el Ejemplo 2. Para calcular la suma de cuadrados del error para el modelo con cuatro variables basta con recordar la identidad del

ANOVA, esto es SST = SSR + SSE -recordar que SST no depende del número de variables en el modelo, siempre es igual mientras no se agreguen más datos - de donde *SSE* (*X1* , *X2*, *X3*, *X4*) = 140.018 - 132.007 = 8.011

Calculando entonces la F parcial para evaluar la reducción en SSE debido a la introducción de *X4*, se tiene:

F ((X4| *X1* , *X2*, *X3*) = 11.087 / 8.011 = 1.384

La región de rechazo es {*F* > *F* (a = 0.05; 1, 45) = 4.08} notar que se reducen los gl ya que se introduce una nueva variable, *gl (error)* = *n-p* = 50–5. El valor de F no está en RR, en consecuencia, la reducción en SSE ante la introducción de *X4* es no-significante, por lo que el modelo en cuatro variables sólo es más complejo pero no más útil para explicar los cambios en el tiempo de atención a clientes.

4.6 Modelos con variables cualitativas

Con frecuencia los predictores relevantes para un modelo de regresión resultan ser variables cualitativas, es decir variables cuyos "valores" corresponden a categorías nominales bien definidas. Algunos ejemplos de variables cuantitativas importantes en la construcción de modelos de regresión en la investigación de mercados son: marca de productos, sector económico de la empresa, género o profesión del consumidor.

La introducción al modelo de regresión de variables independientes cualitativas se hace a través del empleo de variables indicador o dummy en inglés. Para ilustrar el procedimiento considere el siguiente caso:

X1 = tipo de organización (0 = pública; 1 = privada).

X2 = valor total de la empresa declarado en libros (en millones de pesos)

Y = relación deuda/capital de la organización

El modelo de regresión lineal múltiple a ajustar es el siguiente:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$$

geométricamente el modelo anterior no resulta ser un plano en el espacio sino dos modelos, uno para cada tipo de organizaciones.

Es decir si *X1* =1, entonces:

$$Y_i = \beta_0 + \beta_1 (1) + \beta_2 X_2 + \varepsilon_i$$

$$Y_i = (\beta_0 + \beta_1) + \beta_2 X_2 + \varepsilon_i$$

Mientras que si *X1* = 0, se tiene:

$$Y_i = (\beta_0) + \beta_2 X_2 + \varepsilon_i$$

Es decir se tienen dos rectas paralelas, con β_2 representando el incremento en la relación deuda/capital por incremento de una unidad en el valor de la organización. Mientras que el parámetro β_1 "mide" la diferencia entre las relaciones deuda/capital dependiendo del tipo de organización que se trate. De acuerdo con esta discusión si se desea probar que existe una diferencia en la relación deuda/capital para los tipos de organizaciones, la hipótesis a probar sería la siguiente:

$$H_o: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

El estadístico de prueba para esta prueba individual es la t-Student ya establecida en (19), esto es

$$t = \frac{b_1}{S(b_1)} \qquad con \quad RR = \{|t| > t(\alpha/2, n-p)\}$$

En el caso descrito las dos rectas de regresión ajustadas tienen la misma pendiente, esto implica que, independientemente del tipo de organización, se tiene el mismo incremento en el monto de la deuda cuando se incrementa *X2* = volumen del negocio. Cuando este supuesto no es válido en la práctica y lo más razonable es suponer que la empresa pública tenderá a incrementar más el monto de su deuda con respecto a lo que lo haría una empresa privada, entonces el modelo anterior es insatisfactorio. Para describir adecuadamente la situación anterior, lo que conviene es incluir en el modelo de regresión un término de interacción.

Una interacción, en estadística, significa que el efecto de una variable independiente sobre la respuesta depende de los niveles de otra u otras variables. En el caso particular del ejemplo que se discute, la presencia de una interacción indicaría que el incremento en la relación deuda/capital, cuando aumenta el valor del negocio, no es el mismo para los dos tipos de empresa. Para considerar interacciones entre las variables independientes es necesario incluir en el modelo de regresión un término que es el producto de las variables que interactúan, es decir *X1*X2*. El modelo de regresión con interacción para dos variables sería el siguiente:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon_i \qquad (35)$$

Cuando una de las dos variables es cualitativa y tiene sólo dos categorías, resulta simple visualizar el efecto que tiene la introducción del término *X1*X2*. Basta con sustituir los valores (0,1) que asume la variable indicar *X1* en el modelo anterior. Si *X1* =1 (empresa privada) el modelo de regresión correspondiente es:

$$Y_i = \beta_0 + \beta_1(1) + \beta_2 X_2 + \beta_{12}(1) X_2 + \varepsilon_i$$

$$Y_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_{12}) X_2 + \varepsilon_i$$

Mientras que si se trata de una organización pública (*X1* = 0) se tiene:

$$Y_i = \beta_0 + \beta_2 X_2 + \varepsilon_i$$

En este caso el modelo bivariado corresponde a dos rectas de regresión no paralelas. Cuando *X1* =1, la pendiente de la recta asociada es $\beta_2+\beta_{12}$ en tanto que si X1= 0, la recta asociada tiene pendiente de β_{12}. Esta situación se ilustra gráficamente en la Figura 4.3.

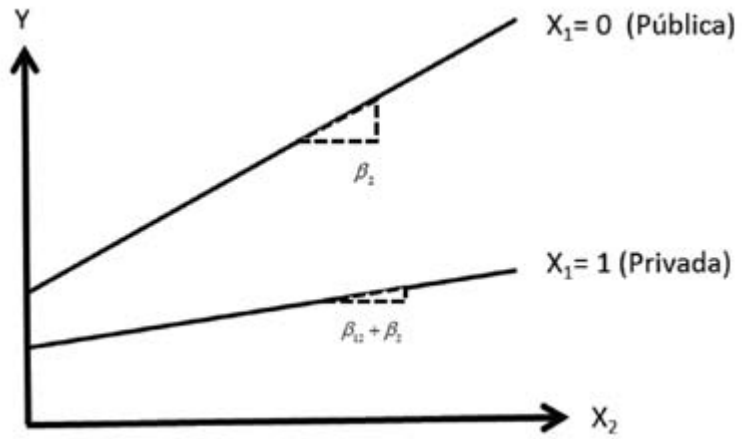


Figura 4.3 Modelo de regresión bivariada con dos interacciones

Para probar por la presencia de interacción entre las variables independientes, el coeficiente de interés es β_{12}. La hipótesis a probar sería:

$$H_o: \beta_{12} = 0$$
$$H_a: \beta_{12} \neq 0$$

Actividad de repaso

4. Regresión lineal múltiple

Calificar cada declaración como falsa o verdadera

1. ___ Los términos en la diagonal en la matriz de varianza-covarianza son las varianzas de cada uno de los términos de error.

2. ___ Algunos problemas asociados a la multicolinealidad son los cambios sustanciales en los estimados de los coeficientes de la regresión

3. ___ En presencia de multicolinealidad las desviaciones estándar de los coeficientes del modelo de regresión múltiple se incrementa considerablemente.

4. ___ Valores del Factor Inflacionario de Varianza (VIF) mayores o iguales a cinco sugieren una multicolinealidad moderada

5. ___ El uso del método de estimación de mínimos cuadrados no garantiza que los estimadores sean insesgados

6. ___ Si el coeficiente de determinación tiene un valor pequeño, esto significa que las variables independientes no son causa de la variación de la respuesta

7. ___ Un coeficiente de determinación bajo indica que el modelo construido a partir de las variables independientes, éstas explican poco la variabilidad de la respuesta

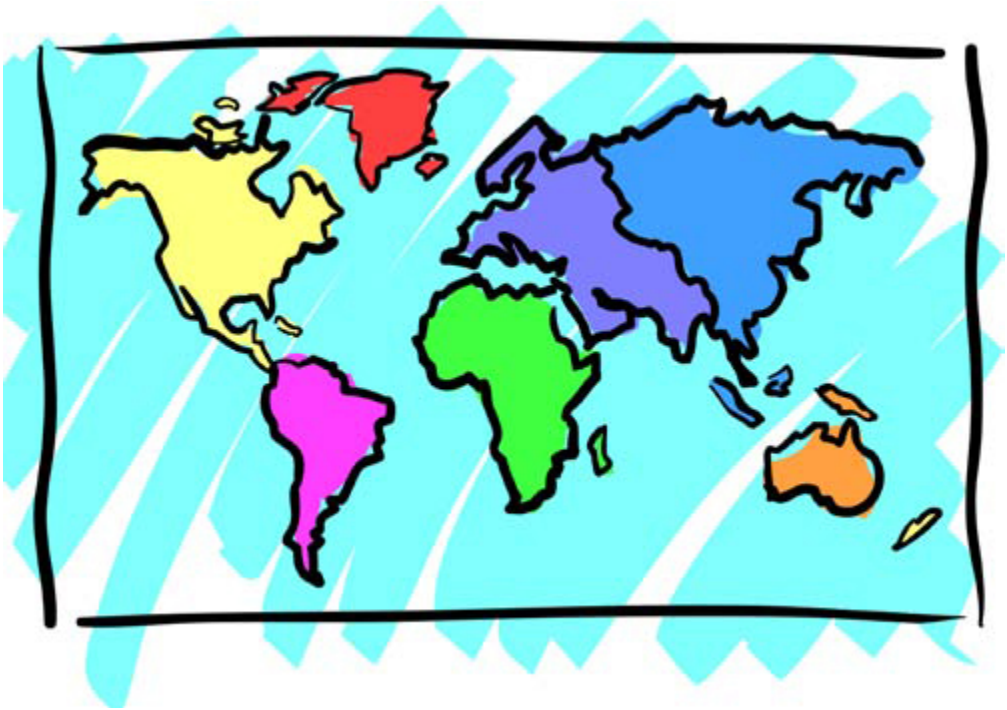
8. ___ Los residuos estandarizados son aquellos residuos que una media de 1 y una desviación estándar de 0.

9. ___ El método de eliminación progresiva (backward) selecciona todas las variables independientes en el análisis de regresión y posteriormente va eliminando aquellas variables que no suponen una contribución significativa a la predicción.

10. ___ La transformación crea una nueva variable y elimina la característica indeseable para dicha variable, permitiendo así, que contribuya favorablemente al modelo de regresión lineal múltiple
Respuestas:V,V,V,F,F,F,V,F,V,V

4.7 Aplicación del análisis de regresión con apoyo computacional

Un asesor presidencial desea mostrar evidencia cuantitativa en relación a las áreas que tendrán que fortalecerse en México para lograr un mayor crecimiento económico. Con este propósito seleccionó al azar un grupo de países y recopiló información sobre las variables relacionadas al nivel de preparación de la población local y de la actividad de varios sectores de la economía. Como respuesta o variable dependiente se eligió el Producto Interno Bruto de cada país en 1996, reportado en miles de millones de dólares americanos.



Las variables independientes, también para el año de 1996 son las siguientes:

X1 = Densidad poblacional (número de habitantes por kilómetro cuadrado)

X2 = Alfabetismo (proporción de la población que sabe leer y escribir)

X3 = Población económicamente activa (como porcentaje del total de residentes)

X4 = Ingresos anuales del sector agrícola (en millones de dólares americanos)

X5 = Ingresos anuales del sector manufacturero (en millones de dólares americanos)

X6 = Ingresos anuales por exportaciones (en millones de dólares americanos)

X7 = Preparación especializada (total de individuos que cursan educación media superior y superior)

Aquellas variables que determinen tener un impacto positivo sobre el PIB representarán las áreas de oportunidad para mejorarlo.

Sigue los siguientes pasos para aplicar el análisis de regresión. Compara tus respuestas con las proporcionadas por los autores.

a) Obtén la matriz de correlación. Con base al análisis de esta matriz, realice una selección preliminar de variables independientes y ajuste un modelo de regresión múltiple incluyendo sólo a aquellas variables estadísticamente correlacionadas con el PIB

b) Utiliza el procedimiento Stepwise, para hacer la selección automática de un modelo de regresión para el PIB. ¿Coincide el modelo seleccionado por Stepwise con el propuesto en el primer inciso? ¿Se diagnostican problemas graves por multicolinealidad?

c) Propón al asesor presidencial un modelo de regresión que permita describir la importancia que cada uno de los indicadores

económicos y de población tiene sobre el PIB de un país.

d) Realiza un análisis de **residuos** para el modelo que sugirió en c. ¿Se declaran desviaciones relevantes a los supuestos estadísticos para la regresión?

Respuestas:

a) Se emplea MINITAB para realizar el análisis. La secuencia de comandos a utilizar es:

La ventana de diálogo se muestra en la Figura 4.4. Se incorporan todas las variables candidatas al modelo y en la ventana de diálogo se selecciona que la matriz de correlación muestre los valores P (display p-values) para cada relación entre las variables.

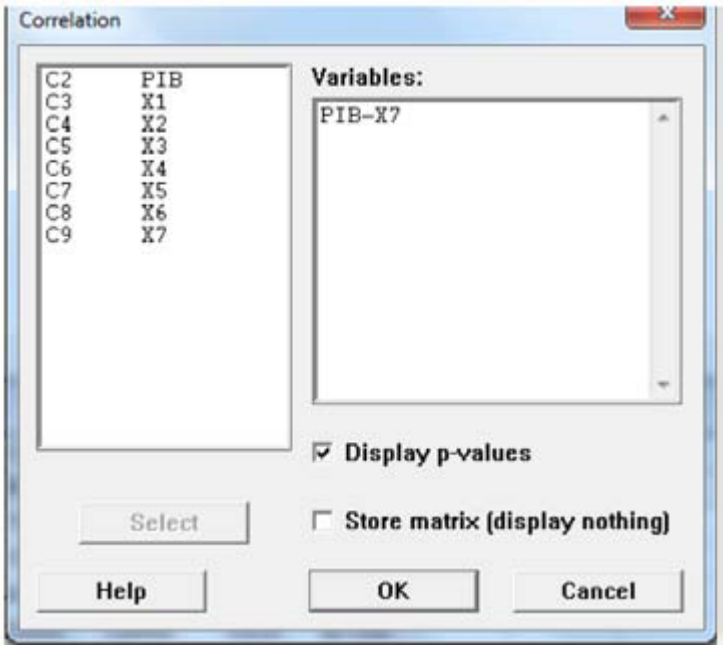


Figura 4.4 Ventana de diálogo de MINITAB para obtener la matriz de correlación

Una vez que se cierra la ventana de diálogo de la Figura 4.4 al dar clic en OK, se mostrará la **matriz de correlaciones** para las variables, como lo muestra la figura de la Tabla 4.4.

Correlations: PIB, Dens_pob, Alfabetismo, Pob_econ_ac, ing_agric, ing_ manuf,ing

	PIB	Dens_pob	Alfabeti	Pob_econ	ing_agri	ing_ man	ing_expo
Dens_pob	0.293						
	0.155						
Alfabeti	0.292	-0.069					
	0.157	0.743					
Pob_econ	0.329	0.192	0.393				
	0.108	0.357	0.052				
ing_agri	0.617	0.366	0.123	0.272			
	0.001	0.072	0.557	0.188			
ing_ man	0.982	0.380	0.314	0.339	0.635		
	0.000	0.061	0.127	0.097	0.001		
ing_expo	0.682	0.301	0.304	0.467	0.316	0.668	
	0.000	0.143	0.140	0.019	0.124	0.000	
est_medi	0.811	0.126	0.081	0.317	0.608	0.719	0.549
	0.000	0.549	0.699	0.123	0.001	0.000	0.005

Cell Contents: Pearson correlation

Tabla 4.4 Matriz de correlaciones para las variables

Primera etapa

Se analiza el vector de coeficientes de correlación, primera columna de R, entre la variable dependiente y las independientes para determinar cuáles variables están significativamente asociadas con la respuesta. Se realiza la Prueba de Hipótesis, que los coeficientes de correlación son cero. El estadístico de prueba es la t-Student y las hipótesis a probar son:

Ho: ρ = 0(Xi y Y no tiene relación lineal)

H_1: ρ ≠ 0 (Xi y Y tienen relación lineal)

En el listado de MINITAB, además del coeficiente de correlación se obtienen los valores del nivel de significancia estimado (P's), que indican que sólo las variables densidad poblacional, alfabetismo y población económicamente activa no están relacionadas estadísticamente con la respuesta, cuando el máximo riesgo de error del Tipo I aceptable es 10%.

Otra alternativa para esta primera etapa es calcular manualmente el estadístico de prueba, t-Student:

Para el coeficiente de correlación r(X1, Y), t = 1.4696

Para el coeficiente de correlación r(X2, Y), t = 1.4641

Para el coeficiente de correlación r(X3, Y), t = 1.6708

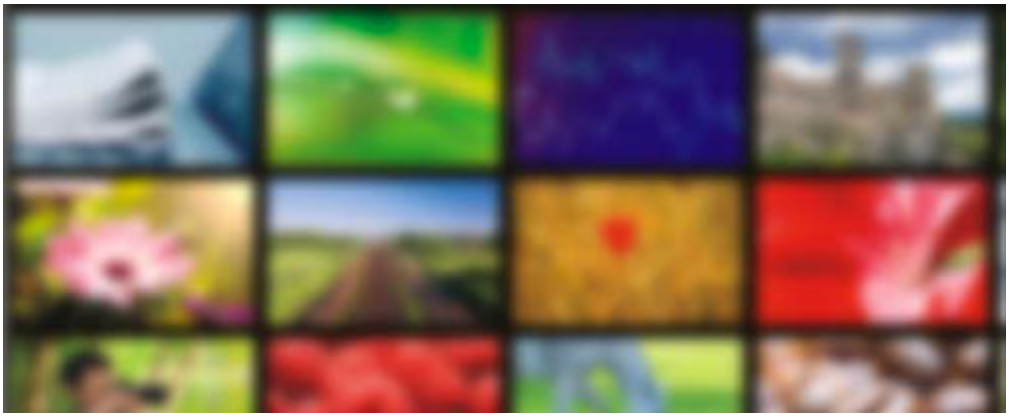
Para el coeficiente de correlación r(X4, Y), t = 3.7600

Para el coeficiente de correlación r(X5, Y), t = 24.9337

Para el coeficiente de correlación r(X6, Y), t = 4.4722

Para el coeficiente de correlación r(X7, Y), t = 6.6480

La RR = {|t |> t (0.025; 22) = 2.0686} de donde se concluye que las correlaciones de las variables X1, X2 y X3 con la respuesta no son significantes en tanto que las variables X4, X5, X6 y X7 están correlacionadas estadísticamente y por lo tanto guardan una relación lineal con la variable respuesta (PIB) y deberán incluirse en el modelo de regresión.



Segunda etapa

Análisis de la sub-matriz de coeficientes de correlación entre parejas de predictores para detectar posibles problemas de multicolinealidad. Las correlaciones simples altamente significantes son las siguientes:

Para r(X4,X5) = 0.635

Para r(X4,X7) = 0.608

Para r(X5,X6) = 0.668

Para r(X5,X7) = 0.719

Para r(X6,X7) = 0.549

Los datos anteriores alertan sobre potenciales problemas debido a la multicolinealidad, pero se deben valorar los niveles de VIF para cada una de las variables y así plantear si existe multicolinealidad.

Se emplea nuevamente MINITAB para realizar el análisis de regresión, la secuencia de comandos es la siguiente:

Stat → Regression → Regression

La ventana de diálogo solicita que se incorporen la variable respuesta y las variables predictoras. Como resultado del análisis anterior, de incorpora el PIB a la variable respuesta y las variables X4,X5,X6 y X7, como se muestra en la Figura 4.5

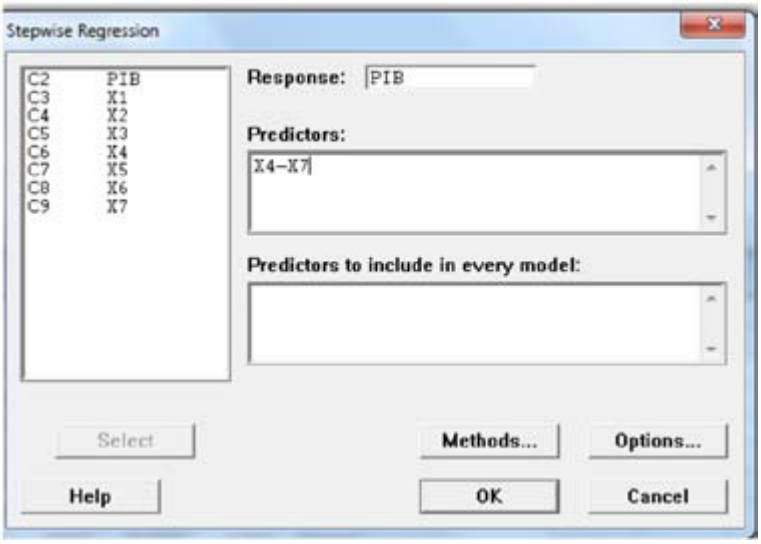


Figura 4.5 Ventana de diálogo de MINITAB para obtener el análisis de regresión.

Tercera etapa

Mediante el Análisis de Varianza se prueba la siguiente hipótesis

H0: Todos los coeficientes de regresión son cero vs.

H1: Al menos uno de los coeficientes de regresión es diferente de cero

El estadístico de prueba apropiado es F = 313.855, la región de rechazo para un nivel de significanci a del 5% es RR = {F > F(α; p-1,n-p)} , para este caso particular se tiene

RR = {F > F(0.05; 6,19) = 2.6283}

Se rechaza la hipótesis nula y se concluye que al menos uno de los coeficientes de regresión parciales es diferente de cero.

Cuarta etapa

El objetivo es definir el modelo de regresión múltiple con el mínimo número de variables predictoras. Del análisis conducido en la primera etapa se identificó que las variables densidad poblacional, alfabetismo y población económicamente activa no están estadísticamente asociadas con la respuesta por lo cual no tendrían que incluirse en el modelo de regresión ya que sus coeficientes en el modelo serán declarados cero. Para verificarlo, se ajustó el modelo de regresión en todas las variables. Los valores del estadístico t-Student en el listado permiten probar la hipótesis individual de que el i-ésimo coeficiente de regresión es cero.

The regression equation is

PIB = 163 - 0.809 Den_Pob + 1.04 Alfabetismo - 4.47 Pob_ eco_activ
- 0.00210 ing_agric + 0.00299 ing_manuf + 0.00242 ing_exp
+0.000134 estud

Predictor	Coef	SE Coef	T	P	VIF
Constant	162.8	327.2	0.50	0.625	
Den_Pob	-0.8089	0.4942	-1.64	0.120	1.5
Alfabeti	1.035	2.956	0.35	0.731	1.5
Pob_eco	-4.464	4.895	-0.91	0.373	1.5
ing_agri	-0.002097	0.001335	-1.57	0.135	2.1
ing_manu	0.0029912	0.0001517	19.72	0.000	3.9
ing_exp	0.002419	0.003715	0.65	0.524	2.3
estud	0.00013368	0.00002285	5.85	0.000	3.0

S = 182.8 R-Sq = 99.2% R-Sq(adj) = 98.8%

PRESS = 4195005 R-Sq(pred) = 93.84%

Tabla 4.5 análisis del modelo de regresión múltiple

Como se esperaba, los coeficientes de regresión de las variables X1 densidad poblacional, X2 alfabetismo y X3 población económicamente ectiva son declarados cero. Sin embargo, los coeficientes de las variables X4 ingresos del sector agrícola y X6 ingresos por exportaciones también son declarados cero, aun cuando en la etapa 1 se había concluido que estas dos variables individualmente guardan una alta asociación lineal con la respuesta. La explicación es que, a pesar de que los VIF's indican una multicolinealidad no tan severa, una vez que las variables X4, X5 y X7 se han incluido en el modelo de regresión, agregar a este modelo la variable X6 ya no implica una contribución significante. Igual situación se da para el caso de X4.

Si se ignora el análisis de correlación inicial, el modelo de regresión múltiple que se propone considerará sólo a aquellas variables para las cuales sus coeficientes de regresión son estadísticamente significativos. La estimación y evaluación de este modelo se da a continuación:

The regression equation is

PIB = - 122 + 0.00287 ing_manuf +0.000130 estud

Predictor	Coef	SE Coef	T	P	VIF
Constant	-122.28	51.10	-2.39	0.026	
ing_manu	0.0028656	0.0001232	23.25	0.000	2.1
estud	0.00013027	0.00002119	6.15	0.000	2.1

S = 203.6 R-Sq = 98.7% R-Sq(adj) = 98.5%

PRESS = 3970839 R-Sq(pred) = 94.17%

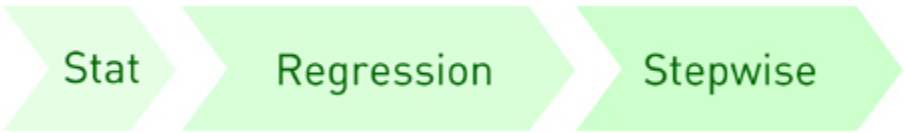
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	6721508900	33607544	811.00	0.000
Residual Error	22	911667	41439		
Total	24	68126756			

Tabla 4.6 análisis del modelo de regresión múltiple considerando solo las variables con coeficientes de correlación significativos

b) Este modelo está ignorando a dos variables fuertemente correlacionadas con la respuesta: ingreso por agricultura e ingreso por exportaciones. Puede ajustarse un modelo en las cuatro variables, pero los coeficientes de regresión no resultarán apropiados para establecer inferencias para el impacto de los predictores sobre el PIB debido a los problemas de multicolinealidad.

Se emplea MINITAB para realizar el algoritmo de Stepwise, la secuencia



de comandos es la siguiente:

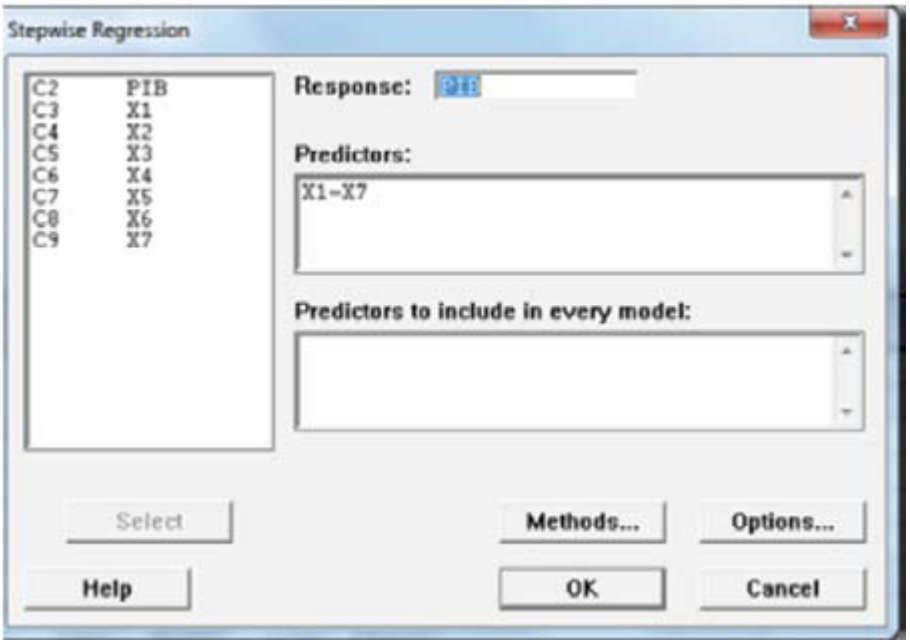


Figura 4.6 Ventana de diálogo de MINITAB para realizar el algoritmo de Stepwise,.

La ventana de diálogo solicita que se incorporen la variable respuesta y las variables predictoras.

Una vez realizada el llenado de la ventana de diálogo, MINITAB desplegará información similar a la de la Tabla 4.7, como se muestra a continuación:

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is PIB on 7 predictors, with N = 25

Step	1	2	3	4
Constant	7.542	-122.275	-76.514	-9.676
ing_man	0.00341	0.00287	0.00297	0.00305
T-Value	24.69	23.25	24.77	25.53
P-Value	0.000	0.000	0.000	0.000
est_medi		0.00013	0.00014	0.00013
T-Value		6.15	7.17	6.72
P-Value		0.000	0.000	0.000
ing_agri			-0.0030	-0.00024
T-Value			-2.39	-1.96
P-Value			0.026	0.064
Dens_pob				-0.85
T-Value				-1.96
P-Value				0.065
S	328	204	185	173
R-Sq	96.36	98.66	98.95	99.12
R-Sq(adj)	96.20	98.54	98.80	98.94
C-p	53.2	8.3	4.5	3.0

Tabla 4.7 Resultados del análisis Stepwise

El modelo propuesto por STEPWISE coincide en elegir como primeras variables a X5 y X7 como se planteó en el inciso anterior. De la variable X4 se había identificado que tenía correlación significante con la respuesta, por lo que se justifica que se haya agregado. Sin embargo, agregar X1 no es satisfactorio ya que hay otras variables como X4 y X6 que tienen una mayor asociación con la respuesta. Esto hace evidente que el uso de métodos automáticos para selección de los modelos de regresión debe hacerse con precaución, pues es difícil manejar modelos de regresión en donde hay conjuntos de variables que exhiben patrones de asociación.

c) La diferencia entre los coeficientes determinación para los modelos con dos, tres y cuatro variables es mínima, por lo que, combinado este criterio con los resultados de las pruebas de

hipótesis para los coeficientes de correlación y la regresión empleando Stepwise, se recomienda como mejor modelo al que considera a las variables X4, X5, X7. Si se desea incluir también a la variable X6 ingreso por exportaciones, por su alta asociación con el PIB, es necesario emplear otro método de estimación (no OLS) para el modelo de regresión.

El modelo estimado es el siguiente:

PIB = -76.514 + 0.00297 (Ingreso por manufactura) + 0.00014 (estudiantes a nivel medio y medio superior) - 0.0030 (Ingreso por agricultura)

Este modelo tiene asociado un coeficiente de determinación de 98.8%. Note que en el modelo anterior el coeficiente de regresión de la variable "ingreso por agricultura" es negativo, lo que se interpretaría como que un incremento en el ingreso por agricultura decrece el valor del PIB. Esto no es congruente ni con el coeficiente de correlación positivo de esta variable con la respuesta (0.617) ni con una realidad económica. No hay un error estadístico, sino una estimación deficiente del valor del coeficiente debido a la multicolinealidad. En consecuencia otra opción para sugerir es el modelo en sólo dos variables: ingreso por manufactura y estudiantes a nivel medio y medio superior, el cual se reportó previamente indicándose que tiene el inconveniente de omitir variables relevantes al PIB.

a) Desviaciones a los supuestos:

La forma más simple de evaluar los supuestos es realizar un análisis gráfico de los residuos. Se emplea nuevamente MINITAB para obtener los gráficos de los residuos. El comando es la siguiente:

Se muestra la ventana de diálogo como la de la Figura 4.7 y se selecciona la opción de presentar las cuatro gráficas de residuos en

una.

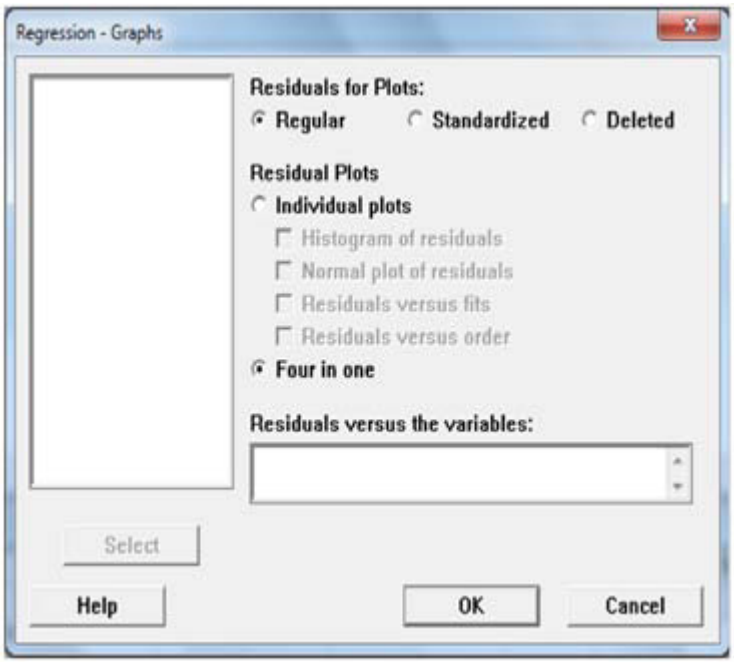


Figura 4.7 Ventana de diálogo de MINITAB para análisis de residuos en una sola.

A continuación, se realiza un análisis de residuos a partir de las gráficas que se presentan en la Figura 4.8

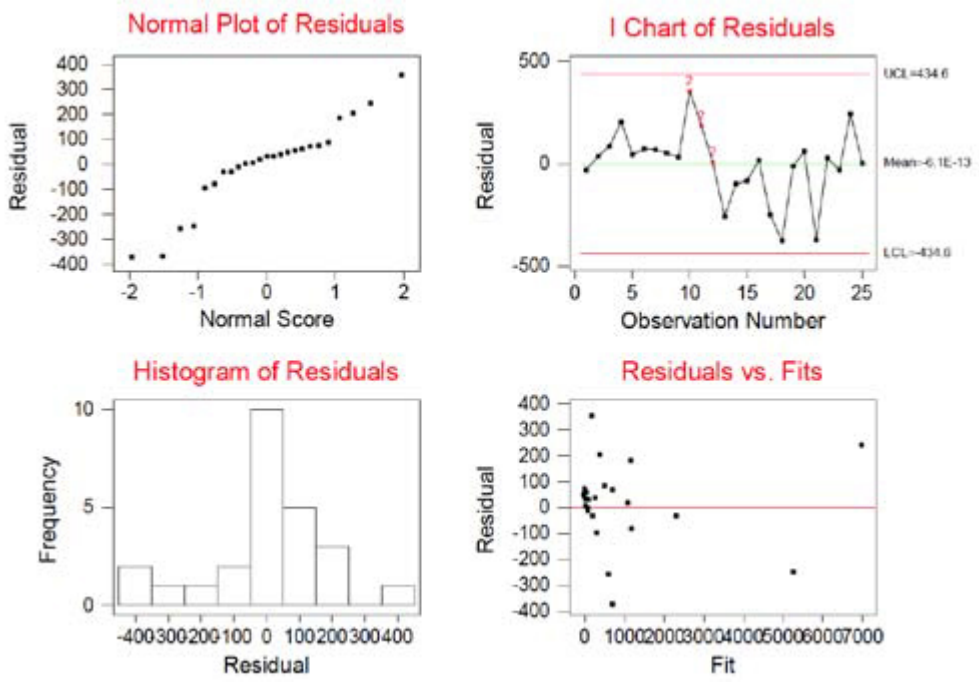


Figura 4.8 Gráficas del análisis de los residuos del modelo

» Independencia

La gráfica de residuos vs. tiempo, observation number, no revela patrones de asociación entre los residuos del modelo, sin embargo se marca una secuencia de tres residuos de igual signo, si la gráfica es sospechosa, lo recomendable es realizar pruebas estadísticas adicionales.

i) Probar por no autocorrelación de primer orden entre los términos del error. Esto es un problema potencial en este caso debido a que los datos se obtuvieron en diferentes periodos de tiempo.

Prueba de Hipótesis:

H_0: $\rho_1 = 0$

H_1: $\rho_1 \neq 0$

La regla de decisión para la prueba de hipótesis es

Si $D > d_U$, concluir H_O

Si $D > d_L$, concluir H_1

Si $d_L < D < d_U$ la prueba es inconclusa

Para un nivel de significancia de 0.05 y k = 2 los valores de d_U = 1.55 y el de d_L = 1.21.

Si el valor del estadístico Durbin-Watson es

Durbin-Watson statistic D = 1.43

Por lo tanto, la prueba es inconclusa. Se recomienda buscar pruebas alternativas como el análisis de la Función de Autocorrelación y la Prueba de Ljung y Box las cuales pueden ser consultadas por el lector interesado en Hanke (2006).

» Varianza no constante de los errores

En la gráfica de los residuos vs. los valores ajustados (fits) puede observarse que la varianza de los errores es la misma ante diferentes valores ajustados. Esta no es una prueba formal, pero la gráfica no resulta sospechosa de Heterocedasticidad.

» Normalidad para los errores

La gráfica de residuos vs. scores normales, Normal Plot of residuals, muestra alteraciones, los residuos no tienden a alinearse a lo largo de una recta imaginaria en la gráfica. Se tiene además que el histograma exhibe sesgo hacia la izquierda y un pequeño conjunto de observaciones mayores que el resto (rectángulo a la derecha). Se sugiere aplicar de nuevo pruebas estadísticas formales para evaluar el supuesto de normalidad.

Planteamiento de hipótesis

H_0: Los errores se distribuyen normalmente, versus

H_1: Los errores no se distribuyen normalmente

Para realizar esta prueba, utilizar la rutina Normality Test dentro del menú Basic Statistics en MINITAB. Los resultados de la prueba son gráficos y numéricos, además pueden emplearse varios estadísticos de prueba: Kolmogorov-Smirnov, Shapiro-Wilk y Anderson-Darling. Los resultados de la prueba de Kolmogorov se reportan en la Figura 4.9, como el valor de P < 0.05 esto lleva a rechazar la hipótesis de que los errores se distribuyen normalmente.

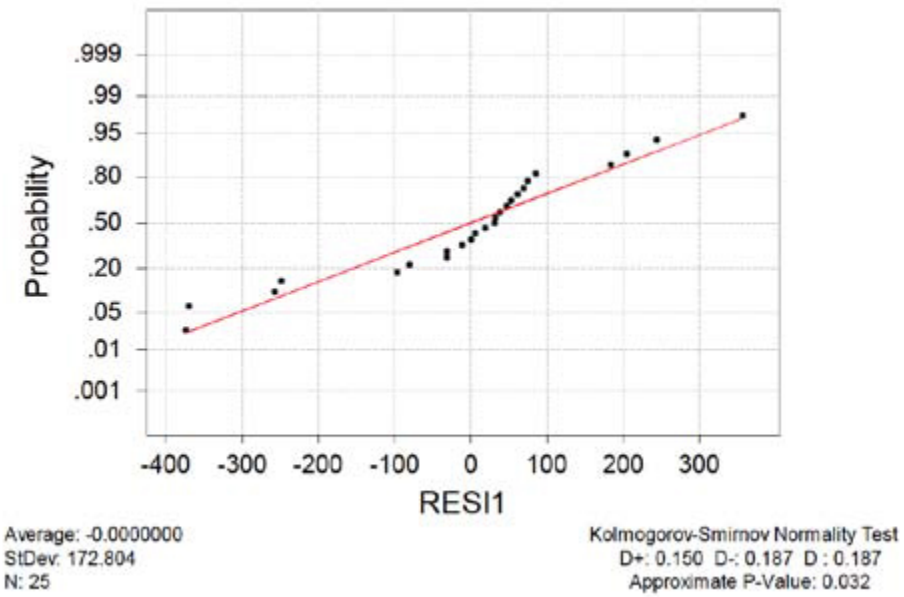


Figura 4.9 Gráfica de la Prueba de Normalidad para los residuos del modelo

El anterior análisis muestra que existen desviaciones a la normalidad de los errores. Sin embargo el análisis de varianza es robusto es este tipo de desviaciones por lo que se puede considerar que, ya sea el modelo en dos o tres variables, permitirá predecir el comportamiento del PIB, el problema es que ninguno de los dos modelos considera el total de variables relevantes para determinar el PIB, además de que los estimados para los coeficientes de regresión no son lo bastante precisos como para establecer políticas respecto a cuál es la contribución de cada variable (ingreso de manufactura, ingreso de agricultura, estudiantes a nivel medio y medio superior) sobre el PIB.

Actividad de repaso

4. Regresión lineal múltiple

RELACIONAR COLUMNAS

Dos o más variables independientes están correlacionadas entre ellas y su coeficiente de correlación es significante.	○	○	Coeficiente de regresión lineal simple
Representa el porcentaje de la variación total de la respuesta a través del modelo de regresión múltiple, penalizando la contribución de variables que no sean estadísticamente significantes.	○	○	Análisis Exploratorio de Datos (EDA)
Este tipo de estimación tiene importantes ventajas sobre la estimación puntual: se asocia a una probabilidad de que el intervalo contenga el valor poblacional o parámetro, lo cual es garantía de la calidad de la inferencia y se establecen un rango de posibles valores para dicho parámetro.	○	○	Coeficiente de regresión parcial
La diferencia entre los valores reales y predichos de la variable dependiente.	○	○	El coeficiente de determinación
Mide el cambio neto esperado en la respuesta ignorando la presencia de otras variables que también provocan cambio sobre ésta.	○	○	El coeficiente de determinación ajustado
Descripción de datos en los que la varianza del término de error (ε) aparece constante sobre un rango de variables independientes	○	○	La estimación de intervalos de confianza para la respuesta media
Método de selección de variables, donde las variables independientes adicionales se seleccionan en términos del aumento del poder explicativo que puede añadirse al modelo de regresión.	○	○	Homocedasticidad
Mide el cambio marginal sobre la respuesta suponiendo que es posible incrementar sólo la i-ésima variable y mantener fijas a todas las demás.	○	○	Estimación por pasos en la regresión múltiple
Se interpreta como el porcentaje de la variación total de la respuesta a través del modelo de regresión múltiple.	○	○	Residuo
Agrupa aquellos métodos de análisis principalmente de tipo gráfico que se utilizan para explorar las características de los datos y sus patrones de comportamiento.	○	○	Coeficiente de regresión lineal simple

Ejercicio integrador

4. Regresión lineal múltiple

Instrucciones: Lee la siguiente información y responde a las preguntas. Consulta en la barra lateral derecha las soluciones propuestas por los autores.

En una primera parte del estudio se construyó un modelo de regresión donde la respuesta fue Y = horas promedio diarias que la persona está conectada a la red, medida directamente después de solicitar y pagar al participante por conectarse al sistema de la empresa de Investigación de Mercados que realizó el estudio. Al momento de conectarse, el usuario proporcionó información sobre las siguientes variables independientes:

X1 = Género de la persona (1 si para hombres, 0 para mujeres)

X2 = Horas de trabajo o estudio por semana

X3 = Número de integrantes de la familia

X4 = Ingreso familiar mensual (en miles de pesos)

X5 = Años de escolaridad

El departamento de Investigación de Mercados, ajustó un modelo de regresión lineal múltiple en las cinco variables anteriores. El analista del departamento proporcionó el siguiente listado parcial obtenido durante su análisis. En la matriz de correlación, los coeficientes que involucran a X1 no son apropiados para análisis, por lo cual se sustituyeron sus valores por la siglas NA = no aplicable

Resultados del Análisis de varianza y matriz de correlación

ANOVA				Matriz de Correlación				
	gl	SS	Y	1				
Regresión	5	19.972		X1	NA	1		
Error	52	11.888		X2	-.71	NA	1	
				X3	.08	NA	.22	1
				X4	.56	NA	.34	.13
				X5	.48	NA	.41	-.21

Variables incluidas en el modelo de regresión múltiple

Variable	Coeficiente	Desv. estándar	t	P	VIF
Constante	1.41411	0.1634	8.65	0.000	
X1	0.61739	0.30445	2.03	0.001	1.4
X2	-0.73971	0.31911	-2.31	0.010	1.0
X3	-0.05106	0.05023	1.02	0.423	1.2
X4	0.12234	0.05781	2.12	0.017	1.7
X5	0.19461	0.05057	3.85	0.000	1.1

Preguntas

a) Analiza la matriz de correlación y de conclusiones relevantes respecto a las variables independientes potencialmente útiles para explicar a Y y también considera problemas potenciales por multicolinealidad (note que se reportan los VIF's).

b) Prueba la hipótesis de que el modelo de regresión contribuye a explicar los cambios en Y. Calcula también R^2 ajustado y reflexiona sobre su resultado.

c) ¿Apoyan los datos la hipótesis de que los hombres son usuarios más intensos de Internet que las mujeres?

d) Construye un intervalo de confianza para el incremento esperado en el número de horas de uso de una persona si el nivel de escolaridad de ésta se incrementará en una unidad.

e) Con base al modelo de regresión estimado, define el perfil de los usuarios intensos de Internet vs. los usuarios poco frecuentes.

RESPUESTAS

Revisa las respuesta correspondientes al ejercicio integrador del capítulo 4

http://www.editorialdigitaltecdemon-terrey.com/materialadicional/id019/cap4/respuestas4.pdf

Conclusión capítulo 4

4. Regresión lineal múltiple

A través del uso del análisis de regresión se cumple con al menos dos propósitos esenciales:

1. determinar las variables críticas que explican la variable respuesta de interés y

2. pronosticar los valores futuros de la respuesta ya sea mediante estimadores puntuales o por intervalo de confianza.

El primero de estos propósitos es muy importante ya que con él se verifica cuál es el conjunto de variables que explican satisfactoriamente a una respuesta. El segundo permite anticipar el resultado de ciertas acciones de mercadotecnia.

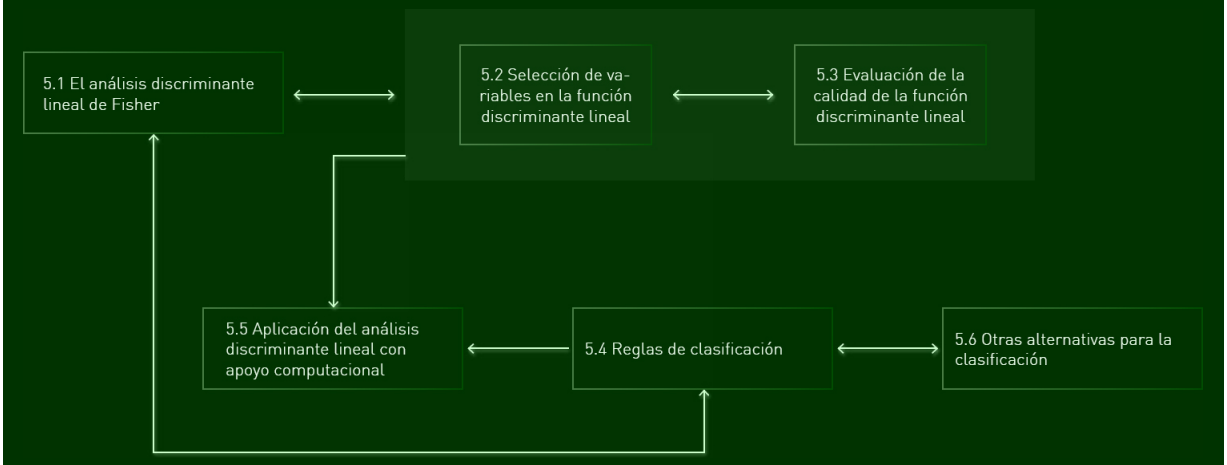
Para construir un modelo de regresión múltiple apropiado es necesario que se realice un análisis exploratorio de los datos para garantizar el establecimiento de relaciones pertinentes y válidas, así como para detectar problemas de asociación entre las variables independientes (multicolinealidad) que deriven en problemas importantes en la interpretación y calidad de los coeficientes del modelo de regresión. Tampoco se debe olvidar la importancia de evaluar los supuestos estadísticos básicos que garantizan la calidad de las inferencias estadísticas derivadas a partir del modelo.

Si bien se han desarrollado algoritmos que permiten la selección de variables a incluir en un modelo de regresión lineal múltiple, éstos deben de usarse con precaución y no deben tratarse como "cajas negras" donde el usuario sólo conoce los datos de entrada y acepta el modelo de salida. Por el contrario se invita a los usuarios a comprender las bases, operación y supuestos con los que trabajan estos algoritmos de selección para así hacer un uso correcto de ellos. Por último hay que distinguir entre el análisis estadístico de regresión y el método de mínimos cuadrados ordinarios, que es únicamente una técnica, simple de utilizar, para la estimación de los parámetros del modelo de regresión.

Lo atractivo de este método no está solo en su simplicidad sino en que, bajo los supuestos estadísticos para el modelo, proporciona estimadores que coinciden con aquellos derivados mediante métodos de estimación estadística. Sin embargo, ante la presencia de datos extremos o cuando hay violaciones a los supuestos, la calidad del método se deteriora. Por lo tanto se debe tener presente que existen otras alternativas de análisis que permiten sortear prácticamente todas las limitaciones de una regresión lineal.

Capítulo 5
Análisis discriminante multivariado

Organizador temático



5. Análisis discriminante multivariado

Introducción

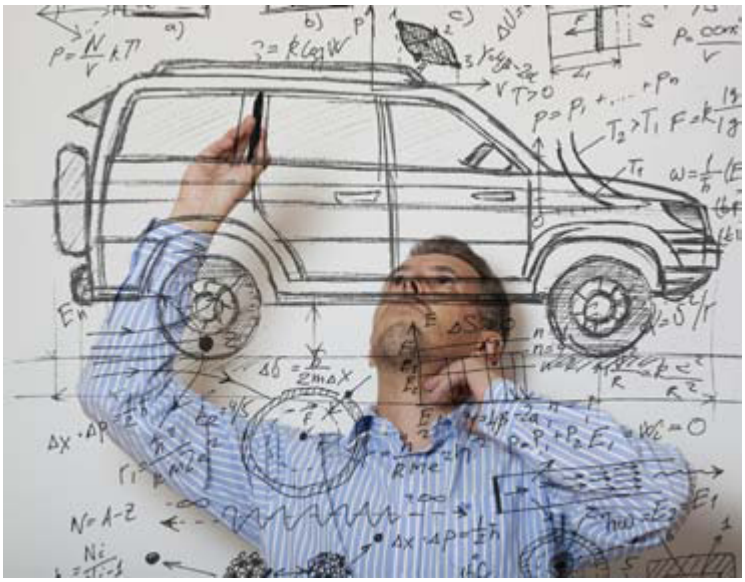
El análisis discriminante se refiere a aquellos métodos multivariables diseñados para separar y/o asignar elementos a un cierto número de grupos previamente especificados. En la práctica estos dos objetivos -separar y clasificar- se traslapan por lo que una función que discrimina puede utilizarse para clasificar mientras que una regla de clasificación sugiere un procedimiento de distinción.

Gracias a los avances en el área de estadística, el dise ño de computadoras y software, el análisis discriminante, al igual que otros métodos multivariables, se ha desarro llado y refinado. La primera fase del análisis discriminante es el trabajo seminal de Fisher (1936)

quien formuló una propuesta para el problema de discriminación basada en conceptos algebraicos. Posteriormente, Welch (1939), Rao (1948) y Anderson (1958) abordaron el problema desde una perspectiva probabilística. En una tercera fase, el pro blema de discriminación y clasificación se estudió desde el punto de vista de la teoría de decisiones, surgieron entonces aplicaciones importantes motivadas en el diagnóstico clínico (Swets y Pickett, 1982). El desarrollo de nuevos métodos para discriminación y clasificación sigue siendo de interés, las propuestas más recientes para la clasificación de individuos utilizan métodos multicriterio, conjuntos borrosos y métodos de inteligencia artificial como las **redes neuronales** (Cruz-Hernández, 2011).

En el contexto de inteligencia de mercados las aplicaciones de los métodos de discriminación y clasificación son numerosas, la mayoría de éstas se enfocan a identificar aquellos aspectos distintivos del segmento meta de la empresa y a predecir el comportamiento de compra, por Ejemplo marcas preferidas, del consumidor. El estudio reportado por Robertson y Kennedy (1968) que se resume en el recuadro es evidencia del largo historial de aplicaciones del análisis discriminante multivariado (ADM) en mercadotecnia.

La difusión y adopción de nuevos productos es importante para garantizar la posición competitiva de una empresa. El interés por precisar las características de los consumidores innovadores es relevante para el diseño de estrategias de promoción y penetración para productos novedosos. El segmento de clientes innovadores se refiere a aquel 10% de los consumidores que adoptan primero un producto que perciben como novedad.



En general se asume que los consumidores innovadores difieren en sus características de los no-innovadores, estas diferencias lleva a los innovadores a asumir el riesgo de adquirir nuevos productos. Varias características se han sugerido para distinguir o reconocer a los clientes innovadores de los no-

innovadores. Las que se consideraron en este estudio fueron: interés en comprar o probar artículos nuevos porque se perciben como innovadores; movilidad fuera de su grupo social y en el trabajo; un mayor grado de participación con la comunidad; un amplio rango de intereses de consumo; preocupación por el reconocimiento; una actitud cosmopolita y un estándar económico superior. A partir de los datos recolectados en una encuesta telefónica aplicada a 60 consumidores innovadores -identificados por haber sido de los primeros en comprar un novedoso electrodoméstico- y a 40 noinnovadores, se identificaron como características distintivas del cliente innovador a su interés personal por comprar o probar artículos nuevos y a su mayor movilidad fuera de sus grupos sociales y de trabajo. Otras variables que distinguen en menor grado al consumidor innovador fueron su situación económica privilegiada, su mayor participación con la comunidad y una actitud cosmopolita. A partir de este perfil puede promoverse mejor un nuevo producto para que sea adoptado lo más pronto posible y así acelerar el proceso de su difusión en el mercado.

5.1 El análisis discriminante lineal de Fisher

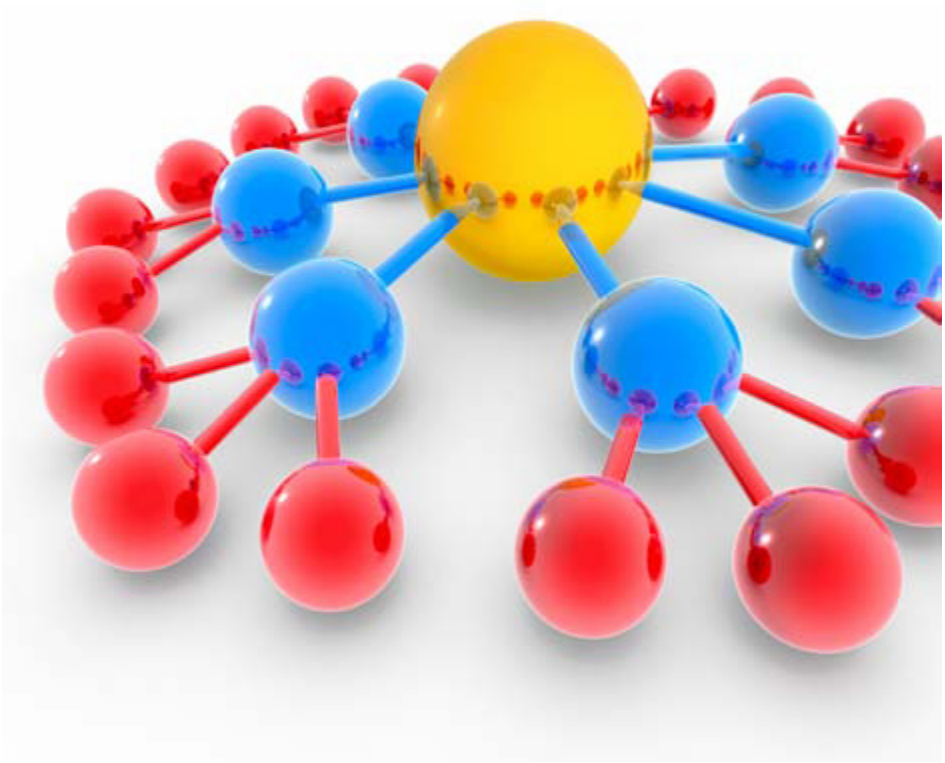
El propósito del análisis discriminante es maximizar la separación o distinción entre grupos de individuos cuyo grupo de pertenencia -marca favorita, situación crediticia, postura respecto a una campaña social o política, entre otros- es conocido a priori. En este sentido, el análisis discriminante multivariado (ADM) difiere del análisis de conglomerados ya que en este último los grupos a los que se asignan los individuos no son conocidos sino que se especifican a posteriori, esto es después del análisis. Cuando puede identificarse a las variables o atributos que caracterizan a los grupos y crear una función capaz de distinguir con precisión a los miembros de distintos grupos, también es factible asignar o clasificar nuevos objetos a los segmentos ya conocidos. Por lo cual los métodos de discriminación y clasificación están fuertemente vinculados: una función que discrimina puede ser útil para clasificar, mientras que una regla de

clasificación sugiere un procedimiento para la distinción de objetos. Los objetivos del análisis discriminante y de clasificación se formulan como sigue:

1. Describir algebraica y gráficamente las diferentes características o atributos de individuos que provienen de distintos grupos.

2. Determinar aquellas variables que mejor distinguen o discriminan individuos provenientes de distintos grupos.

3. Definir reglas que permitan la asignación o clasificación de nuevos individuos en los grupos especificados.



La propuesta de Fisher para la discriminación de dos grupos consiste en utilizar una función lineal de las variables discriminantes, aquellas elegidas para caracterizar a los grupos, tal que se maximice la distancia entre los grupos con respecto a la variabilidad de la **función discriminante.** Esta idea se describe gráficamente en

la Figura 5.1 para el caso de dos variables discriminantes, en esta figura los objetos u observaciones están incluidos en alguno de dos grupos, representados como elipses. En el espaciobidimensional, una combinación lineal de las variables *X*1 y *X*2 corresponde a una recta. En la figura se observa que el **discriminante lineal** #1, FDL (1) hace una mejor separación entre los dos grupos que el segundo discriminante FDL (2), en el sentido de que las **proyecciones** de los datos sobre cada función resultan en dos poblaciones aproximadamente acampanadas que en el caso de FDL (1) están bien separadas mientras que respecto a FDL (2) resultan muy traslapadas.

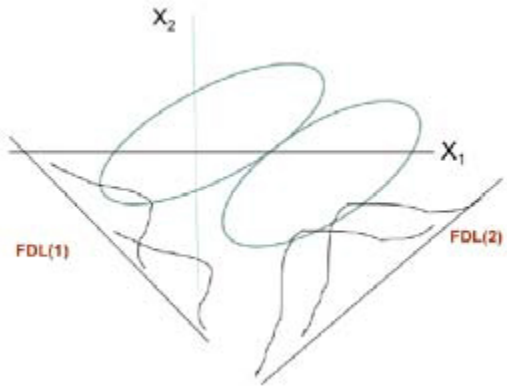


Figura 5.1 Construcción de una función discriminante lineal.

En el análisis discriminante clásico se requiere, además de que los *g* grupos de objetos sean conocidos, que las variables que se usen para discriminar estén en una escala de intervalo o razón. El ADM no-métrico, es decir aquel en que las variables discriminantes están en escalas nominales u ordinales, utiliza conceptos y métodos diferentes, algunos de ellos se describen en la última sección de este capítulo. Los datos necesarios para realizar un análisis discriminante métrico se obtienen a través de un muestreo estratificado, es decir primero se identifican los grupos y luego se procede a seleccionar aleatoriamente individuos dentro de cada grupo (Scheaffer, Mendenahll y Ott, 1996, p. 127), por lo cual los datos tienen el formato ($X1$,$X2$,…Xp, G_i) donde i = 1,2, … g. El número de individuos que se elige de cada grupo suele hacerse no-proporcional al tamaño del grupo o estrato, ya que en aquellos

casos en que hay grupos minoritarios se tendría un número escaso de observaciones lo que repercutirá en la calidad del análisis. Una sugerencia empírica en el caso de ADM es tener por cada grupo un total de 20+5*p* datos, recomendación que también se hizo en el caso del análisis de regresión.

Para el caso de dos grupos, la propuesta de Fisher de encontrar una combinación lineal de las variables discriminantes $Y = l'x$ que maximice la distancia entre los dos grupos se expresa como sigue:

$$\max_l \frac{DISTANCIA^2(PROYECCIONES)}{Var(Y)} = \frac{[l'(\mu_1 - \mu_2)]^2}{l'\Sigma l} \qquad (1)$$

donde $\mu_{G1} = l'\mu_1$ y $\mu_{G2} = l'\mu_2$ son las proyecciones de cada una de las medias de los grupos sobre la recta que repre senta a la función discriminante

Var (Y) = l'Σl es la varianza de la FDL

μ_1 y μ_2 son los centroides de cada grupo y Σ es la matriz de varianza-covarianza la cual se asume común a los dos grupos.

La solución al problema de optimización anterior (ver Johnson y Whicher, 1998 para la demostración) es: l = Σ−1(μ1 -μ2). En una situación práctica, los parámetros que caracterizan a los grupos (medias y matriz de varianzacovarianza) no son conocidos, por lo cual el discriminante lineal se estima a partir de datos muestrales como en (2).

$$Y = \hat{l}'x = (\bar{x}_1 - \bar{x}_2)'S_p^{-1}x \qquad (2)$$

La magnitud de las entradas del vector $\hat{l}$ se relaciona con la importancia relativa de cada variable para discrimi nar entre los grupos. En ausencia de colinealidad severa, los coeficientes estandarizados de las variables en DL, esto es las entradas del

vector $\hat{l}$, miden el poder discriminante o la importancia de cada variable para hacer la distinción entre los grupos.

El máximo valor que asume el cociente (1), es decir lo más que pueden separarse los centroides de los dos grupos es igual a la **distancia de Mahalanobis**, esta medida de distancia ya fue mencionada en el contexto de análisis de conglomerados. Para calcular esta distancia se utiliza la siguiente expresión:

$$D^2 = (\bar{x}_1 - \bar{x}_2)' S_p^{-1} (\bar{x}_1 - \bar{x}_2) \qquad (3)$$

Si bien el discriminante lineal Y induce la máxima separación D^2 entre grupos, esta separación puede no ser suficiente para distinguir grupos. Para determinar si se ha logrado una discriminación significante entre grupos o una separación importante entre ellos es necesario probar la siguiente hipótesis:

H_0: Separación insatisfactoria vs. H_1: Grupos bien separados

Bajo los supuestos de:

1) el vector de variables discriminantes *x*sigue la distribución normal multivariada y

2) la matriz de varianza-covarianza (verdadera) es igual para los dos grupos, el estadístico de prueba para la hipótesis anterior es una *F* con *p*=número de variables discriminantes y n_1+n_2-p-1 grados de libertad (McLachlan, 1992). El estadístico de prueba se calcula según se indica en (4).

$$F = \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - 2)p} \left(\frac{n_1 n_2}{n_1 + n_2} \right) D^2 \approx F(p, n_1 + n_2 - p - 1) \qquad (4)$$

La región de rechazo para un nivel de significancia α especificado por el analista tiene la siguiente forma RR = {$F>F(\alpha; p, n_1+n_2-p-1)$}.

Si se emplea el discriminante lineal *Y* se puede derivar una regla de clasificación como sigue:

i) Calcular los puntajes discriminantes para cada una de los centroides en cada grupo, esto es

$$\bar{y}_1 = \hat{l}'\bar{x}_1, \bar{y}_2 = \hat{l}'\bar{x}_2$$

ii) Calcular su promedio $m = \frac{1}{2}(\hat{l}'\bar{x}_1 + \hat{l}'\bar{x}_2)$

iii) Asignar a un nuevo elemento, cuyas características se describen en el vector x_0, al grupo 1 si su puntaje o score discriminante $y_0 - m \geq 0$ y al grupo 2 cuando $y_0 - m \leq 0$. Esto equivale a asignar al nuevo individuo a aquel grupo a donde quede más próximo.

Si el número de elementos en cada grupo no es el mismo, m se define como el promedio ponderado de los puntajes discriminantes medios, las ponderaciones son el número de elementos en cada grupo.

¿Sabías qué?

Sir Ronald Fisher (1890-1962), notable estadístico inglés nombrado Caballero en 1952, hizo importantes contribuciones a la estadística al aplicar sistemáticamente los métodos estadísticos en estudios de genética y biología. Entre sus contribuciones está el diseño de experimentos en agricultura, el método de análisis de varianza y el de máxima verosimilitud. Su propuesta para el análisis discriminante surgió también del área de biología.

Ejemplo 1. El estudio hecho por la Procuraduría Federal del Consumidor (PROFECO) sobre 16 marcas y 36 modelos de computadoras portátiles comparó las marcas con base en cinco

variables, las dos más importantes fueron: aplicaciones de productividad -aptitud del equipo para atender las aplicaciones más comunes como procesador de palabras, hoja de cálculo y navegación por Internet- y desempeño del equipo para multimedia.

Se ha sugerido que los usuarios de laptops valoran de diferente manera los atributos anteriores dependiendo de sus necesidades profesionales de cómputo, por lo cual un grupo de distribuidores de computadoras decidieron realizar un estudio de investigación de mercados cuyo objetivo fue determinar si los profesionales en áreas de ingeniería y de negocios tienen distintas necesidades en cuanto a la productividad y desempeño multimedia de su computadora portátil. En caso de que ambos grupos de profesionales aprecien de manera distinta estos criterios, los distribuidores van a identificar los modelos más idóneos para cada grupo y los promoverán de manera diferenciada, enfocándose en los beneficios que cada profesional valora.

Treinta ingenieros y 30 administradores expresaron -sobre una escala de 0 - 10, donde 0 = para nada importante y 10 = absolutamente importante- la importancia que para ellos tienen las citadas cualidades. La información para cada grupo se resumió en las siguientes medidas descriptivas.

$$\bar{x}_1 = \begin{bmatrix} 8.5 \\ 4.8 \end{bmatrix}, \quad \bar{x}_2 = \begin{bmatrix} 7.2 \\ 5.9 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} 3.1 & 2.8 \\ 2.8 & 3.4 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 3.5 & 3.1 \\ 3.1 & 4.6 \end{bmatrix}$$

El discriminante lineal de acuerdo a Fisher se calcula a partir de la información anterior mediante el siguiente procedimiento:

1) Combinar las matrices de varianza-covarianza muestrales bajo el supuesto de que ambos grupos de usuarios, ingenieros y administradores, exhiben el mismo grado de variabilidad y patrones de asociación en sus percepciones. Para este caso se tiene

$$S_p = \frac{(30-1)S_1 + (30-1)S_2}{30+30-2} = \begin{bmatrix} 3.3 & 2.95 \\ 2.95 & 4 \end{bmatrix}$$

2) Invertir la matriz de varianza-covarianza combinada, en este ejemplo

$$S_p^{-1} = \begin{bmatrix} 0.8894 & -0.6559 \\ -0.6559 & 0.7337 \end{bmatrix}$$

3) Calcular el discriminante lineal multiplicando al vector de diferencias entre centroides por la inversa de la matriz de varianza-covarianza combinada, esto es:

$$Y = l'x = (\bar{x}_1 - \bar{x}_2)' S_p^{-1} x = (1.3, -1.1) \begin{bmatrix} 0.8894 & -0.6559 \\ -0.6559 & 0.7337 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 1.8777x_1 - 01.6598x_2$$

En el espacio bidimensional la función anterior representa una recta, las proyecciones de los datos sobre esta recta resultan en la mayor separación posible. En particular, las proyecciones de los centroides de cada grupo sobre el discriminante lineal, denominados puntajes o scores discriminantes promedio, son iguales a: $\bar{y}_1 = l'\bar{x}_1 = 1.8777(8.5) - 1.6598(4.8) = 7.99341$, $\bar{y}_2 = l'\bar{x}_2 = 3.72662$, de donde la máxima distancia o separación alcanzada es de 7.993413.72662 = 4.26679.

Para responder a la pregunta de si los grupos de ingenieros y administradores valoran de manera diferente las características de una laptop, es necesario demostrar que la distancia anterior implica una buena separación entre grupos. El estadístico de prueba en este caso es igual a:

$$F = \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - 2)p}\left(\frac{n_1 n_2}{n_1 + n_2}\right)D^2 = \frac{(30 + 30 - 2 - 1)}{(30 + 30 - 2)2}\frac{(30)(30)}{(30 + 30)}4.26679 = 31.4492$$

Si el nivel de significancia se especifica como 5%, la región de rechazo es {*F*> *F*(.05; *p*=2, *n*1+*n*2−*p*−1=57) =3.15}. Como la *F* calculada está en la región de rechazo, es decir excede el valor crítico se concluye que a través de la función discriminante lineal se logra una buena distinción entre grupos. Esto implica que ingenieros y administradores se distinguen en cuanto al valor que asignan a las cualidades del equipo de cómputo que utilizan en sus actividades profesionales.

Es importante notar que en este ejemplo el propósito del análisis fue distinguir entre los grupos, no la clasificación de nuevos individuos, ya que para el distribuidor no es relevante “predecir” si un usuario es ingeniero o administrador sino establecer el perfil de necesidades del usuario para así recomendarle el equipo más apropiado.

Para el caso de más de dos grupos, la generalización del método de Fisher requiere de maximizar el cociente de la variación entre los g grupos respecto a la variación de los individuos dentro de cada

grupo. Si se asume que todos los grupos tienen la misma matriz de varianza-covarianza Σ, el problema se formula como en (5):

$$\max_l \frac{SS\ entre\ grupos}{Var(Y)} = \frac{l'Bl}{l'Wl} \qquad (5)$$

En la expresión anterior W es el estimado de la matriz de varianza-covarianza común a los grupos (Σ), este estimado se denomina matriz de varianza dentro de grupos (W de within) ya que captura la variabilidad que hay entre los individuos que pertenecen a un mismo grupo. La matriz B, por otra parte, se conoce como la matriz de productos cruzados entre grupos (B de between) la que cuantifica el grado de variabilidad que hay entre los centroides de los g grupos. Estas matrices se calculan como sigue:

$$B = \sum_{i=1}^{g} (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'; \quad W = \sum_{i=1}^{g} (n_i - 1)S_i = S_p \qquad (6)$$

Donde S_i es la matriz de varianza-covarianza de cada uno de los grupos, es decir W combina la información de cada uno de los grupos bajo el supuesto de que todos tienen la misma varianza.

El vector l, formado por el conjunto de coeficientes para las variables discriminantes, que minimiza el cociente (5) corresponde al eigenvector asociado con el mayor **eigenvalor** de la matriz $W^{-1}B$. La función $Y=\boldsymbol{l'x}$ calculada bajo esta propuesta se conoce como el ***Primer Discriminante***. Es posible probar (McLachlan, 1992) que la matriz $W^{-1}B$ tiene a lo más $s = mín\ \{p,g-1\}$ eigen-valores diferentes, es decir, que pueden calcularse hasta s discriminantes lineales para separar a $g>2$ grupos a partir de múltiples variables. Cuando $g = 2$, s = mín $\{p,2\text{-}1\}=1$, es decir, hay un único discriminante lineal que puede calcularse para distinguir entre dos grupos.

Cada uno de los discriminantes lineales es ortogonal respecto a los otros discriminantes. Esto es que en el espacio bi o tridimensional los discriminantes lineales son rectas perpendiculares

entre sí. Si bien es posible calcular hasta s discriminantes lineales para separar a los grupos, el principio de parsimonia -que se aplica para el modelado estadístico y también en el caso de ADM- establece como objetivo distinguir a los grupos usando el menor número posible de funciones discriminantes lineales. Para identificar aquellos discriminantes que son útiles para distinguir los grupos se utiliza como criterio el porcentaje con que el *k*-ésimo discriminante contribuye a la variabilidad entre grupos, esta contribución o aportación a la separación entre grupos se calcula como:

$$\lambda_k / \text{Suma}(\lambda_i) \qquad (7)$$

En la expresión (7) λ_k es el *k*-ésimo eigen-valor de la matriz $W^{-1}B$ que es también la varianza del *k*-ésimo discriminante. Al dividir el eigen-valor entre la suma de todos los eigen-valores, que representa la variación total registrada entre los grupos, se obtiene la contribución porcentual de cada función discriminante. Similar al caso de análisis factorial, el primer discriminante lineal es el que contribuye más a la separación entre grupos, mientras que la contribución de los siguientes discriminantes lineales va decreciendo gradualmente, ya que $\lambda_1 > \lambda_2 \ldots > \lambda_s$. Si hay g=3 grupos, existen dos discriminantes lineales, si el primero contribuye con un 85% a la separación entre grupos, la contribución del segundo es únicamente del 15% la decisión recomendada es utilizar únicamente un discriminante lineal para separar los grupos. Este caso se ilustra gráficamente en la Figura 5.2, note que los centroides de los tres grupos son colineales, en este caso el único discriminante que contribuye a la separación es Y1, mientras que el segundo no permite hacer ninguna distinción entre los grupos.

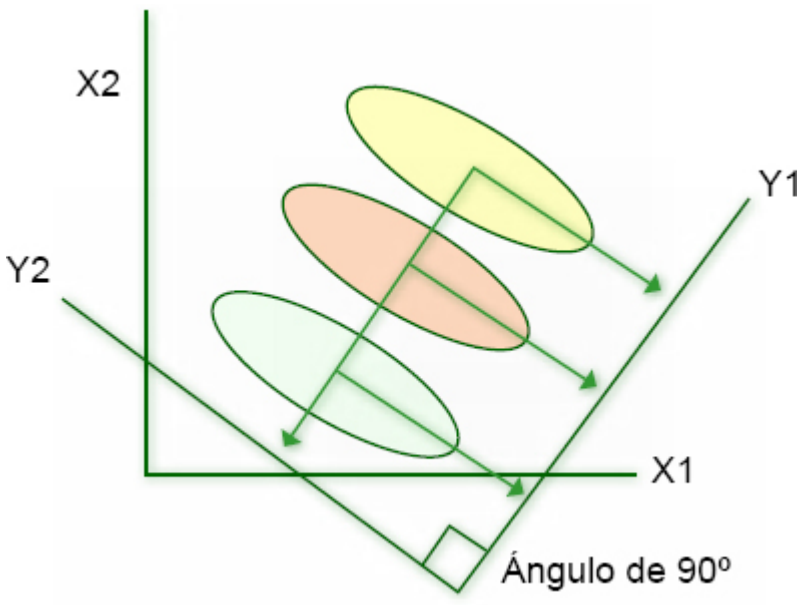


Figura 5.2 Discriminación de tres grupos con un único discriminante lineal

Cuando un único discriminante lineal es suficiente para lograr una buena distinción entre varios grupos, la regla de clasificación presentada en la sección anterior es aplicable, es decir se calcula el puntaje discriminante Y_0 = para una nueva observación x_0, y el individuo se asigna a aquel grupo para el cual la distancia al cuadrado entre Y_0 y el puntaje discriminante medio del grupo es la menor de todas. Sin embargo, cuando una separación significante entre grupos requiere de más de un discriminante lineal hay múltiples distancias que pueden calcularse, dependiendo del discriminante que se utilice para determinar la distancia entre un nuevo objeto y el centroide de cada uno de los grupos. En este caso lo que procede es obtener funciones de clasificación lineales (ver sección 5.4) para cada grupo. Al emplear estas funciones, se asigna un nuevo dato a aquel grupo para el cual su puntaje o score discriminante es el mayor de todos.

Ejemplo 2. Las tendencias del consumidor cambian con el tiempo y se relacionan con su estilo de vida y sus actitudes. En el caso del consumo de alimentos y bebidas, la tendencia actual es hacia el consumo de productos saludables lo que ha llevado a los fabricantes a introducir productos con menor contenido de

ingredientes como grasa, carbohidratos y sodio; así como alimentos funcionales, es decir aquellos que aparte de sus características nutricionales tienen componentes como vitaminas, fibras o micro-organismos que contribuyen a mejorar la salud. Entre las categorías de alimentos que han registrado un mayor crecimiento y ampliación en la variedad de productos está el yogur (El Semanario, enero 2009). Una de las empresas productoras de alimentos, líder en el mercado, realizó un estudio cuyo propósito fue determinar si los consumidores de diferentes tipos de yogur difieren en su perfil psicográfico y demográfico. De ser el caso, la empresa diseñará estrategias de promoción y venta diferenciadas por tipo de producto ya que actualmente promueve el yogur de manera genérica.

Cuatro tipos de yogures fueron considerados: yogur batido normal, yogur batido reducido en grasa, yogur bebible y yogur funcional. Veinticinco consumidores regulares de cada tipo de yogur (que consumen yogur al menos tres veces a la semana y al menos en dos ocasiones un mismo tipo sin importar el sabor o la marca) fueron encuestados. La información recolectada para cada consumidor incluye a las siguientes variables:

X1 = habitualmente compro nuevos productos para probar (1= total desacuerdo a 5 = total acuerdo)

X2 = tengo una dieta muy saludable (1 = total desacuerdo a 5 = total acuerdo)

X3 = me fijo mucho en la calidad de un producto antes de comprarlo (1 = total desacuerdo a 5 = total acuerdo)

X4 = cuando compro un producto busco que tenga buen precio (1 = total desacuerdo a 5 = total acuerdo)

X5 = práctica de alguna actividad física o deporte (1 = menos de 1 vez al mes a 5 = al menos 3 veces por semana)

X6 = edad (en años)

X7 = ingreso familiar neto (escala de 8 categorías según niveles socioeconómicos D, C-, C, C+, B-, B, B+ y A) mayor valor de la escala es mayor categoría de ingreso).

En la Tabla 5.1 se reportan los coeficientes estandarizados de los tres discriminantes que es posible calcular para el caso de g = 4 grupos de consumidores de yogur; los eigen-valores asociados y la contribución porcentual de cada discriminante a la separación total entre los cuatro grupos; más el estadístico de prueba?-Wilks, entre paréntesis está el valor P asociado a la prueba.

Variable	Primer discriminante	Segundo discriminante	Tercer discriminante
Prueba de nuevos productos	-0.044	0.102	0.681
Dieta saludable	-0.723	0.697	0.078
Calidad	0.074	0.288	-0.048
Precio	0.387	0.158	-.528
Actividad física	0.740	0.061	0.379
Edad	0.163	-0.585	0.446
Ingreso	0.172	0.719	0.125
Eigenvalor	4.35	3.41	1.07
% de varianza	49.26	38.62	12.12

Tabla 5.1 Discriminantes lineales para grupos de consumidores de yogur

Elaboración a partir del listado de SPSS

A partir de la información en la Tabla 5.1 se concluye que la separación entre los cuatro grupos se podría hacer con únicamente dos discriminantes lineales ya que el tercer discriminante contribuye únicamente con un 12.12% a la separación entre grupos.

Actividad de repaso

5. Análisis discriminante multivariado

Revisa la información, contesta las preguntas.

Entre las capacidades críticas que permiten a una empresa mantener su posición competitiva están las siguientes tres: capacidades de innovación, de relación y de absorción (la habilidad de la empresa para identificar, obtener y utilizar el conocimiento externo en su beneficio) . Un estudio entre empresas que operan en la zona industrial del centro de México tuvo como objetivo demostrar que estas capacidades distinguen realmente a las empresas exitosas (G1 = buena posición financiera y perspectivas de crecimiento) de las no-exitosas (G2 = situación financiera problemática y estancamiento). Las capacidades de las empresas se midieron como índices entre 0 y 100 puntos, un mayor valor del índice corresponde a mejores capacidades.

A partir de la información recolectada entre 40 empresas, 20 de cada grupo, se calculó el discriminante lineal propuesto por Fisher, el cual es igual a Y = 0.452 innovación + 0.311 relación + .543 absorción.

a) ¿Cuál de los tres tipos de capacidades discrimina más entre empresas exitosas y no-exitosas?

b) Si los centroides de los dos grupos son: probar que la función discriminante separa significativamente al grupo de empresas exitosas del de no-exitosas.

Respuestas:

a) Capacidades de absorción porque tiene el mayor coeficiente en la combinación lineal Y

b) Es necesario calcular el estadístico F presentado en (4). Para lo anterior es necesario calcular la distancia de Mahalanobis que es

igual a la diferencia entre los puntajes discriminantes promedio por grupo.

$$D2 = 18.564,\ F = 58.6232 \text{ y } RR = \{F>F(.05, 3, 36 = 2.84\}.$$

La hipótesis de separación insatisfactoria entre los grupos se rechaza, con lo que se concluye que a partir de las tres variables se logra una buena discriminación para las dos categorías de empresas.

5.2 Selección de variables en la función discriminante lineal

Además de evaluar hasta donde la FDL permite distinguir o discriminar entre los grupos, es también relevante determinar cuáles de las variables que se propusieron para discriminar realmente contribuyen a distinguir los grupos. Por ejemplo en el diagrama de la Figura 5.3, la primera variable X1 no permite diferenciar a los grupos puesto que los promedios de los grupos respecto a esta variable casi coinciden. En contraste, para la variable X2 ambos grupos difieren de manera importante.

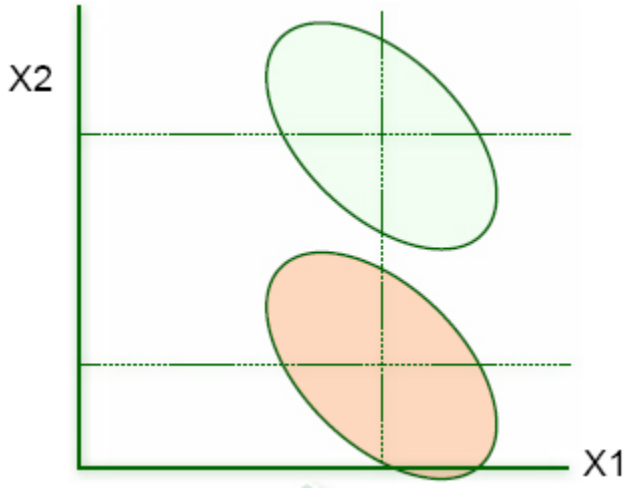


Figura 5.3 Discriminación entre dos grupos a través de una única variable

Dada una situación como la anterior, otra tarea importante en ADM es identificar aquellas variables que son buenos discriminantes y eliminar de la función discriminante lineal a aquellas variables que no contribuyen significativamente a la separación de los grupos (Perreault, Behrman y Armstrong, 1979). Para determinar cuáles variables utilizar para la discriminación o la clasificación lo que procede es establecer si las medias de los grupos difieren significativamente entre sí; tal comparación requiere de una prueba objetiva que pueda aplicarse para cualquier número de variables discriminantes ya que la comparación visual a través de un gráfico como el de la Figura 5.3 es subjetiva y únicamente viable si p=2 ó 3.

La hipótesis a probar se formula en seguida y requiere de probarse para cada una de las p variables discriminantes que incluye el vector $\boldsymbol{x}$, haciendo referencia la k-ésima variable donde k=1,2,…, p se tiene:

$$H_0 : \mu_{1k} = \mu_{2k} = ... = \mu_{gk} \quad vs. \quad H_1 : \mu_{ik} \neq \mu_{jk} \ para\ al\ menos\ una\ i \neq j$$

Si la hipótesis nula no se rechaza significa que las medias de los g grupos no difieren significativamente con respecto a la variable X_k. Por el contrario, si la hipótesis nula es rechazada, la conclusión es que hay al menos dos grupos cuyas medias son estadísticamente diferentes en relación a la variable X_k, esto implica que la variable permite discriminar o distinguir a los grupos.

La hipótesis anterior, para el caso de g= 2 grupos, es una hipótesis básica que bajo el supuesto de normalidad, independencia y varianza constante puede probarse utilizando como estadístico de prueba a la t-Student (Malhotra, 2008, p. 480). En el caso de g> 2 grupos es necesario realizar un Análisis de Varianza de un factor (Malhotra, 2008, p. 506), los detalles de este análisis se proporcionan en el capítulo 6 de este eBook. Sin embargo, como el uso de los métodos multivariantes siempre se realiza con apoyo computacional, no es necesario conocer los detalles numéricos del ANOVA de un factor para establecer las conclusiones necesarias ya que el software estadístico proporcionará el valor del estadístico de

prueba *F* y el valor *P* correspondiente. Estas cantidades se interpretan en una forma similar que en el caso del ANOVA para regresión que se discutió en el capítulo anterior.

La prueba de hipótesis anterior permite hacer un "cernido" del conjunto de variables y elegir únicamente aquellas que son útiles para separar y clasificar. La realización del ANOVA de un factor para cada variable discriminante es equivalente al análisis de correlación descrito en el capítulo 4 y cuyo propósito fue también identificar aquellas variables potencialmente útiles para explicar una respuesta de interés. El ADM es otro método de dependencia multivariante pero a diferencia del análisis de regresión, en el caso de discriminación la variable "dependiente" no es una variable métrica, de escala de intervalo ó razón, sino que es un identificador del grupo al que pertenecen las observaciones. Es decir que, mientras el vector de variables "independientes" o discriminantes ***x*** incluye solo variables métricas, la respuesta "*Y*" es una variable en escala nominal con 1,2,…*g* categorías. Otro criterio que se ha sugeri do para identificar las variables discriminantes críticas es el índice de potencia discriminante relativo que es similar a la comunalidad de una variable en el análisis factorial (Perreault et al., 1979). Más que este índice, lo que se utiliza para corroborar la importancia que cada variable tiene en la discriminación son las correlaciones de cada variable con el discriminante lineal, estas correlaciones son proporcionales a los coeficientes de las variables en el discriminante y son parte de los listados de salida de los paquetes estadísticos que tienen implementado el ADM.

Ejemplo 3. La aportación económica del sector servicios en México exhibe una tendencia creciente desde hace varios años, el sub-sector turístico es uno de los de mayor impacto. Entre los destinos más favorecidos por el turismo nacional están las playas, que son promovidas por las agencias turísticas mediante diversos paquetes. Un consorcio de agencias turísticas de la zona centro de México hace promoción de paquetes de playa antes del inicio de cada temporada vacacional.

Muchos de los clientes a quienes se remiten folletos y se les hacen llamadas promocionales finalmente no concretan una compra. Con el objetivo de dirigir mejor sus esfuerzos promocionales, el consorcio solicitó la realización de une estudio para determinar cuál es el perfil del cliente más proclive a responder a la promoción de paquetes de playa para así promoverlos únicamente entre quienes están más interesados en la compra. Cincuenta clientes fueron clasificados como: 1 = respondieron a la última promoción de paquetes de playa (21) y 2 = no respondieron (29). La información recolectada para cada cliente incluyó a las siguientes variables:

X1 = Ingreso familiar mensual neto (en miles de pesos)

X2 = Edad del jefe de familia

X3 = Número de integrantes en la familia

X4 = Cantidad gastada durante las últimas vacaciones familiares

X5 = Promedio de vacaciones anuales según las salidas registradas en los últimos cinco años

X6 = Importancia asignada a una oferta integral para las vacaciones (hotel + transporte + entretenimiento) en una escala de 0 = nada importante a 10 = totalmente importante

En este caso hay únicamente dos grupos: clientes que respondieron a la promoción y clientes que no respondieron. Entonces, la comparación entre las medias de estos grupos puede realizar se empleando la prueba t-Student para muestras independientes, el lector recordará haber discutido los detalles de esta prueba en un curso de estadística básica. La comparación de dos medias es una de las varias pruebas disponibles en la opción Basic Statistics de MINITAB, la secuencia de comandos es la siguiente:

Una vez abierta la ventana de diálogo, basta con indicar en cuáles columnas se encuentran codificados los datos para los dos tipos de empresa (subscripts) y los datos para cada uno de los indicadores propuestos como discriminantes (samples). En la Tabla 5.2 se resumen los resultados de esta prueba de hipótesis para la comparación entre dos medias.

Variable	t-Student	Valor P	Conclusión
Ingreso familiar	-3.59	0.000	Clientes que responden tienen un mayor ingreso familiar
Edad jefe de familia	2.30	0.013	Clientes que responden son más jóvenes que no-respondientes
Número integrantes familia	-0.88	0.192	No hay diferencias
Cantidad gastada últimas vacaciones	-2.45	0.008	Clientes que responden gastaron más en la última vacación
Promedio de vacacio-nes por año	-2.99	0.002	Clientes que responden toman más vacaciones por año
Importancia oferta integral	-2.57	0.006	Clientes que responden asignan más importancia a oferta integral

Tabla 5.2 Comparación entre clientes que respondieron y no respondieron a la promoción de paquetes de playa.

Elaboración a partir del listado de MINITAB

Las comparaciones entre los grupos respecto a las seis variables elegidas para definir el perfil del cliente llevan a concluir que, excepto por el número de integrantes en la familia, todas las variables permiten diferenciar a los grupos. Las pruebas de hipótesis son, por lo tanto, un análisis preliminar que permite establecer cuántas y cuáles variables son necesarias para alcanzar una buena clasificación de los clientes e implementar las promociones dirigidas.

Otra opción para la selección de variables críticas para discriminar entre grupos y/o clasificar individuos en grupo es emplear el procedimiento **Stepwise**. En el caso de ADM este procedimiento subóptimo de selección de variables opera de manera similar que en el caso del análisis de Regresión Lineal Múltiple; la decisión de entrar o eliminar una variable se basa en el potencial que ésta tiene para discriminar entre los grupos.

El problema principal cuando se utiliza este procedimiento automático de selección de variables es que no hay garantía de que se elija el mejor subgrupo de “discriminantes”, en particular cuando hay colinealidad entre las variables discriminantes. Para definir cuáles variables se incluirán, o eliminarán, en la función discriminante lineal se utiliza el estadístico lambda de Wilks (Λ-Wilks). Una vez que se tiene un discriminante lineal con al menos dos variables, el estadístico Λ-Wilks para cada variable representará qué modelo reducido -en una única variable si únicamente hay dos discriminantes-se obtendrá si se elimina a la variable. Entre más importante sea la variable para discriminar (clasificar) entre los grupos mayor será el valor Λ-Wilks ya que al eliminarse a la variable del discriminante lineal se deteriora su capacidad para distinguir entre los grupos. El estadístico Λ-Wilks es, por lo tanto, el criterio estadístico para determinar hasta dónde una función discriminante lineal permite una buena separación o distinción entre grupos. En

las secciones 5.3 y 5.6 se amplía la información sobre este estadístico y se presentan ejemplos de su uso.

5.3 Evaluación de la calidad de la función discriminante lineal

En el caso de un método conclusivo como el ADM lineal es posible realizar pruebas estadísticas formales para verificar si los discriminantes lineales contribuyen significativamente a la separación o variabilidad entre los grupos. Según se describió en la sección anterior, el estadístico de prueba que se utiliza para probar la significancia de la función discriminante lineal es la lambda de Wilks (Λ-Wilks), este estadístico calcula como en (8)

$$\Lambda = \frac{Suma\ de\ cuadrados\ entre\ grupos}{Suma\ de\ cuadrados\ total} = \frac{|B|}{|B+W|} \qquad (8)$$

Cuando hay s>1 discriminantes lineales, la Λ-Wilks se utiliza para probar si los s discriminantes contribuyen significativamente a la separación entre grupos. El procedimiento, en el caso del paquete estadístico SPSS es de la siguiente manera:

i) Se determina si los discriminantes 1,2,…,s son significantes

ii) Se elimina sucesivamente a uno de los discriminantes y se evalúa si con los restantes aún se logra una distinción significante entre grupos. Por ejemplo, se elimina al primer discriminante y se procede a determinar si los discriminantes lineales restantes 2,…,s permiten aún separar a los grupos. Después se elimina tanto al primer como segundo discriminante y se prueba si los discriminantes lineales 3,…,s aún distinguen entre los grupos. El último valor de Λ-Wilks se asocia al último (s) discriminante lineal. Similar al caso de Λ-Wilks en el procedimiento Stepwise de selección de variables, un valor mayor para el estadístico significa que al eliminarse al discriminante lineal se reduce la capacidad de distinción o separación entre los grupos.

Ejemplo 4. En referencia al Ejemplo 2 donde se describe el proyecto de investigación de mercados realizado con el propósito de perfilar los segmentos de los consumidores de yogur de varios tipos (batido regular, batido bajo en grasa, líquido y funcional) se procedió a probar la significancia de los tres discriminantes lineales. Los resultados de la prueba estadística se reportan a continuación, el valor en cada celda es la lambda de Wilks y el número en paréntesis el valor P correspondiente.

Prueba para las funciones	**(1 hasta 3)**	**(2 hasta 3)**	**(3)**
	.187(0.000)	.405(0.000)	.859(.201)

Con base en la tabla anterior se concluye que con únicamente dos discriminantes puede lograrse una separación significativa de los grupos. Cuando el primer discriminante se elimina, el valor de la Λ-Wilks crece debido a que este discriminante es el que más contribuye a la separación. Si tanto el primero como el segundo discriminante son eliminados, la Λ-Wilks ya no es significante (P = 0.201) lo que indica que el tercer discriminante lineal no aporta significativamente a la variabilidad entre los grupos de consumidores de yogur.

5.4 Reglas de clasificación

El propósito de la clasificación es asignar nuevos individuos a los grupos especificados en términos de sus características, descritas en el vector multivariante **x**. En la construcción de una regla de clasificación se busca minimizar la diferencia entre la pertenencia observada de los individuos a los grupos y la asignada mediante una regla de clasificación;, el objetivo general para una **función** de clasificación es minimizar alguna medida de la "pérdida" asociada a clasificaciones erróneas, es decir cuando los individuos son asignados incorrectamente. Bajo la perspectiva de teoría de decisiones, una regla que minimiza la pérdida esperada por mala asignación se denomina **Regla Bayesiana** o Regla óptima.

En el desarrollo de un método de clasificación se aprovecha la información que se tiene sobre los individuos para revisar las probabilidades de asignación a priori, es decir la probabilidad de que un individuo pertenezca a cierto grupo sin información adicional excepto por el conocimiento sobre la distribución de los individuos en los grupos. El problema de clasificación se formula como sigue: sean $f_1(\mathbf{x})$ y $f_2(\mathbf{x})$ las funciones de densidad de probabilidad asociadas con el vector p-dimensional **x** de los grupos 1 y 2 respectivamente. Dadas las regiones de clasificación R_1 y R_2, las probabilidades de clasificación incorrecta son:

$$\Pr(\textit{asignar a } G_2 \mid G_1) = \int_{R_2} f_1(x)dx \quad y \quad \Pr(\textit{asignar a } G_1 \mid G_2) = \int_{R_1} f_2(x)dx$$

el propósito es definir las regiones R_1 y R_2 de manera que, según se dijo antes, se minimice alguna medida de la pérdida asociada con la clasificación incorrecta de los individuos. La forma más simple para esta medida es el total de clasificaciones incorrectas (TCI), esto es:

$$\text{TCI} = \Pr(\text{total de clasificaciones incorrectas}) = \Pr(G_2 \mid G_1) + \Pr(G_1 \mid G_2) \qquad (8)$$

La Figura 5.4 describe gráficamente el caso particular de clasificación con una única variable, el área sombreada bajo la curva

de cada función de densidad representa la probabilidad de la asignación incorrecta de un individuo; por Ejemplo el área identificada como $R(1|2)$ corresponde a la probabilidad de que un individuo del grupo 2 se localice en la región 1 y por lo tanto sea asignado, erróneamente, al primer grupo. El propósito que establece la expresión (8) es minimizar el área total mediante la selección del punto que define a las regiones R_1 (asignar al grupo 1) y R_2 (asignar al grupo 2).

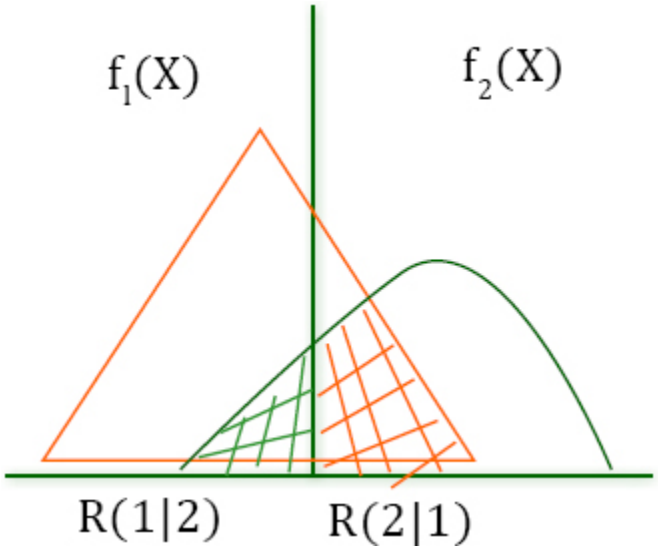


Figura 5.4 Representación gráfica de la probabilidad total de clasificaciones incorrectas

Es posible demostrar (ver Johnson y Wichern, 1998) que bajo el criterio de minimizar el total de clasificaciones incorrectas una nueva observación $\mathbf{x}_0$ será asignada a las regiones R_1 y R_2 de la siguiente manera:

$$R_1\text{: Clasificar en grupo 1 si } f_1(\boldsymbol{x}_0) / f_2(\boldsymbol{x}_0) \geq p_2 / p_1$$
$$R_2\text{: Clasificar en grupo 2 si } f_1(\boldsymbol{x}_0) / f_2(\boldsymbol{x}_0) \geq p_2 / p_1 \qquad (9)$$

donde p_1 y p_2 son las probabilidades de pertenencia a priori para los grupos 1 y 2. Estas probabilidades pueden determinarse a través de información disponible sobre la estructura de los grupos en la población de interés. Si se toma como ejemplo el caso de los grupos

de consumidores de yogur, suponer que el producto que contribuye más a las ventas es el yogur batido normal cuya participación en el mercado de yogur se estima en 40%, el yogur bebible está en segundo lugar con una participación del 30%, el batido bajo en grasa tiene un 20% de participación y el restante 10% es del yogur funcional. Por lo tanto la probabilidad a priori de que un consumidor cualquiera de yogur prefiera el yogur bebible sería de 0.3. Notar que estas probabilida des no consideran el perfil del individuo sino únicamente la información a priori sobre la estructura de mercado de los productos de yogur.

Otra medida alternativa para la "pérdida" asociada con una clasificación incorrecta resulta de considerar los costos de la mala clasificación. En términos de esta medida el problema de clasificación se formula como definir las regiones R_1 y R_2 que minimicen el costo (total) esperado por una mala clasificación (CEM), esto es:

$$\text{CEM} = p_1\, c_{2|1}\, Pr(\text{asignar a } G_2 |\, G_1) + p_2\, c_{1|2}\, Pr(G_1 |\, G_2) \quad (10)$$

Donde $c_{2|1}$ y $c_{1|2}$ son respectivamente los costos de asignar incorrectamente a individuos del grupo 1 al grupo 2 y de asignar a individuos del grupo 2 al 1 por error. La regla que minimiza CEM, clasificará a una nueva observación x0 en las regiones R_1 y R_2 de la siguiente manera:

$$R_1\text{: Clasificar en grupo 1 si } f_1(\boldsymbol{x}_0)/f_2(\boldsymbol{x}_0) \geq p_2\, c_{1|2}/p_1\, c_{2|1}$$

$$R_2\text{: Clasificar en grupo 2 si } f_1(\boldsymbol{x}_0)/f_2(\boldsymbol{x}_0) \geq p_2\, c_{1|2}/p_1\, c_{2|1} \quad (11)$$

Las soluciones anteriores al problema de clasificación requieren de especificar las funciones de densidad $f_1(\mathbf{x})$ y $f_2(\mathbf{x})$. Si estas funciones se asumen conocidas, el problema estadístico es especificar los parámetros de estas funciones de densidad, por lo

tanto el desarrollo de la regla de clasificación se hace desde un enfoque paramétrico. Cuando las funciones de densidad de los grupos no se especifican, resultan reglas de clasificación no-paramétricas, bajo este enfoque una de las reglas más conocidas se basa en el uso del método kernel para estimar las funciones de densidad de los grupos. El método está disponible en el software comercial SAS (Statistical Analysis System).

Si se asume que las funciones de densidad son normales multivariadas con matrices de varianza-covarianza, la regla de clasificación para el caso de dos grupos con probabilidades y costos de clasificación incorrecta diferentes resulta ser (12)

Asignar al grupo 1 cuando

$$(\bar{x}_1 - \bar{x}_2)' S_p^{-1} x_o - 1/2(\bar{x}_1 - \bar{x}_2)' S_p^{-1}(\bar{x}_1 + \bar{x}_2) \geq \ln \frac{c(1|2)p_2}{c(2|1)p_1} \qquad (12)$$

La regla de clasificación anterior se conoce como *Regla de Clasificación de Anderson* (1958). Esta regla de clasificación es equivalente a clasificar con la función discriminante lineal de Fisher, según se describió al final de la sección 5.1, cuando las probabilidades a priori y los costos por mala clasificación son iguales. Este importante resultado permite apreciar por qué los problemas de discriminación y clasificación están tan cercanamente interrelacionados.

Es relevante indicar que cuando se clasifican individuos en $g>2$ grupos, si las probabilidades a priori y los costos de clasificación incorrecta son iguales para todos los grupos, la regla de clasificación paramétrica óptima coincide también con el método de discriminantes lineales propuesto por Fisher, lo que reitera la pérdida de distinción entre los problemas de discriminación y clasificación (Arroyo, 1997).

Si las distribuciones de probabilidad de los grupos son conocidas, la asignación de un nuevo elemento al *i*-ésimo grupo se realiza de

acuerdo a sus probabilidades a posteriori. El lector recordará haber discutido este tema en su curso básico de estadística. Esta probabilidad a priori se entiende como una probabilidad revisada después de tomar en cuenta la información adicional sobre las características de los individuos que se van a clasificar por lo cual se representa como $p_i(G_i|x)$ para hacer explícito que la asignación dependerá de las variables discriminantes en el vector??. Si se emplean estas probabilidades, un nuevo individuo será asignado a aquel grupo para el cual su probabilidad a posteriori es la mayor, bajo el supuesto de normalidad esto implica asignar al individuo al grupo *i* cuando $\max_i \ln p_i f_i(x)$. Esto es equivalente a asignar al individuo al *i*-ésimo grupo si su score discriminante lineal di(x) es el mayor en el grupo *i*. Este score discriminante se determina como sigue:

$$d_i(x) = \bar{x}_i' S_p^{-1} x_0 - \frac{1}{2} \bar{x}_i' S_p^{-1} \bar{x}_i + \ln p_i \qquad (13)$$

Donde S_p es la matriz de varianza-covarianza que resulta de combinar la información de todos los grupos ya que se asume que la verdadera matriz Σes la misma para todos. Cuando se utiliza software estadístico (por ejemplo MINITAB), estos puntajes se calculan fácilmente a partir de las funciones discriminantes lineales entre grupos. No confundir estas funciones con los discriminantes lineales, mientras que estas últimas son combinaciones lineales que separan a los grupos, las funciones discriminantes lineales definen los puntajes a partir de los cuales se clasifican o asignan individuos a los grupos.

Cuando las matrices de varianza-covarianza de los grupos no son iguales, los puntajes discriminantes dados en (13) se modifican de la siguiente forma:

$$d_i^Q(x_o) = -\frac{1}{2} \ln |S_i| - 1/2 (x_o - \bar{x}_i)' S_i^{-1} (x_o - \bar{x}_i) + \ln p_i \qquad (14)$$

Notar que en la expresión (14) la matriz de varianza-covarianza combinada S_p se sustituye por la matriz correspondiente para cada grupo. La regla de clasificación anterior se conoce como regla discriminante cuadrática; de acuerdo con (14) un nuevo individuo con perfil de características x_0 se asigna al grupo i si su puntaje discriminante cuadrático es el máximo para el grupo i en relación con los otros grupos.

La evaluación de una regla de clasificación se realiza en términos de la medida de error asociada con las clasificaciones incorrectas. Para reglas como (13) y (14), cuyo propósito es minimizar la tasa de clasificaciones incorrectas, la calidad de su desempeño se evalúa al comparar la tasa de clasificaciones correctas alcanzada con la tasa de clasificaciones correctas que se obtendría si se clasifica al azar, esto es, empleando únicamente las probabilidades a priori. Si los mismos datos que se emplearon para construir la función de clasificación son asignados a los grupos, a través de las reglas (13) ó (14), se obtendrá la tasa de error aparente (APER por sus siglas en inglés, APparent Error Rate). Esta tasa de error y su complemento, la tasa de clasificaciones correctas aparente, reciben esta denominación porque son estimados "optimistas" del verdadero desempeño de la función de clasificación. Esto significa que tienden a sobre-estimar la calidad de la función (McLachlan, 1992).

Para calcular estas tasas se construye una matriz de clasificaciones como la que se muestra en la Tabla 5.3, para el caso de dos grupos.

Verdadero grupo de pertenencia	**Asignado a grupo 1**	**Asignado a grupo 2**
1	A	B
2	C	D
Número de datos	n1	n2

Tabla 5.3 Matriz de clasificaciones observadas y asignadas

A partir de la matriz anterior se calcula la tasa de clasificaciones correctas la cual es igual a (+ *d*)/(*n*1 + *n*2).

Un estimador insesgado de la verdadera tasa de error o de clasificaciones correctas puede obtenerse a través del procedimiento de "uno fuera" (Leave-one-out Error Rate) (McLachlan, 1992). Este método de estimación consiste en eliminar una observación del conjunto de datos y construir la regla de clasificación con los *N*-1 datos restantes. Con esta regla, basada en *N*-1 datos, se clasifica el dato eliminado; el proceso se repite hasta que todas las observaciones se hayan clasificado después de eliminarlas sucesivamente del conjunto de datos. El procedimiento involucra un considerable número de cálculos sin embargo, con los recursos computacionales actuales es posible completarlos rápidamente. Otras opciones que producen estimados insesgados de las tasas de error y de clasificaciones correctas son el método de "Jacknife" -que consiste en eliminar *k* observaciones en lugar de solo una del conjunto de datos- y de "Bootstrap" -que consiste en generar m muestras aleatorias de tamaño *N* con reemplazo a partir de las *N* observaciones. Para ambos métodos con cada muestra generada, ya sea al eliminar *k* observaciones o elegida al azar, se procede a construir la regla discriminante y a calcular APER. El promedio de las tasas de clasificaciones incorrectas de todas las muestras es un estimador insesgado de la tasa de error.

Una vez que se ha determinado la tasa de clasificaciones correctas, ésta se compara con el criterio de asignación proporcional por azar $C_{pro} = \sum_{i=1}^{k} p_i^2$ Una regla de clasificación se califica como de mal desempeño cuando su tasa de clasificaciones correctas no excede considerablemente el valor de Cpro. Para formalizar la evaluación de una regla de clasificación derivada bajo el supuesto de normalidad, lo más apropiado es probar la siguiente hipótesis:

H_0: La tasa de clasificaciones correctas que produce la regla de clasificación es similar a la obtenida si se clasifica al azar, versus

H_1: La tasa de clasificaciones correctas que produce la regla de clasificación es mejor a la obtenida si se clasifica al azar.

El estadístico de prueba correspondiente se denomina *Q* de Press para evaluar el desempeño de la regla de clasificación y se calcula según la expresión (15).

$$Q = \frac{(N - ng)^2}{N(g-1)} \approx \chi^2(g-1) \qquad (15)$$

donde *N* = total de datos y *n* = total de clasificaciones correctas

Ejemplo 5. Respecto al Ejemplo 3 que presenta el problema de clasificar a los clientes de las agencias de servicios turísticos con el propósito de realizar promociones dirigidas para maximizar la tasa de respuesta, compra, a la promoción de paquetes de playa, se utilizó una función de clasificación lineal para asignar a los clientes a los grupos de respondientes y no-respondientes. Los cálculos correspondientes se realizaron usando el paquete estadístico MINITAB, que implementa tanto las reglas de clasificación lineal (13) como cuadrática (14). MINITAB no realiza el cálculo de discriminantes lineales ya que está enfocada directamente al problema de clasificación.

La secuencia de comandos de MINITAB para trabajar clasificación es:

La ventana de diálogo que se abre es similar a la que se muestra en la Figura 5.5.

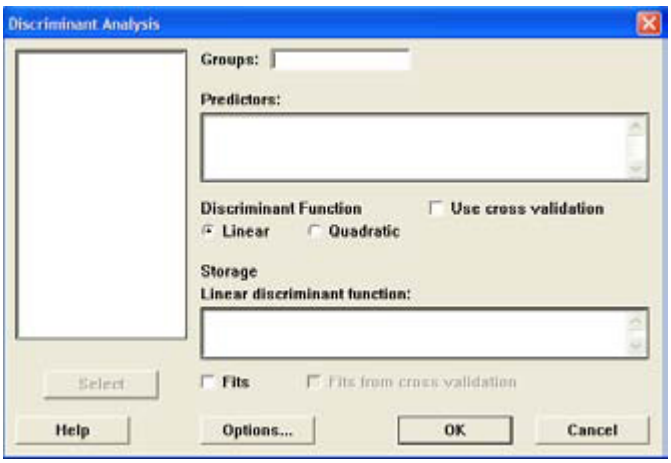


Figura 5.5 Ventana de diálogo para la clasificación con MINITAB

En el recuadro Groups hay que declarar la columna que contiene los datos de pertenencia de los individuos; las variables discriminantes se listan en la ventana de Predictors. Es posible elegir entre una **función de clasificación lineal** (el default) o cuadrática seleccionando el recuadro correspondiente. Para calcular un estimado insesgado de la tasa de clasificaciones correctas hay que elegir la opción Use cross validation y para declarar los valores de las probabilidades a priori es necesario utilizar el botón de Options para abrir una ventana de diálogo adicional como la que se muestra en la Figura 5.6.

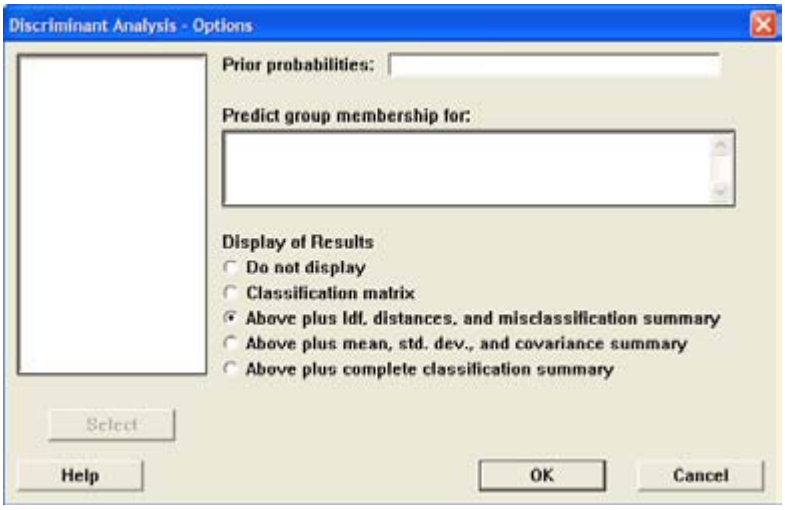


Figura 5.6 Ventana de Opciones para la clasificación con MINITAB

Con base en la información histórica disponible para este Ejemplo, 21/60 = 35% de los clientes responden a una promoción, entonces las probabilidades a prior son de .35 .65 (ingresar en MINITAB separadas por un espacio). En el recuadro con la leyenda *Predict group membership* pueden anotarse los datos para un nuevo individuo, el programa calculará las probabilidades de pertenencia a posteriori y asignará al individuo al grupo que corresponda. El listado de salida se presenta a continuación, note que la variable "número de integrantes de la familia" no se utilizó para la clasificación puesto que previamente se demostró que no difiere entre grupos.

Summary of Classification with Cross-validation

	True Group	
Put into Group	1	2
1	19	9
2	2	20
Total N	21	29
N correct	19	20
Proportion	0.905	0.690

N = 50 N Correct = 39 Proportion Correct = 0.780

Linear Discriminant Function for Groups

	1	2
Constant	-101.97	-177.30
Ingreso	0.24	0.80
Promedio vacaciones anuales	-4.40	3.19
Edad jefe familia	4.22	3.21
Gasto última vacación	0.00	0.02

	1	2
Importancia oferta integral	8.79	10.92

En la primera parte del listado se reporta la matriz de clasificaciones obtenida a través de validación cruzada. En este Ejemplo el valor de Cpro es de $.35^2 + .65^2 = 0.545$ mientras que la tasa global de clasificaciones correctas según el listado de MINITAB es de (19+20)/50 = 0.78 la cual excede considerablemente el valor de Cpro, de donde la regla de clasificación mejora la clasificación al azar.

Para formalizar esta observación se calcula el estadístico de prueba *Q* de Press, el cual resulta igual a $Q = \frac{(N - ng)^2}{N(g-1)} \approx \chi^2(g-1)$. La región de rechazo para un nivel de significancia del 5% se establece como $RR = \{Q > \chi^2_{.05}(1) = 3.843\}$, el valor de *Q*-Press calculado lleva a concluir que la función de clasificación empleada resulta en un mayor número de clasificaciones correctas que si se utiliza una asignación aleatoria de clientes a grupos.

La segunda parte del listado reporta las funciones de clasificación lineales calculadas según la expresión (13). Si los datos de un nuevo individuo se sustituyen en cada una de estas funciones, el individuo será asignado al grupo donde obtenga un mayor puntaje o score discriminante.

¿Sabías qué?

El análisis discriminante juega un papel muy importante en los programas de reconocimiento facial de empleados en investigaciones criminales y forenses.

Actividad de repaso

5. Análisis discriminante multivariado

RELACIONAR COLUMNAS

Se denomina así a aquella combinación lineal de las variables "independientes" que separa más entre g grupos ya conocidos.	○ ○	Eigenvalor
Regla de clasificación óptima bajo los supuestos de normalidad, homogeneidad de varianza y costos de clasificación incorrecta iguales.	● ●	Redes neuronales
Es la suma de los cuadrados de las probabilidades a priori y sirve como referencia para comparar la calidad de una regla de clasificación.	○ ○	Primer discriminante
Es la tasa de clasificaciones incorrectas producidas por una regla de clasificación, se calcula al clasificar los mismos datos que se emplearon para construir la regla.	● ●	Uno fuera
Regla paramétrica de clasificación que asume normalidad pero relaja el supuesto de homogeneidad de varianza.	○ ○	Matriz de confusión
Procedimiento no-paramétrico para la clasificación de objetos basado en conceptos de inteligencia artificial.	● ●	Regla de Anderson
A través de esta cantidad se determina la contribución que cada discriminante lineal hace a la variabilidad (separación) entre grupos.	○ ○	Análisis discriminante métrico
Método de estimación de la tasa de clasificaciones correctas (incorrectas) que consiste en eliminar un dato a la vez para asignarlo a un grupo usando la regla de clasificación.	● ●	Tasa de error aparente
Método de dependencia multivariado en el cual las variables independientes están en escalas de intervalo o razón y la dependiente en una escala nominal.	○ ○	Cpro
Es un arreglo bi-dimensional que muestra la clasificación observada de los datos contra la asignación obtenida al aplicar una regla de clasificación	● ●	Regla de clasificación cuadrática

Actividad de repaso (1)

5. Análisis discriminante multivariado

Instrucciones: Realiza las operaciones y da clic en el botón de Respuesta para revisar las alternativas propuestas por los autores

En el proyecto de segmentación para consumidores de yogur se consideraron como alternativa a los siguientes grupos: grupo 1= alto consumo de yogur, al menos tres veces por semana, grupo 2 = consumo intermedio de yogur, 1-2 veces por semana, y grupo 3= consumo infrecuente de yogur, menos de una vez por semana. Un subconjunto de las características demográficas y psicográficas de los consumidores, las que resultaron mejores discriminantes, fueron utilizadas para clasificar a los consumidores en estos grupos definidos en términos de la intensidad de consumo del producto. Asumiendo probabilidades a priori de .2, .3 y .5 se obtuvieron los siguientes resultados para la clasificación de individuos.

Grupo de clasificación	Grupo verdadero de pertenencia		
Alto	21	3	0
Intermedio	3	36	11
Infrecuente	1	11	14
Total de datos	25	50	25
Datos clasificados correctamente	21	36	14
Porcentaje de clasificaciones correctas	84%	72%	56%

Siguiente

5.5 Aplicación del análisis discriminante lineal con apoyo computacional

El siguiente problema de aplicación del análisis discriminante multivariante se resolverá con el apoyo del software estadístico SPSS ya que el paquete MINITAB, que es el utilizado a lo largo de este eBook, no realiza el cálculo de discriminantes lineales. Para mayores detalles sobre cómo utilizar SPSS para discriminación y clasificación se recomienda consultar el ejemplo detallado que presentan Calvos-Silvosa y Rodríguez-López (2003).

Ejemplo 5. Para que las empresas mexicanas puedan realizar la exportación de sus productos, el gobierno mexicano a través de Bancomext ofrece créditos, a tasas preferentes, destinados a cubrir los gastos asociados con la puesta en marcha de los planes de exportación empresariales.

Para decidir que empresas son candidatas a recibir un crédito, el área de Bancomext responsable de la autorización ha propuesto que únicamente si la empresa es solvente, cuenta con un producto para el cual hay demanda en el extranjero y ha identificado los canales que le permitirán exportarlo a un costo aceptable, se le autorice el crédito.

Para que la propuesta sea autorizada, es necesario demostrar ante las autoridades de Bancomext que los indicadores sugeridos

para calificar a las empresas son justos y apropiados. Para ello se seleccionaron al azar datos de empresas a las que Bancomext apoyó económicamente en el pasado. Las empresas se calificaron como “exitosas” (después de dos años están exportando y han empezado a liquidar su crédito) y “no exitosas” (a dos años de recibir el crédito no han alcanzado su volumen meta de exportación ni han empezado a liquidar su crédito). Las solicitudes, elegidas al azar, de 25 empresas del primer grupo y 21 del segundo fueron revisadas para obtener información sobre las razones financieras que usa el score Z de Altman. Este es un índice financiero que se usa para predecir cuando una empresa se acerca a la insolvencia. Aparte de los indicadores financieros, se revisó el plan de exportación de la empresa, que fue calificado por los cinco responsables del proyecto en cuanto a su calidad. Los datos disponibles incluyen a las siguientes variables:

X1 = capital neto de trabajo contra pasivo de corto plazo

X2 = utilidades retenidas acumuladas contra activos totales

X3 = capital contable contra pasivos totales

X4 = utilidades antes de intereses e impuestos contra actividades totales

X5 = ventas contra activos totales

X6 = calificación promedio (0-10) asignada por el equipo de Bancomext al plan de exportación de la empresa.

La primera actividad del proceso de aplicación del análisis discriminante es definir aquellas variables que son importantes para distinguir entre los grupos. Si se emplea SPSS para realizar el análisis, la secuencia de comandos a utilizar es:

Analyze → Classify → Discriminant

La ventana de diálogo para declarar variables y las ventanas adicionales que permiten obtener detalles para el procedimiento se muestran en la Figura 5.7.

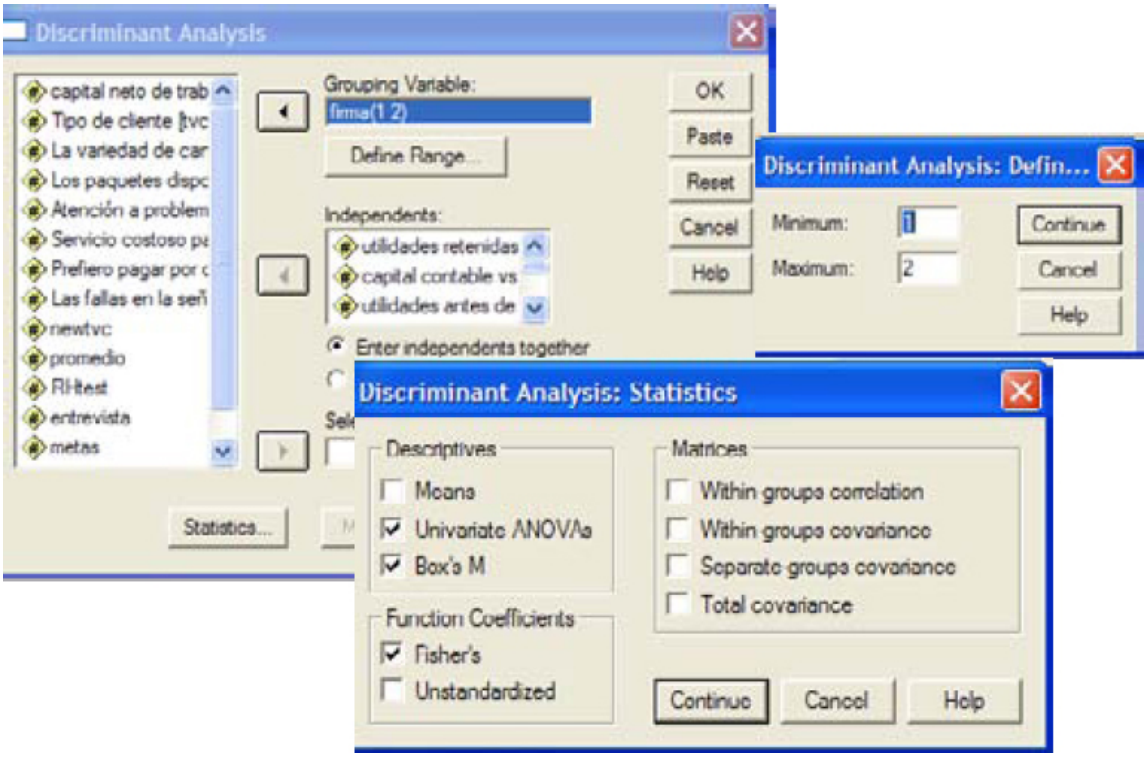


Figura 5.7 Ventana de diálogo principal de SPSS para el análisis discriminante

En el recuadro intitulado Grouping variable se declara el nombre de la variable que identifica el grupo de pertenencia de cada individuo; el número de grupos se indica con la opción Define Range, para este Ejemplo hay únicamente 2 grupos de donde *minimum* = 1 y *maximum* = 2. SPSS permite utilizar todas las variables en la construcción del discriminante lineal o bien realizar una selección automática del mejor subconjunto de variables discriminantes mediante el uso del método *Stepwise*. Si se utiliza esta última opción se debe considerar el riesgo potencial discutido en la sección 5.2 en relación a que algunas variables críticas no queden incluidas en el discriminante lineal. La recomendación en este texto es identificar aquellas variables para las cuales los grupos difieren y usar este subconjunto para el cálculo del discriminante lineal.

Para identificar las variables que son potenciales discriminantes hay que seleccionar la opción *Univariate ANOVAs* dentro de la

opción Statistics que figura en la ventana de diálogo de SPSS. Los resultados de este análisis se resumen en la Tabla 5.4.

	Wilks' Lambda	F	df1	df2	Sig.
Capital neto de trabajo vs activos totales	.885	5.729	1	44	.021
Utilidades retenidas acumuladas contra activods totales	.899	4.922	1	44	.032
Capital contable vs pasivos totales	.863	6.998	1	44	.011
Utilidades antes de intereses/impuesto vs activos totales	.516	41.323	1	44	.000
Ventas vs activos totales	.762	13.433	1	44	.001
Calificación promedio al plan de exportación	.540	37.433	1	44	.000

Tabla 5.4 Prueba de igualdad entre las medias de los grupos

Para identificar cuáles variables discriminantes contribuyen significativamente a la distinción entre los grupos, SPSS utiliza como estadístico de prueba a la lambda de Wilks que, con base a como está definida, recordar expresión (8), varía entre [0;1]. Entre menor el valor de Λ-Wilks, más contribuye la variable a la discriminación. A partir de los valores de la Λ-Wilks se calculan estadísticos F que resultan más familiares y para los que es factible calcular el valor P correspondiente. De las entradas de la tabla anterior se concluye que todas las variables (X1 a X6) contribuyen significativamente a la distinción de los dos grupos de empresas. En consecuencia, las seis variables deben considerarse en el cálculo del único discriminante lineal y la construcción de la regla de clasificación requerida para calificar a los aspirantes a un crédito de Bancomext.

El siguiente paso del análisis es calcular el discriminante lineal y verificar si genera una buena distinción entre grupos. Puesto que todas las medias de todas las variables elegidas difieren de un grupo a otro, el discriminante lineal está basado en seis variables. Las porciones del listado de salida de SPSS relevantes para este paso del análisis se muestran enseguida.

Function	Eigenvalgue	% of Variance	Cumulative %	Canonical Correlation
1	4.350(a)	100.00	100.00	.021

Tabla 5.5 Eigenvalores para los discriminantes lineales

a First 1 canonical discriminant functions were used in the analysis.

En la Tabla 5.6 se reportan los resultados de la prueba sobre la capacidad del discriminante lineal para inducir una separación significante entre los grupos. El estadístico de prueba Λ-Wilks es pequeño lo que indica que el discriminante lineal separa efectivamente a los grupos. A partir del estadístico se calcula una ji-cuadrada que permite determinar la significancia de la prueba, el P-valor o nivel de significancia estimado de la prueba es de 0.000 permitiendo rechazar la hipótesis nula de que el discriminante lineal no separa significativamente a los grupos de empresas exitosas de las noexitosas.

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.187	68.761	6	.000

Tabla 5.6 Significancia del discriminante lineal a través de la lambda de Wilks

Los coeficientes estandarizados o pesos de las variables independientes -los cinco indicadores financieros y la calificación obtenida en el plan de exportación- en el discriminante lineal son las entradas de la tabla 5.7. En ausencia de fuerte colinealidad, estos coeficientes permiten identificar a los mejores discriminantes que en este caso fueron: las utilidades netas versus los activos totales (X4), la calificación promedio del plan de exportación (X5) y el capital contable de la empresa versus sus pasivos totales (X3). Los coeficientes de correlación de cada variable con el discriminante lineal son una forma alternativa para determinar cuáles variables

son los mejores discriminantes y establecer una jerarquía (Calvo-Silvosa y Rodríguez-López, 2003). La matriz de estructura de la Tabla 8 muestra a las variables ordenadas según su potencial discriminante: X4 > X5 > X2 > X3 > X1 > X6.

	Function 1
Capital neto de trabajo vs activos totales	.249
Utilidades retenidas acumuladas contra activos totales	.366
Capital contable vs pasivos totales	.529
Utilidades antes de intereses/impuesto vs activos totales	-.914
Calificación promedio del plan de exportación	.566
Ventas contra activos totales	.455

Tabla 5.7 Coeficientes canónicos estandarizados para la función discriminante

	Function 1
Utilidades antes de interes/impuesto vs activos totales	-.465
Calificación promedio del plan de exportación	.442
Ventas contra activos totales	.268
Capital contable vs pasivos totales	.191
Capital neto de trabajo vs activos totales	.173
Utilidades retenidas acumuladas contra activos totales	.160

Tabla 5.8 Matriz de estructura

Variables ordered by absolute size of correlation within function.

En la Tabla 5.9 se reportan los puntajes discriminantes promedio, esto es $\bar{y}_1 = \hat{l}'\bar{x}_1 = -2.226$ y $\bar{y}_2 = \hat{l}'\bar{x}_2 = 1.870$. La diferencia entre los dos puntajes discriminantes promedio hace evidente la distinción lograda entre los grupos.

Tipo de empresa	**Function** 1
No exitosa	-2.226
Exitosa	1.870

Tabla 5.9 Puntajes discriminantes para los centroides de los grupos

El objetivo principal en este problema es clasificar las nuevas solicitudes de crédito que presentan las empresas a Bancomext, para obtener las funciones de clasificación es necesario elegir la opción *Classify* en la ventana principal de SPSS. Abierta la ventana adicional, que se muestra en la Figura 5.8, el analista tiene dos opciones para especificar las probabilidades a priori: 1) declararlas iguales, 0.5 en este ejemplo ya que hay dos grupos, o 2) especificarlas proporcionales a la cantidad datos de cada grupo. Esta segunda opción asume que se ha utilizado un muestreo estratificado proporcional al tamaño; es decir que la proporción de datos en cada grupo es la misma que la proporción poblacional. Como para este ejemplo no se tiene información sobre las probabilidades a priori, esto es qué tan probable es que una empresa fracase o no en sus intentos de exportar, las probabilidades a priori se asumieron iguales. También es necesario que el analista elija la opción *Leave-one-out classification* para estimar las tasas de clasificación correctas adecuadamente ya que, de no elegirse esta opción, el programa únicamente reportará las tasas aparentes.

Varios diagramas que describen gráficamente la separación entre grupos pueden obtenerse a través de SPSS: 1) *Combined-groups* muestra todos datos en el espacio reducido de los discriminantes lineales; 2) *Separate-groups* muestra por separado los puntajes discriminantes de los datos para cada grupo y 3) *Territorial map* describe las regiones de clasificación para los grupos.

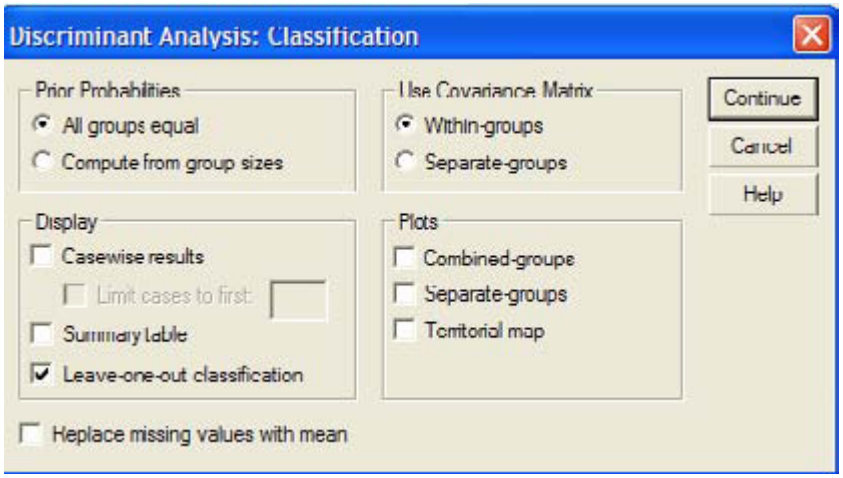


Figura 5.8 Clasificación de individuos con SPSS

El primer cuadro de resultados -después de aquellos que resumen el número y tipo de datos utilizados- que se muestran en la tabla 5.10, corresponden a las funciones de clasificación lineales descritas en la expresión (13). Cabe recordar que para clasificar a una nueva empresa en un grupo se requiere de sustituir sus datos en cada función; la empresa quedará asignada al grupo en donde obtenga el mayor puntaje.

	Tipo de empresa	
	No exitosa	Exitosa
Capital neto de trabajo vs activos totales	.249	5.823
Utilidades retenidas acumuladas contra activos totales	.366	5.354

	Tipo de empresa	
	No exitosa	Exitosa
Capital contable vs pasivos totales	.529	1.497
Utilidades antes de interes/impuesto vs activos totales	-.914	-0.48
Rating plan de exportación	.566	6.972
Ventas contra activos totales	.455	.888

Tabla 5.10 Coeficientes para las funciones de clasificación

Fisher's linear discriminant functions

La última parte del listado numérico es la **matriz de clasificación** de la Tabla 5.11. SPSS reporta en una misma tabla las tasas de clasificación aparentes y aquellas calculadas con el método de uno-fuera. Cuando todos los datos se usan para calcular las funciones de clasificación lineales para luego ser asignados a los grupos, la tasa de clasificaciones correcta para el primer grupo (no-exitosa) es del 100% y para el segundo (exitosa) del 92% de donde la tasa global de clasificaciones correctas es del 95.7% que es un estimado optimista del verdadero desempeño de la regla de clasificación. Al

corregir este estimado optimista mediante el método de uno-fuera, la tasa de clasificación decrece a un 93.5%.

Tipo de empresa			Predictr Group membership		Total
Original	Count	No exitosa	21	0	21
		Exitosa	2	23	25
	%	No exitosa	100.0	.0	100.0
		Exitosa	8.0	92.0	100.0
Cross-validated(a)	Count	No exitosa	20	1	21
		Exitosa	2	23	25
	%	No exitosa	95.2	4.8	100.0
		Exitosa	8.0	92.0	100.0

Tabla 5.11 Resultados de la clasificación. Resultados(b,c)

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 95.7% of original grouped cases correctly classified.

c. 93.5% of cross-validated grouped cases correctly classified.

Para establecer si la regla de clasificación lineal es satisfactoria para predecir el éxito de las empresas que aspiran a exportar, se requiere del cálculo de Cpro y la *Q*-Press. Para probabilidades a priori iguales, Cpro=.5, puesto que 93.5 >> 50, se establece que la regla de clasificación lineal asigna significativamente mejor a las empresas que una asignación al azar. Para formalizar la conclusión se calcula *Q*-Press = 34.78, este valor calculado se compara contra el valor crítico , se concluye que la regla de clasificación lineal supera en calidad a la clasificación al azar o utilizando

exclusivamente la información a priori, sin considerar la información financiera y del plan de exportación de las empresas solicitantes.

¿Sabías qué?

Uno de los problemas en la que al análisis discriminante tiene gran aplicación es el de calificación crediticia, esto es ¿cómo evaluar si una persona o empresa es solvente y por lo tanto candidata a recibir algún tipo de crédito? De hecho el score Z que se menciona en el Ejemplo 3, fue validado por el Dr. Edward I. Altman a través de un análisis discriminante.

La aplicación del análisis discriminante para predecir la solvencia de una entidad no se ha limitado al caso de las empresas, también se ha utilizado para revelar aspectos importantes sobre la economía de un país.

5.6 Otras alternativas para la clasificación

Otro método de estadística multivariante que puede utilizarse para clasificar individuos es la **regresión logística**. Esta metodología consiste en formular modelos, que son una representación matemática, de cómo los individuos eligen entre varias alternativas de productos, marcas o servicios. La probabilidad de elegir una cierta alternativa o de pertenecer a un grupo se modela ya sea como una función de los atributos del producto o del perfil del individuo. Los modelos de elección aleatoria más usados son el modelo multinomial condicional logit, introducido por McFadden y el modelo probit. La diferencia entre los modelos logit y probit radica en la función de densidad que se propone para los errores del modelo; el modelo logit propone funciones logísticas y el probit usa la función de densidad normal. El modelo logit, que es el más utilizado en mercadotecnia (Malhotra, 1984) se define como sigue:

$$\Pr(\textit{elegir opción } k) = \frac{e^{U_k(x)}}{\sum_{i=1}^{g} e^{U_i(x)}}; U(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Donde U_k es una función de utilidad que depende de un conjunto de variables independientes que representan las características de g alternativas disponibles, en el contexto de clasificación estas alternativas corresponden a los grupos definidos a priori, los que representan perfiles de productos o individuos que hacen la selección. Una de las ventajas del uso de los modelos logit sobre las reglas de clasificación paramétricas descritas en la sección 5.4 es que las variables *X's* que se consideran en la función de utilidad pueden ser cualitativas (escala nominal) o cuantitativas (escala de intervalo o razón). Otra ventaja es que en la función de utilidad pueden incluirse interacciones -términos que corresponden al producto entre variables- lo que lo hace más versátil y evita que valores altos en una variable dominen la asignación.

Erosa y Arroyo (2011) utilizaron los modelos logit para explicar el comportamiento de los clientes de una cadena de supermercados ante la ocurrencia de faltantes en anaquel de sus productos favoritos. Las variables género del cliente, lugar de residencia, monto regular de su compra y lealtad a la tienda detallista fueron los mejores predictores de la respuesta del cliente ante faltantes, las cuales se clasificaron en cuatro grupos: grupo 1= sustituye productos, grupo 2 = pospone la compra de productos, grupo 3 = compra los productos faltantes en otra tienda y grupo 4 = cancela su compra total si hay faltantes. Las últimas dos respuestas afectan a la tienda detallista por lo que la información sobre el perfil del cliente más sensible a la ocurrencia de faltantes, es de gran importancia para diseñar esquemas de retención enfocados a estos clientes sensibles.

En la práctica, las propiedades estadísticas de los datos son poco conocidas, por lo cual las reglas paramétricas pueden no ser aplicables, por tal razón se han desarrollado métodos alternativos

para la clasificación usando los conceptos de otras disciplinas. A continuación se describen dos de estas alternativas de clasificación.

Redes neuronales: La base para de los modelos de redes neuronales es la creación de modelos que “reproduzcan” la capacidad del cerebro humano, es decir que puedan emular el proceso humano de razonamiento. Las redes neuronales se describen como un gran número de elementos de procesamiento interconectados entre sí, neuronas, para dar solución a un problema específico, tal como ocurre con las personas, la red neuronal aprende o se “entrena” con la experiencia. La Figura 5.9 describe gráficamente la estructura de una red neuronal.

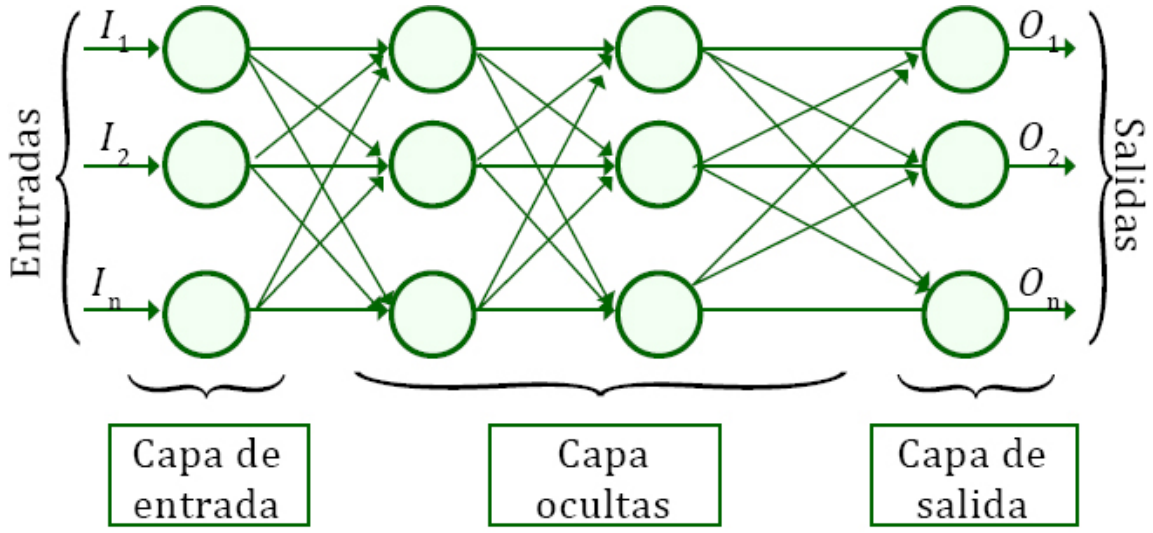


Figura 5.9 Esquema para la estructura de una red neuronal

En la clasificación con redes neuronales, las entradas son las variables definidas para la clasificación, que pueden ser tanto métricas como no métricas, y las salidas consisten en asignar a los objetos a los grupos. Las entradas se procesan empleando funciones de diversos tipos cuyos parámetros se revisan dependiendo de las clasificaciones correctas logradas.

West, Brockett y Golden (1997) emplearon las redes neuronales para modelar decisiones de elección de compra no-compensatorias. Tales selecciones se refieren a situaciones en

que los atributos de un producto o servicio no se compensan, es decir, casos en los que una opción que resulta muy bien evaluada en un atributo pero que no alcanza cierto valor en otro resulta inaceptable y por lo tanto no es elegida. El desempeño de las redes neuronales fue comparado con el del modelo logit y la regla de clasificación lineal y se encontró que bajo un esquema de decisión nocompensatorio las redes neuronales predicen mejor cuál alternativa elegirá un consumidor. A diferencia de los métodos estadísticos de clasificación que permiten identificar las variables que mejor distinguen y clasifican a los individuos, los parámetros de la red neuronal no tienen una interpretación directa. Para corregir este inconveniente, los autores de este estudio calcularon elasticidades que permiten identificar a las variables que mejor discriminan la elección de los consumidores.

Técnicas de ***conjuntos difusos***. Estos métodos introducen un cierto grado de vaguedad en la asignación de los individuos a los grupos. La principal ventaja de este método es la posibilidad de considerar criterios que no son precisos como es el caso de riesgo de trabajar con cierto proveedor o capacidad de una empresa para trabajar en colaboración con sus socios de negocios. Los criterios pueden ser cualitativos, cuantitativos o ambos y se recurre también a la opinión de expertos para calificar los diferentes criterios. El método hace uso de variables lingüísticas para generar números (difusos) predeterminados. La principal desventaja de este método es la dificultad de comprensión y aplicación que representa.

Un conjunto difuso es aquel en el que la barrera entre pertenencia y no pertenencia al conjunto no es absoluta sino difusa. Cada elemento tiene un cierto grado de pertenencia a cada grupo el que se expresa como un número entre 0 y 1 que no corresponde estrictamente a un valor de probabilidad. La teoría de conjuntos difusos se utiliza para caracterizar conceptos y conjuntos que tienen una ambigüedad o imprecisión inherente como el atributo antes citado de aversión al riesgo, la actitud de servicio de un proveedor o su prestigio. El diagrama de la Figura 5.10 describe un ejemplo de cómo se declara la pertenencia para un conjunto borroso, mientras

en el diagrama (a) el objeto pertenece de manera absoluta, grado de pertenencia 1, al grupo si la variable x que describe sus características se encuentra dentro del intervalo [5;7], en el caso (b) el nivel de pertenencia del objeto varía dependiendo de sus características, se tiene el mayor grado de pertenencia cuando x=6.

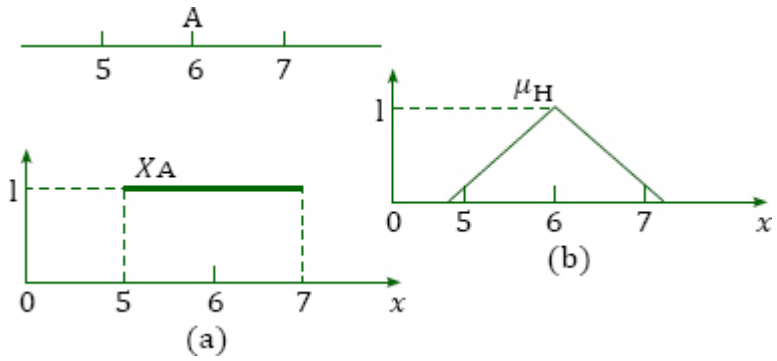


Figura 5.10 Nivel de pertenencia para un conjunto no difuso (a) y uno difuso (b).

Guerrero Dávalos y Terceño Gómez (s.f.) utilizaron los conjuntos borrosos para elegir proveedores de servicios de tercerización, outsourcing. Las empresas que tercerizan ceden las decisiones, la realización y el control sobre alguna actividad propia a terceros por lo que es crítico contar con un esquema de evaluación de potenciales proveedores que asegure que la actividad tercerizada se realizará apropiadamente. Los criterios para evaluar el desempeño de los proveedores incluyen atributos como capacidad para satisfacer las necesidades de la empresa compradora, compatibilidad en las filosofías empresariales y asistencia ofrecida. Como estos atributos tienen un cierto grado de subjetividad e incertidumbre, los autores usaron conjuntos borrosos para realizar la evaluación de proveedores de servicios para la administración de recursos humanos.

Un grupo de expertos evaluó a los proveedores en términos de múltiples criterios; a partir de esta información se determinó un índice de adecuación basado en conjuntos borrosos que permitió valorar los proveedores candidatos a través de dos pasos: i) comparación del proveedor con el ideal y ii) jerarquización de los

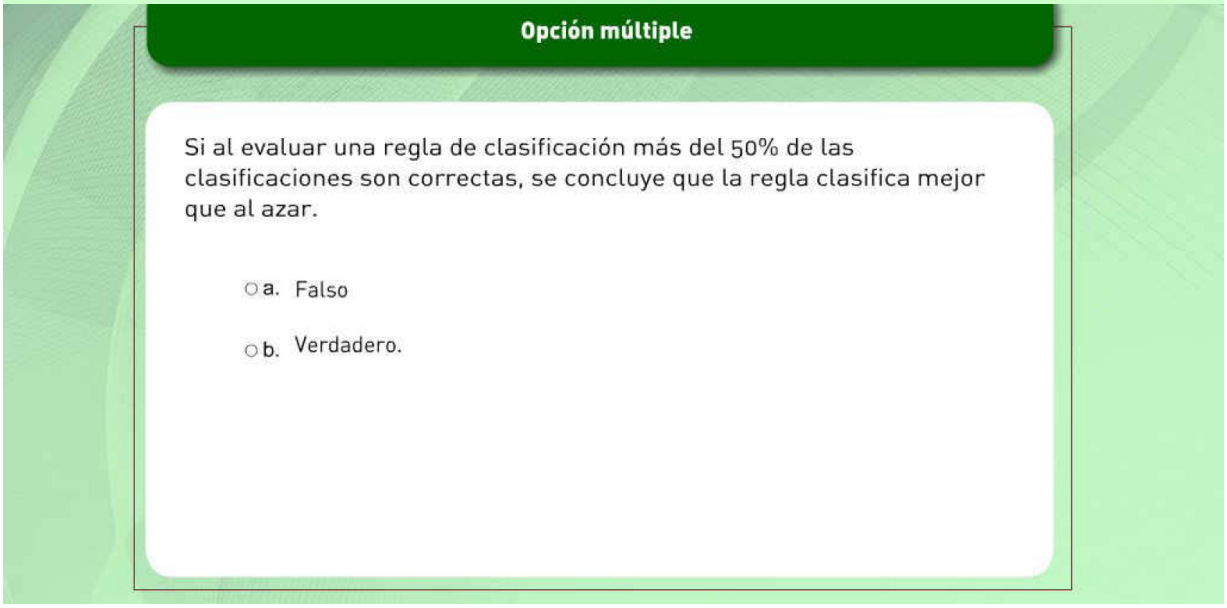
proveedores según la desviación que tienen respecto al ideal. La jerarquía construida permitió realizar la selección del mejor proveedor.

¿Sabías qué?

Para apoyar la automatización industrial se han desarrollado novedosos algoritmos de visión computacional y reconocimiento de patrones que se utilizan en sistemas automáticos de inspección y control de calidad.

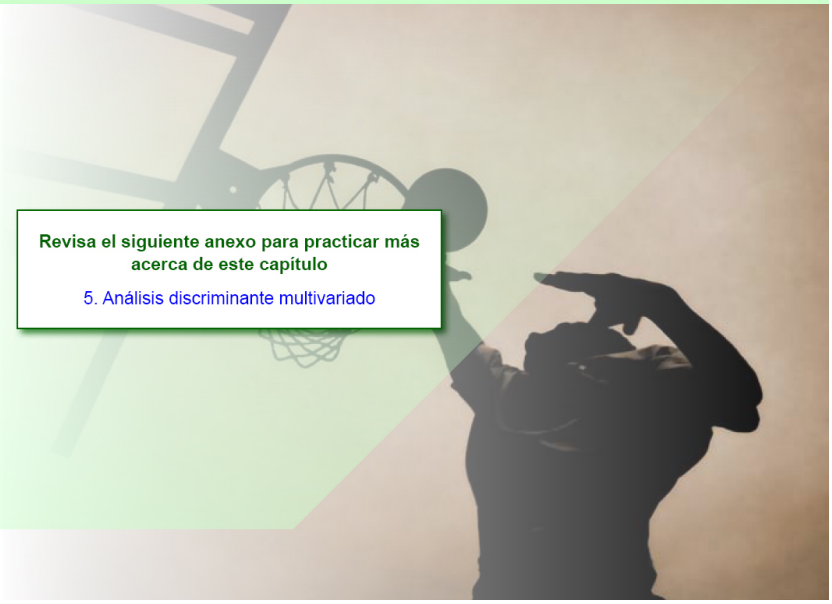
Actividad de repaso

5. Análisis discriminante multivariado



Ejercicio integrador

5. Análisis discriminante multivariado



Revisa el siguiente anexo para practicar más acerca de este capítulo

5. Análisis discriminante multivariado

Conclusión capítulo 5

5. Análisis discriminante multivariado

El análisis discriminante multivariante tiene como propósito la separación o distinción de individuos en grupos especificados a priori a partir de un conjunto de variables.

El problema de discriminación fue inicialmente abordado por Fisher quien lo resolvió al calcular aquella combinación lineal de las variables que separa más a los grupos. A partir de esta función lineal que discrimina significativamente a los grupos puede derivarse una regla de clasificación, también lineal, a través de la cual se busca asignar a nuevos individuos a los grupos identificados, de donde los problemas de discriminación y clasificación están fuertemente vinculados. Pero mientras en problema de discriminación implica perfilar y/o distinguir entre varios segmentos de individuos, el problema de clasificación tiene como objetivo predecir o establecer el grupo del que proviene un nuevo objeto a partir de la información sobre un cierto número de variables independientes.

Desde una perspectiva estadística, la clasificación de individuos requiere del cálculo de probabilidades a posteriori, esto es, probabilidades revisadas de pertenencia a los grupos y condicionadas por la información que se tiene sobre un objeto con respecto a p variables cuantitativas.

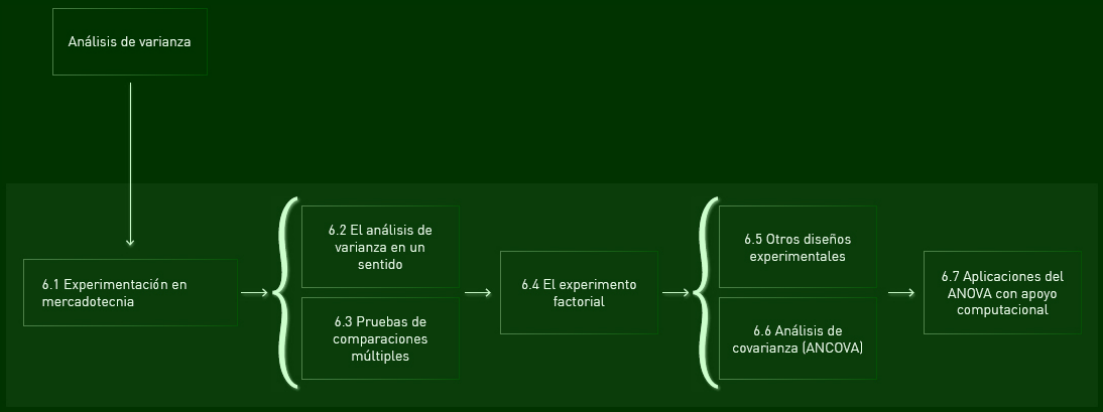
Aquellas reglas de clasificación que minimizan una medida del error o el costo de clasificar erróneamente a los individuos se denominan reglas Bayesianas u óptimas. El problema de construcción de reglas de

clasificación en estadística se ha abordado desde las perspectivas paramétrica y no- paramétrica. Entre las reglas paramétricas de mayor que asumen normalidad para las variables que se utilizan para clasificar, están la regla, función, de clasificación lineal, la cuadrática y la **regla de Anderson.** Una alternativa más reciente, también desde la perspectiva de estadística paramétrica es el uso de modelos de regresión logística, el más utilizado es el llamado modelo logit.

El problema de clasificación ocurre en múltiples disciplinas, desde biología hasta finanzas, dada su importancia y la necesidad de contar con mejores reglas de clasificación y de reconocimiento de patrones, se han desarrollado nuevas propuestas que han podido implementarse gracias a los avances en las ciencias computacionales. Estos métodos de clasificación no-estadísticos incluyen el uso de redes neuronales, conjuntos difusos y métodos de superación basados en múltiples criterios (por Ejemplo PROMETHEE TRI). Estos métodos ofrecen ciertas ventajas sobre los métodos estadísticos tradicionales como son: poder utilizar variables tanto cualitativas como cuantitativas, considerar la ambigüedad o incertidumbre de ciertos atributos, modelar decisiones no-compensatorias y trabajar con grandes cantidades de datos. A pesar de estos avances metodológicos, los métodos de discriminación y clasificación estadísticos siguen utilizándose con éxito debido a su buen desempeño, amplia difusión y facilidad de uso y comprensión.

Capítulo 6
Análisis de varianza

Organizador temático



6. Análisis de varianza

Introducción

El análisis de varianza es una metodología genérica introducida por R. A. Fisher en el contexto del análisis de experimentos agrícolas la cual consiste en descomponer una suma de cuadrados, que representa la variabilidad total observada de una respuesta, en componentes que puedan asignarse a fuentes de variabilidad identificables. La pregunta crítica que se busca responder a través del análisis de varianza es si una respuesta de interés (Y) difiere dependiendo de los niveles o “valores” que asumen un conjunto de variables independientes o factores (Xs).

En el área de mercadotecnia las variables X que se estudian en este análisis multivariante son típicamente variables en escalas nominales tales como color, marca o tipo de empaque para un producto con categorías definidas; o bien variables con categorías

ordinales tales como bajo, medio o alto nivel de picante para una botana o en el grado de consumo de cereales para el desayuno. Si bien las variables *X*s suelen ser cualitativas, el análisis de varianza o ANOVA (por sus siglas en inglés ANalysis Of VAriance) también puede incluir variables cuantitativas tales como cantidad de producto (250, 300 ó 500 mililitros de bebida por envase) o de calorías (100, 125 ó 150 kilocalorías).

En el ANOVA se asume que la variabilidad observada sobre la respuesta de interés se atribuye a:

1) Variabilidad sistemática como resultado de la presencia de diferentes condiciones representadas por el uso de distintos niveles para los varios factores cuya influencia es el interés central del estudio.

2) Variabilidad adicional en la respuesta como resultado de la influencia de variables externas o de ruido cuyo efecto requiere ser controlado o atenuado de alguna manera ya sea porque se encubre el efecto que tienen los factores sobre la respuesta, o porque la variabilidad que inducen las variables de ruido representa un problema de calidad. En este último caso, el propósito de la experimentación es diseñar procesos y productos resistentes o robustos respecto a los cambios de variables externas no-controlables como pueden ser las condiciones de uso para un producto (Taguchi, Chowdhury y Wu, 2005).

3) Variabilidad aleatoria provocada por causas no identificables, la cual se asocia con el llamado **error experimental** que describe la variación en la respuesta aun cuando ésta se haya registrado en las mismas condiciones.

El propósito del ANOVA se describe entonces como la medición y comparación de cada una de estas fuentes de variación con el fin de determinar hasta dónde los factores de interés, al modificarse sistemáticamente, resultan en diferencias significativas para la respuesta. Cuando la variación aleatoria es muy grande, las diferencias sistemáticas no se podrán detectar con facilidad, es por eso que el control de las variables externas a través de un buen

diseño experimental resulta esencial. A este punto es importante remarcar que si bien el ANOVA se puede utilizar con datos obtenidos de una encuesta, su uso se asocia principalmente al análisis de datos experimentales, esto es, de datos que fueron generados a través de la manipulación y control de variables para observar cómo se afecta una cierta respuesta.

Si bien el uso de encuestas y de datos secundarios se vincula más con los estudios de inteligencia de mercados, hay que recordar que la experimentación en mercadotecnia también es posible, sobre todo bajo las condiciones controladas de un laboratorio al cual acuden los consumidores potenciales para evaluar nuevos productos, servicios o mensajes publicitarios. La gran ventaja de un experimento es su alta validez interna, que se refiere a la capacidad para inferir, a partir de la información generada por el estudio, qué hay una relación causal entre el conjunto de variables independientes *X* y la respuesta *Y* (Malhotra, 2004). Dado que en un experimento hay manipulación y control de condiciones, es posible tanto establecer precedencia, el cambio en *X* lo que resulta en *Y* y no viceversa, así como descartar explicaciones alternas para los cambios de *Y* aparte de *X*. Los esquemas de control utilizados sobre las variables externas que pudieran afectar a la respuesta, dan lugar a diferentes diseños experimentales.

El ANOVA es por tanto una técnica muy versátil ya que permite analizar datos generados a partir de variados diseños experimentales y para factores de distinta naturaleza según lo demuestra el siguiente estudio realizado por Prendergast, Tsang y Chan (2010):

> El país de origen de un producto (COO del inglés COuntry of Origin) es utilizado como clave extrínseca por los consumidores bajo el supuesto de que la posición competitiva del país en el mercado determina la calidad de los productos que fabrica. Sin embargo el fenómeno de globalización ha llevado a las empresas multinacionales a fabricar sus productos en países de bajo costo por lo cual la validez de la clave país de origen es

cuestionable. Sin embargo, se argumenta que los consumidores han sustituido el COO por la clave país de origen de la marca, asumiendo que aun cuando un producto sea manufacturado en diferentes países, la marca garantiza su calidad. Además del país de origen de la marca de un producto (COB del inglés Country of Origin of Brand), también se está establecido que el nivel de involucramiento, esto es la motivación y atención que amerita la compra de un producto, influye sobre las decisiones de compra. Cuando este involucramiento es alto, se esperaría que el COB tuviera mayor influencia ya que los consumidores utilizarían esta clave extrínseca simple para compensar su bajo involucramiento en la elección del producto. En contraste, aquellos consumidores altamente involucrados con la compra preferirían utilizar información más precisa que el COB. El factor COB fue manipulado a través de anuncios impresos en revistas de computadoras. Dos países de origen de la marca, Japón y Corea del Sur fueron anunciados en las etiquetas de computadoras tipo notebook que promovía la revista, los productos ostentaban además la etiqueta de “hecha en Taiwan” para controlar por el efecto del país de manufactura. El nivel de involucramiento del consumidor que compró una computadora notebook se midió a través de una multi-escala semántica diferencial de siete reactivos.

A partir de los puntajes alcanzados en la escala, los participantes se categorizaron en dos grupos: alto y bajo involucramiento. Un total de 198 consumidores fueron entrevistados asignándoles al azar un anuncio del producto ya sea con marca japonesa o coreana. El propósito fue generar cuatro condiciones experimentales (bajo y alto involucramiento, fuerte y débil COB) con al menos 40 consumidores en cada grupo. La intención de compra del producto fue la variable de respuesta. El Análisis de Varianza aplicado a los datos llevó a establecer que el país de origen de la marca no influye en la intención de compra cuando el nivel de involucramiento del comprador es alto, pero si tiene un efecto significante cuando el

involucramiento es bajo (la marca japonesa resultó en mayores intenciones de compra).

¿Sabías qué?

Sir Ronald Fisher formalizó el análisis de varianza en su artículo sobre el tema de genética (año) *The correlation between relatives on the supposition of Mendelian inheritance.* La primera aplicación del análisis de varianza se publicó en 1921 y el análisis se difundió extensamente después de que Fisher lo incluyó en su libro Statistical methods for research workers el cual se publicó en 1925.

6.1 Experimentación en mercadotecnia

La experimentación o manipulación de condiciones de prueba se asocia tradicionalmente con ciencias como la Física, Química o Biología. Si bien los orígenes del diseño experimental estadístico están en las áreas de agricultura y genética, y algunos de los avances más interesantes fueron sugeridos por estadísticos como G. Box en el contexto de aplicaciones en la industria química, el uso del diseño experimental se ha extendido hacia las áreas de ciencias sociales y administrativas. En el caso de mercadotecnia, la

experimentación ha sido muy aprovechada para el diseño de productos y servicios. De hecho, el análisis conjunto (ver Capítulo 7), que hace uso extensivo de la experimentación, ha sido desarrollado por los mercadólogos a partir de las propuestas iniciales en estadística.

Como se mencionó en la introducción de este capítulo, un diseño experimental implica elaborar un plan para:

1. manipular los valores o **niveles** de variables relevantes denominadas factores;

2. determinar las respuestas de interés sobre las que se medirá su efecto y

3. establecer estrategias para el control del efecto de variables externas que también pueden influir en la respuesta.

Los objetivos de un buen diseño experimental son:

a. reducir el número de pruebas o experimentos y

b. minimizar la variabilidad aleatoria o no explicada de la respuesta, esto es el llamado error experimental.

El primer objetivo se orienta hacia la reducción en costos del estudio, mientras el segundo garantiza que se determine adecuadamente el efecto de los factores sobre la o las respuestas de interés. Para atender a este segundo objetivo es necesario hacer un buen control de las llamadas variables externas o de ruido; que se realiza a través de dos estrategias básicas:

a) Agrupamiento de los individuos u objetos (unidades experimentales) que reciben los distintos estímulos en grupos homogéneos dentro de sí y heterogéneos entre sí llamados bloques. Los bloques se forman en términos de variables externas que el investigador de mercados considera que afectan de forma importante a la respuesta. Usualmente las variables que se utilizan para formar los bloques son del tipo cualitativo, esto es, están en escala nominal u ordinal. Como los niveles de los factores de interés

se prueban en cada bloque, la influencia de la variable externa con base en la que se generaron los bloques se cancela, permitiendo una mejor comparación del efecto de los factores sobre la respuesta. Variables típicas que se utilizan para formar bloques son: perfil demográfico de los participantes (género y/o rango de edad), nivel socioeconómico (A, B, C, D) e intensidad en el consumo de un producto (bajo, promedio, alto).

b) Medición de las variables externas que pudieran también afectar a la respuesta para "ajustar" los datos antes de contrastar los niveles de los factores. Estas variables externas reciben el nombre de covariables y se requiere que sean variables cuantitativas, es decir que estén en escalas de intervalo o razón. Cuando en un experimento se incluyen covariables, el análisis correspondiente se denomina Análisis de Covarianza (ANCOVA por sus siglas en inglés, ANalysis of COVAriance). Este análisis es una combinación del análisis de regresión y el de varianza. En el formato usual para este análisis, primero se corrige por el efecto de variables externas y después se realiza la comparación de los niveles del factor en un valor fijo de la variable externa. Variables externas que se definen con frecuencia como covariables en un experimento en mercadotecnia son: aversión al riesgo, nivel de conocimiento sobre un tema y nivel de exposición a la publicidad.

De la discusión anterior se tiene que un experimento puede involucrar uno o varios factores, unifactorial versus factorial, y tomar en consideración varias variables externas. También es posible definir más de una respuesta, en este caso los datos se procesarían a través de un Análisis de Varianza Multivariado (MANOVA por sus siglas en inglés Multivariate ANalysis Of VAriance). Este capítulo se centra en el análisis de varianza univariado, es decir en aquel caso en que, si bien puede haber múltiples factores y variables externas, el efecto de este conjunto de variables se evalúa a través de una única respuesta. Para detalles sobre el MANOVA el lector interesado puede consultar el libro de Johnson y Wichern (1998). Los siguientes ejemplos ilustran el tipo de proyectos para los cuales el ANOVA es la técnica de estadística multivariada que conviene utilizar para analizar los datos generados.

Ejemplo 1. Una compañía que fabrica productos de panificación planea extender su línea de productos para incluir una línea de galletería. La empresa ha segmentado su mercado con base en la edad de sus consumidores y supone que la presentación del producto influye más sobre la decisión de compra entre menor sea la edad del consumidor. Para evaluar la presentación de los nuevos productos que se lanzarán en paquetes económicos, tres productos de cada tipo, se reclutaron 30 participantes que se agruparon de acuerdo a su edad: 10 niños entre 10-13 años; 10 jóvenes de 14-17 años y 10 personas de entre 18-21 años. El producto se presentó en paquetes de celofán transparente y celofán opaco con dibujo, con y sin rejilla para las galletas; los paquetes podían tener además ya sea 6 ó 12 piezas de galleta. Los productos se presentaron a los participantes todos a la vez en una canastilla dándoles oportunidad para que los examinaran por 15 minutos antes de pedirles que asignaran rangos de preferencia a los productos en términos de cuáles les interesaría más comprar. La estructura del experimento descrito es la siguiente:

Tres factores, cada uno a dos niveles: *A* = envoltura de la galleta (nivel 1 = celofán transparente, nivel 2 =celofán opaco); *B* = rejilla (*B*1 = sin, *B*2 = con rejilla); *C* = cantidad de piezas (*C*1 = 6, *C*2 = 12 galletas). El total de posibles productos es 23, esto es 8 diferentes presentaciones.

Variable de respuesta es el ranking asignado a los productos, la escala de esta variable es ordinal.

Variable externa es el participante ya que cada uno de los potenciales consumidores evaluó todos los posibles productos, de esta manera se corrige no únicamente por la edad sino incluso por las diferencias en las preferencias de un individuo a otro. El *Diseño experimental* es por tanto uno de bloques, ya que cada individuo es un bloque para las 8 combinaciones de los factores de diseño del producto.

Ejemplo 2. El área de recursos humanos de una empresa de servicios que opera como distribuidor mayorista de equipo de oficina

ha diseñado cuatro talleres de capacitación para los agentes de ventas de la compañía; el objetivo de los talleres es mejorar las habilidades de venta de los agentes. Los agentes tienen diferentes niveles de experiencia de acuerdo con la antigüedad en la empresa. Veinte agentes de ventas son elegidos al azar y divididos en cuatro grupos de cinco personas. Cada grupo fue asignado aleatoriamente a alguno de los cuatro talleres; el responsable del proyecto de capacitación planea evaluar las ventas realizadas por los agentes, antes y después del taller para determinar cuál de los cuatro es el más efectivo. La estructura del experimento descrito es la siguiente:

Un único factor (experimento unifactorial), el taller de capacitación el cual tiene cuatro niveles.

Variable de respuesta es la diferencia en las ventas reportadas por el agente antes y después del taller de capacitación. Esta variable de respuesta es la apropiada en situaciones como la descrita ya que las ventas de un agente pueden variar no únicamente en función de la capacitación que recibió sino también debido a otras variables como el territorio a donde esté asignado y su perfil personal (habilidades de comunicación, trato con los clientes o capacidad de negociación).

La *variable externa* es la experiencia del agente, que influye sobre el volumen de ventas que puede lograr. En este caso la experiencia se considera una covariable que incrementa el volumen de ventas que puede alcanzar cualquier agente.

Diseño experimental es totalmente al azar. En este diseño no hay un control explícito de las variables externas; la selección y asignación aleatoria de los participantes a los grupos se espera que resulte en grupos de cinco agentes con perfiles similares por lo que las comparaciones de su desempeño deberían reflejar únicamente las diferencias de los talleres de capacitación que recibieron.

¿Sabías qué?

En los años 80, G. Taguchi, estadístico japonés y ganador del premio Deming de calidad, logró despertar el interés de los profesionales en ingeniería por aplicar el diseño experimental a la optimización de productos y procesos. El concepto de variable de ruido que está estrechamente vinculado al de diseño robusto, fue usado por Taguchi para describir a aquellas variables que afectan considerablemente la variación de una respuesta pero que no son controlables, o es muy difícil o costoso controlarlas. En el contexto de mercadotecnia la variabilidad o ruido que inducen estas variables dificulta la evaluación del efecto que tienen los factores sobre la respuesta de interés. En este sentido las variables de ruido, identificadas por los ingenieros, son equivalentes a las variables externas que representan causas plausibles" de la variación adicional observada en una respuesta.

6.2 El análisis de varianza en un sentido

El caso más elemental del ANOVA es cuando hay un único factor bajo estudio y el efecto de las variables externas se controla indirectamente a través de la selección y asignación aleatoria de unidades experimentales, individuos, objetos, o empresas a los niveles del factor de interés. En la terminología de diseño experimental se tendría por tanto el caso de un experimento factorial en un Diseño Completamente al Azar. Los datos recolectados se describen en el siguiente modelo estadístico:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad (1)$$

donde:

Y_{ij} se refiere a la observación o medición para la variable dependiente o respuesta registrada para el *j*-ésimo individuo al i-ésimo nivel del factor

μ_i es la media de la respuesta cuando se utilizó la condición o nivel *i* del factor

ε_{ij} es el término de error que describe la diferencia aleatoria que hay entre cada dato y la media de su grupo, esto es, aquélla que corresponde al nivel *i* del factor. El análisis de varianza es una técnica paramétrica que impone a estos términos de error, supuestos similares a los de un modelo de regresión, esto es, se asume que los errores son variables aleatorias normales, independientes entre sí y de varianza constante.

La Figura 6.1 describe gráficamente el modelo estadístico anterior para el caso de un factor con cuatro niveles (i=1,2,3,4). A cada nivel del factor hay una población de posibles valores para la **respuesta** de interés, la media de esta población es igual a μ_i. Una medición o registro específico de la respuesta se desvía de esta media en una cantidad aleatoria representada por ε_{ij}. El objetivo del **ANOVA en un sentido** es probar la validez de la hipótesis nula siguiente:

H_0: $\mu_1=\mu_2=\mu_3=\mu_4$ versus H_1: $\mu_i \neq \mu_j$ para al menos una $i \neq j$

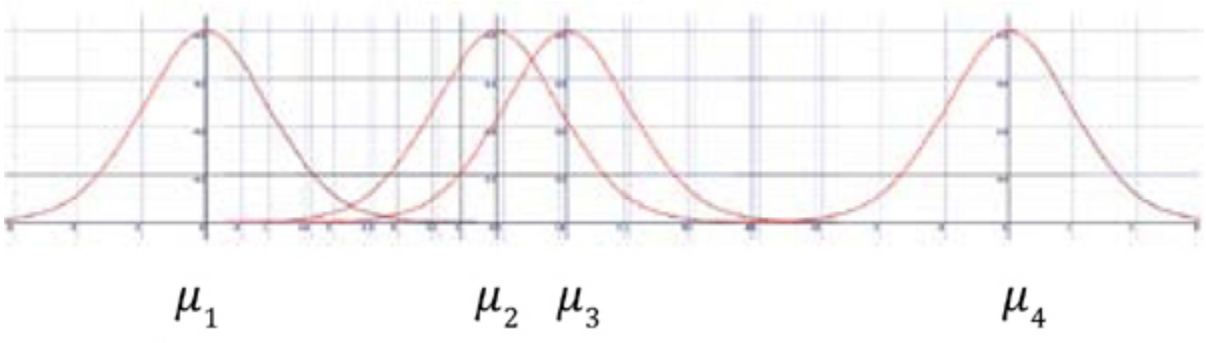


Figura 6.1 Comparación de varias medias para el ANOVA en un sentido

El modelo dado en (1) resulta apropiado para describir el caso simple de una experimentación que involucra a un único factor sin control explícito de variables ruido, pero no es apropiado para describir situaciones más complejas (uso de múltiples factores o/y variables de ruido controladas). Por tanto es necesario re-definir el modelo anterior a uno que se pueda generalizar y utilizar en cualquier caso. Para ello se descompone la media del *i*-ésimo nivel del factor o tratamiento en dos componentes (Montgomery, 2009), esto es $\mu_i = \mu + \tau_i$, lo que resulta en el siguiente modelo estadístico lineal:

$$Y_{ij} = \mu_i + \tau_i \text{ con } \varepsilon_{ij} \sim \text{NID}(0, \sigma^2) \qquad (2)$$

En el modelo (2), el parámetro μ representa a la media global de la respuesta sobre todos los niveles del factor bajo prueba. En tanto que τ_i (i=1,2, …, k posibles niveles para el factor) se refiere al efecto que el *i*-ésimo nivel del factor tiene sobre la respuesta. El modelo anterior es el caso más simple de lo que se denomina un modelo ANOVA en estadística.

Los modelos estadísticos de regresión y ANOVA son dos clases diferentes de modelos lineales que tienen características particulares (Searle, 1997). Los parámetros del modelo (2) se pueden estimar con mínimos cuadrados ordinarios si se impone la restricción adicional de que $\sum_{i=1}^{k} \tau_i$ de otra forma no se puede obtener una solución única para los parámetros. Sin embargo, y a diferencia del caso del modelo de regresión, en un modelo ANOVA más que estimar los parámetros, probar una hipótesis para ellos es lo que resulta por demás interesante (Arroyo, 1994). Las hipótesis previamente formuladas para el modelo (1) se re-escribirían en términos del modelo ANOVA dado en (2) como sigue:

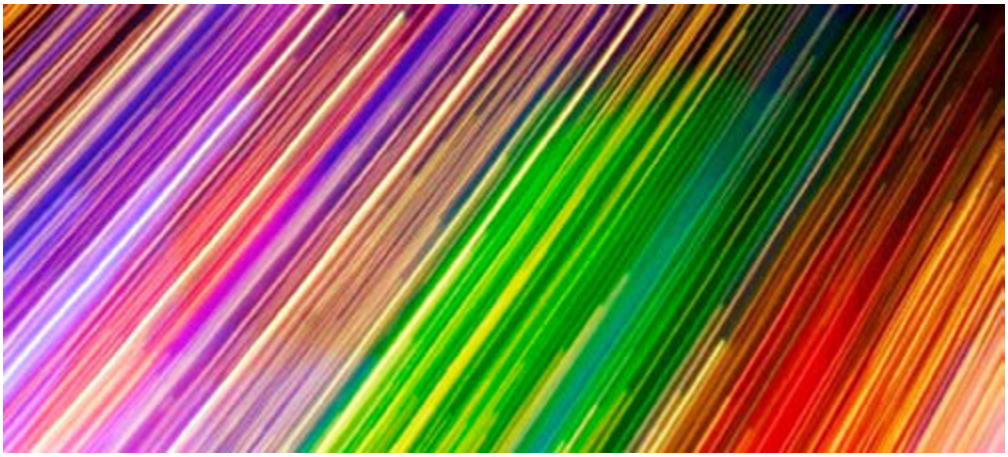
$$H_0: \tau_1 = \ldots = \tau_k = 0 \text{ versus } H_0: \tau_i \neq 0 \text{ para al menos una } i$$

Note que si la hipótesis nula es cierta, entonces todos los datos tienen una misma media (μ) y las diferencias entre ellos son meramente aleatorias, es decir que los distintos niveles del factor no inducen cambios sistemáticos en el valor medio de la respuesta. Mientras que si la hipótesis nula se rechaza a favor de la alternativa esto implica que, para al menos uno de los niveles del factor, la respuesta media resulta ser mayor (menor) que para los otros niveles en una cantidad igual a $\tau_{i.}$

Para probar la hipótesis anterior, la variabilidad total observada en la respuesta se descompone en dos partes:

1) la variabilidad sistemática en la respuesta atribuible al uso de diferentes niveles para el factor en estudio. Esta variación entre grupos de unidades experimentales, que fueron igualmente tratadas, es consecuencia de la manipulaci ón de las condiciones o niveles del factor.

2) la variabilidad adicional observada en la respuesta debido a otras causas no reconocidas. Esta variabilidad dentro de los grupos de tratamientos se atribuye, principalmente, a las diferencias que pueda haber entre las unidades experimentales, por ejemplo individuos con diferentes preferencias, edades o interés en un producto.



La propuesta de Fisher para el ANOVA consiste en cuantificar cada una de estas dos fuentes de variación y compararlas. Si la variabilidad entre grupos es similar a la variabilidad dentro de grupos se concluye que los niveles del factor no inducen cambios sistemáticos en la respuesta, es decir $\tau_i=0$ para toda $i=1,2,\ldots k$; mientras que, si la variabilidad entre grupos es mayor que la variabilidad aleatoria, se concluye que los niveles del factor inducen cambios significativos en la respuesta, esto es $\tau_i \neq 0$para al menos una *i*. Para cuantificar las dos fuentes de variabilidad se calculan sumas de cuadrados las cuales se comparan con una prueba *F* =MSB/MSE donde MSB es el cuadrado medio entre grupos y MSE el cuadrado medio dentro de grupos o error experimental. Bajo los supuestos del modelo ANOVA, independencia, normalidad y varianza constante, la distribución de probabilidad del cociente MSB/MSE es *F* con $v_1 = k-1$, $v_2 = k\ (r-1)$ grados de libertad, asumiendo que se tiene un **experimento balanceado**, esto es que para cada nivel del factor hay un igual número de observaciones o r réplicas. La hipótesis de igualdad entre niveles del factor, tratamientos, se rechaza si MSB>>MSE lo que implica que *F* debe ser significativamente mayor de uno para rechazar H_0. De donde la región de rechazo correspondiente es:

$$RR = \{F > F[\alpha; v_1 = k-1, v_2 = k\ (r-1)] \qquad (3)$$

Donde α es el nivel de significancia que el analista especifica para la prueba y que permite el control del error del tipo *I* que, en este caso, se refiere a declarar diferencias entre los niveles del factor cuando éstas realmente no existen.

La Tabla 6.1 muestra los detalles para el ANOVA en un sentido, el término FC, (CT de Correction Term en inglés) es el factor de corrección por la media y, lo mismo que para el ANOVA de un análisis de regresión es igual al cuadrado de la suma de todos los datos dividido entre el total de ellos, esto es

$$FC = \frac{\left[\sum_{i=1}^{k}\sum_{j=1}^{r} Y_{ij}\right]^2}{rk} = \frac{Y_{\cdot}^2}{rk}$$

Cabe observar que para simplificar la notación con sumatorias se utilizan puntos (.) que indican cuándo se ha sumado sobre un subíndice. Así $Y_{i\cdot}$ indica que se ha sumado sobre j=1,2,…,r esto es, que se han acumulado todas las observaciones para el grupo i, o i-esimo nivel del factor. Para fines computacionales la forma más simple de calcular la suma de cuadrados del error es por diferencia ya que SST = SSB + SSE.

Fuente de variación	GI	Sumas de cuadrados	Cuadrados medios	F
Entre grupos o tratamientos	$k-1$	$\sum_{i=1}^{k} \frac{Y_{i\cdot}^2}{r} - FC$	$SSB/(k-1)$	MSB/MSE
Dentro de grupos o residual	$k(r-1)$	Por diferencia	$SSB/k(r-1)$	
Total	$rk-1$	$\sum_{i=1}^{k}\sum_{j=1}^{r} Y_{ij}^2 - FC$		

Tabla 6.1 Cálculos para el ANOVA en un sentido

Ejemplo 3. El efecto que tienen los colores sobre los consumidores es un tema de interés. La venta de un producto puede depender del color del envase, del logo de la marca o de los colores del local donde se exhibe. El color es una incitación subconsciente a la compra; un consumidor acepta los colores que le agradan mientras que rechaza otros.

Sabiendo esto, una empresa fabricante de alimentos decidió probar el efecto que diferentes colores tienen sobre la aceptación

del consumidor para una nueva bebida energética baja en calorías. Se probaron colores que transmitían las características del sabor de la bebida, rosa y verde pálido, un color que transmitía frescura, azul, y también se probó una bebida transparente que transfiere la idea no sólo de frescura sino de ser un producto natural. Cuatro lotes de la bebida, uno de cada color, se prepararon, envasaron y empacaron en cajas de 12 botellas cada una. Las cajas se distribuyeron al azar en 16 supermercados de distintas cadenas a las cuales el fabricante surte regularmente sus productos. Después de quince días de prueba en el mercado se registró el número de cajas de producto vendidas por cada supermercado, los datos se reportan en la Tabla 6.2.

Empleando las fórmulas de la Tabla 6.1 con estos datos se calculan las correspondientes sumas de cuadrados:

FC = (2070+1680+1280+2270)2/16 = 3,330,625

SST = {1,082,900 + 715,400 + 428,000 + 1,295,700} – FC = 191,375

SSB = (20702 + 16802 + 12802 + 22702)/4 – FC = 144,025

SSE = 191,375 – 144,025 = 47,350

Estas sumas son las entradas del ANOVA que se muestra en la tabla 3.

	Colores			
	Rosa	Verde pálido	Azul	Transparente
	600	350	300	500
	450	400	220	580
	520	450	360	620
	500	480	400	570
Y_i	2,070	1,680	1,280	2,270
$\sum_j Y_{ij}^2$	1,082,900	715,400	428,000	1,295,700

Tabla 6.2 Ventas de cajas de nueva bebida energética

Fuente de variación	gl	Sumas de cuadrados	Cuadrados medios	F
Entre grupos (colores)	3	144025	48008	12.17
Dentro de grupos (error experimental)	12	47350	3946	
Total	15	191375	12.17	

Tabla 6.3 ANOVA para la comparación de distintos colores de una bebida energética

La región de rechazo correspondiente para un nivel de significancia del 5% queda definida como RR = {F > F(3,12) = 3.49}, dado que el estadístico de prueba está dentro de esta región se rechaza la hipótesis nula concluyéndose que las ventas, promedio, de la bebida energética difieren dependiendo del color que ésta tenga.

6.3 Pruebas de comparaciones múltiples

Con frecuencia en un experimento no basta con concluir que hay diferencias entre los tratamientos, sino que el interés central puede ser: contrastar todas las parejas de tratamientos; identificar aquel tratamiento o nivel del factor, bajo el cual se obtiene la mejor respuesta; comparar dos combinaciones lineales de medias entre sí o comparar los grupos tratados bajo cierta condición contra un control. Este tipo de comparaciones, que involucran a varias combinaciones lineales de medias, se denominan pruebas de **comparaciones múltiples**. Para realizar este tipo de comparaciones se requiere un examen cuidadoso del poder y las tasas de error tipo *I* (rechazar la hipótesis nula cuando es verdadera) que aplican para cada una de las varias hipótesis individuales de interés así como del conjunto de éstas. Cuando se realiza una única prueba de significancia a un nivel de significancia

de α, el riesgo de cometer un error del tipo *I* corresponde a una tasa de error individual o tasa de error por comparación (*comparisonwise error rate* en inglés). En caso de que se realicen dos o más comparaciones en un experimento la probabilidad de rechazar al menos una de ellas, cuando en realidad la hipótesis nula es verdadera, se conoce como tasa de error de experimentación sabia (*experimentwise error rate*) la cual también se asocia con el concepto de tasa de error por familia de comparaciones (*familywise error rate*). Esta última tasa de error recibe este nombre porque se refiere a la probabilidad de cometer al menos un error del tipo I al probar una familia o conjunto de hipótesis vinculadas con un mismo tema (Bender Lange, 2001).

Considere el ejemplo de comparar una serie de talleres de capacitación para nuevos agentes de ventas contra el programa de inducción básico diseñado por el área de recursos humanos. La tasa de error global que aplica cuando se tiene un grupo de C hipótesis resulta ser igual a $[1-(1-\alpha)^{C}]$ donde C es el número de comparaciones independientes que se hacen. Si un factor tiene t = 5 niveles y el propósito de la experimentación es comparar todas las posibles parejas de medias (C = 10 comparaciones por parejas en total), esta tasa de error global resulta ser de 0.40126 lo que implica que hay un gran riesgo de declarar por lo menos a una pareja de medias diferentes entre sí, aun cuando éstas sean iguales.

El término post-ANOVA, que se ha usado extensamente para denominar a aquellos métodos que se enfocan a realizar múltiples comparaciones entre medias después de haber obtenido los datos, ha sido criticado fuertemente (ver el East Carolina University Discussion Forum on Multiple Comparisons, s.f.) pues entre más hipótesis involucre un experimento, mayor es el riesgo de cometer un error tipo I al probar una hipótesis individual, no importa si éstas hipótesis se planearon antes de generar la información experimental. Realizar comparaciones entre los varios niveles de un factor únicamente cuando la prueba F global de igualdad entre tratamientos es rechazada (prueba ómnibus) ofrece protección hacia el riesgo de incrementar el error del tipo I, es decir protege contra el

riesgo de declarar diferencias entre medias cuando éstas son iguales.

Las pruebas de comparaciones múltiples como la LSD (*Least Significant Difference*) sugerida por Fisher, consisten en realizar múltiples pruebas t sobre las parejas de medias, que resultan adecuadas siempre y cuando el ANOVA se haya declarado significante (Levin, Serlin y Seaman, 1994). Sin embargo, otras pruebas como la HSD (*Honest Significant Difference*) sugerida por Tukey fueron diseñadas con el propósito de controlar el error del tipo I por lo cual la prueba F resulta redundante. En la prueba de Tukey, la comparación inicial que se realiza entre las medias más extremas, menor versus mayor, resulta equivalente a la prueba ómnibus por lo tanto es apropiado utilizarla sin realizar antes un ANOVA. En general, cuando la hipótesis nula es verdadera, un control débil de la tasa de error del tipo I como el que da la LSD de Fisher se facilita con la prueba F global; sin embargo cuando la hipótesis nula es falsa pero hay varias hipótesis nulas parciales que son verdaderas, la prueba F no mantiene un control fuerte sobre la tasa de error y conviene considerar pruebas cuyo diseño considera el control de la tasa de error global, como es el caso de la HSD de Tukey (Hochberg y Tamhane, 1987).

Las pruebas de comparaciones múltiples pueden clasificarse en varias formas, entre las que destacan las siguientes (Toothaker, 1993):

1. Pruebas planeadas versus no-planeadas. Una comparación es planeada si se formula antes de obtener los datos y no-planeada cuando se realiza después de haber inspeccionado los datos. Las comparaciones planeadas son con frecuencia sugeridas por la estructura del experimento (se incluye un control, se está utilizando un producto o marca nueva en el mercado, por mencionar algunos) o por las preguntas de investigación formuladas por el proyecto. Las comparaciones no-planeadas suelen involucrar comparaciones sugeridas por los datos observados.

2. Pruebas ortogonales y no-ortogonales. Se refiere a si las pruebas que se realizan resultan ser no-correlacionadas, ortogonales, entre sí (ver Ejemplo 5).

3. Comparaciones en pares versus otro tipo de comparaciones. Este criterio se refiere a si el conjunto de pruebas involucra únicamente comparaciones entre dos medias o combinaciones lineales de medias.

4. Tipos de error. Hay dos formas básicas para el control de la tasa de error: sobre cada comparación o prueba individual que se realiza y sobre el conjunto o familia de comparaciones.

Dada la complejidad del problema para controlar correctamente la tasa de error que aplica cuando se realizan múltiples comparaciones, se han desarrollado pruebas que involucran distintas combinaciones de las categorías de clasificación descritas previamente. Por ejemplo hay comparaciones planeadas y ortogonales (contrastes ortogonales) y también no-planeadas que ofrecen control sobre la tasa de error familywise (prueba de Dunnett) o la tasa de error individual (LSD de Fisher). El lector interesado puede consultar una amplia gama de pruebas disponibles en la monografía de Toothaker (1993) y el libro de Steel y Torrie (1980). El trabajo de Tukey es una de las referencias consideradas más importantes en el tema de comparaciones múltiples, además de los artículos y monografías revisión que se publicaron en los 70 (Hochberg y Tamhane, 1987).

El software comercial incluye diversas pruebas de comparaciones múltiples como opciones dentro del análisis de varianza. La prueba de Tukey, según se indicó anteriormente, ofrece un buen control para la tasa global del error experimental cuando todos los pares de medias se comparan entre sí, involucra los siguientes pasos:

1. Ordenar los promedios de los tratamientos de menor a mayor

2. Determinar el valor del llamado rango estudentizado $q(1-\alpha; t =$ # niveles del factor, gle = grados de libertad disponibles para estimar el error experimental). Estos valores, generados por Tukey como

parte del desarrollo de la prueba, están disponibles en las tablas estadísticas de los libros sobre **diseño de experimentos** como el de Montgomery (2009) o Steel y Torrie (1980).

3. Calcular el estadístico de prueba, HSD = [q(α; t, gle)] (MSE/r)1/2 para el caso de igual número de réplicas (r) a cada nivel de factor, esto es un caso balanceado.

4. Iniciar comparando promedios extremos, esto es menor versus mayor; si la diferencia entre dos promedios es mayor a HSD, las medias de los tratamientos bajo comparación se declaran estadísticamente diferentes. Continuar comparando las siguientes medias hasta llegar a la comparación entre pares.

5. Presentar resultados subrayando con una misma línea conjuntos de medias declaradas iguales.

Ejemplo 4. El área de nuevos productos de la empresa fabricante de alimentos ya esperaba que los colores tuvieran diferente aceptación entre los consumidores, por tanto el objetivo de la experimentación se re-especificó a "identificar el color que induce las mayores ventas para el producto." Para atender este objetivo se requiere de una prueba de comparaciones múltiples, la elegida fue la prueba de Tukey. Esta prueba está implementada en MINITAB, la secuencia de comandos para su uso es la siguiente:

Las ventanas de diálogo que se despliegan se muestran en la Figura 6.1. En la ventana principal es necesario declarar la columna en la cual están capturados los datos para la respuesta de interés, en este caso Ventas de producto, en cuanto al factor en estudio éste se identifica como el color del producto. Puesto que el color está registrado en una escala nominal, se utilizan etiquetas de dígitos sucesivos para identificar a los varios colores (1 = rosa, 2 = verde pálido, 3 = azul y 4 = transparente).

El botón de Comparisons despliega las pruebas de comparaciones múltiples disponibles. Aparte de la prueba de Tukey, MINITAB realiza la prueba LSD de Fisher, la t-Dunnett que se aplica cuando uno de los niveles es un control contra el cual se desea comparar el resto de los tratamientos y la prueba de Hsu que compara la media de cada tratamiento contra la mayor o la menor de todas. La tasa de error global o para la familia de hipótesis de interés, excepto para la LSD de Fisher, se declara en el recuadro correspondiente una vez que el analista elije una opción de prueba; el default es 5%. En el caso de la prueba de Fisher, la tasa de error que se especifique corresponde al error del tipo I por cada hipótesis individual; la tasa global que aparece en el listado de salida de MINITAB será mayor puesto que ésta aplica para el conjunto completo de pruebas.

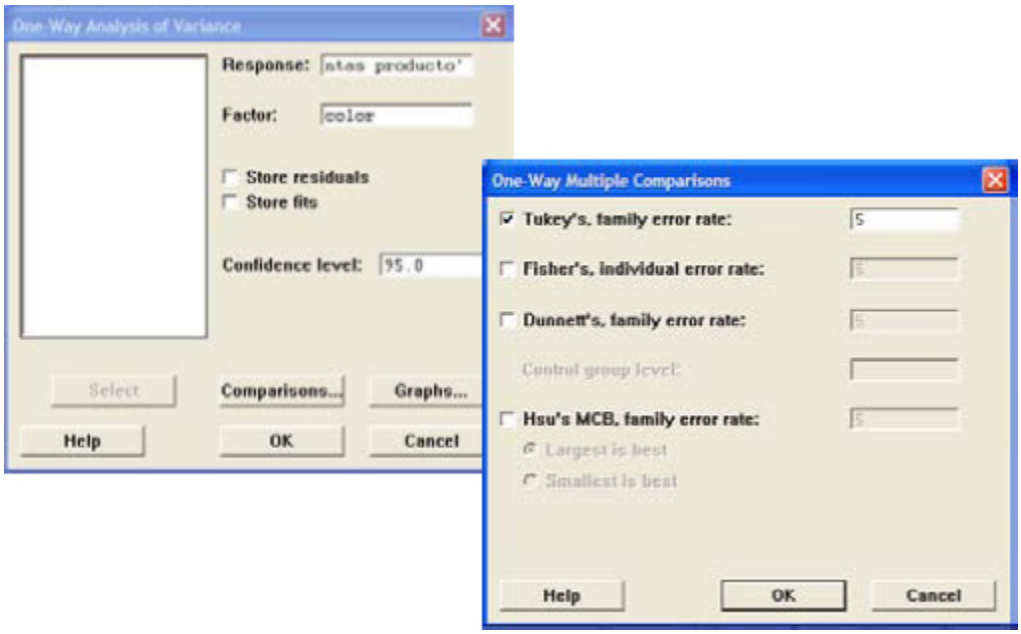


Figura 6.1 Pruebas de comparaciones múltiples en MINITAB

Los resultados de la prueba para este experimento con diferentes colores se dan en el formato de intervalos de confianza para la diferencia entre dos medias; el nivel de confianza global es 95% mientras que el correspondiente a un intervalo individual es superior o igual a 98.83 %. Cuando un intervalo incluye al cero implica que las dos medias son iguales mientras que un intervalo con límites del

mismo signo, + ó -, significa que las dos medias son estadísticamente distintas. En el listado, enseguida de cada intervalo de confianza, se indica la conclusión correspondiente en cuanto a la relación entre medias.

Tukey 95% Simultaneous Confidence Intervals

All Pairwise Comparisons

Individual confidence level = 98.83%

Rosa subtracted from:

	Lower	Center	Upper
verde pálido	-229.41	-97.50	34.41

ROSA = VERDE PÁLIDO, "0" incluído en el intervalo de confianza

azul	-329.41	-197.50	-65.59

ROSA > AZUL, ambos límites negativos lo que significa que promedio(rosa) es mayor que el promedio(azul)

transparente	-81.91	50.00	181.91

ROSA = TRANSPARENTE

Verde pálido subtracted from:

	Lower	Center	Upper
azul	-231.91	-100.00	31.91

VERDE PÁLIDO = AZUL

transparente	15.59	147.50	279.41

VERDE PÁLIDO < TRANSPARENTE

Azul subtracted from:

	Lower	Center	Upper
transparente	115.59	247.50	379.41

TRANSPARENTE > AZUL

Los conjuntos de medias declaradas estadísticamente iguales con base en la prueba de Tukey son: Azul = 320, Verde pálido = 420,

Rosa = 517.5 y Rosa = 517.5, Transparente = 567.5. En este caso los dos subconjuntos no resultaron mutuamente exclusivos ya que la respuesta media para una bebida color rosa se declaró igual a la media de ventas para una bebida transparente, pero también igual a los otros dos colores. A pesar de esto, es posible satisfacer el objetivo del estudio, las mayores ventas para el producto se logran cuando la bebida es color rosa, de sabor toronja, o bien cuando la bebida es transparente.

Entre las pruebas de comparaciones múltiples planeadas o sugeridas por la estructura de los datos o la teoría que sustenta el proyecto de investigación, destaca el uso de contrastes sobre todo los ortogonales (Hochberg y Tamhane, 1987; Steel y Torrie, 1980). Un contraste se define como una combinación lineal de medias de los tratamientos para la cual la suma de los coeficientes de la combinación es igual a cero; dos contrastes son ortogonales entre sí cuando la suma de los productos de sus coeficientes es cero, de lo contrario son no-ortogonales. La suma de cuadrados entre grupos se descompone en un conjunto de hasta (t-1) contrastes ortogonales asociados con un conjunto o familia de hipótesis de interés, lo que resulta interesante para comprender la estructura de las diferencias entre los niveles del factor. Pero no es obligado el aplicar este tipo de descomposición, el analista también tiene la opción de utilizar contrastes no-ortogonales y controlar por el nivel de significancia global, en este caso se recomienda el uso de la prueba de Scheffé (Steel y Torrie, 1980, p. 183).

Ejemplo 5. Considere de nuevo el experimento de los colores para la bebida energética. Dos de los colores sugieren el sabor del producto mientras los otros dos se asocian a otros conceptos, como frescura o naturalidad. Dada esta estructura para el experimento, de los tres grados de libertad entre tratamientos, colores, se podrían sugerir los siguientes contrastes:

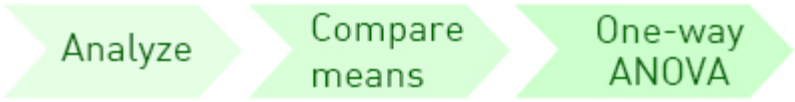
1. Comparar los colores que sugieren sabor con los colores que sugieren otro concepto, esto es H_0: μ (verde pálido) + μ (rosa) = μ (azul) + μ (transparente). Esta hipótesis se puede reescribir como H_0: μ (verde pálido) + μ (rosa) − μ (azul) −μ (transparente) = Q_1 =0. Bajo este formato el conjunto de coeficientes para el contraste Q_1 son: +1, +1, -1 y -1, los que, claramente, suman cero.

2. Contrastar los dos colores asociados al sabor del producto, H_0: μ (verde pálido) = μ (rosa). En este caso los coeficientes del contraste son: +1, -1, 0 y 0 puesto que las medias del color azul y transparente no están involucradas.

3. Comparar los colores no asociados con el sabor, es decir comparar un color que sugiere frescura contra el que sugiere naturalidad en el producto, H_0: μ (azul) = μ (transparente).

Para probar las hipótesis anteriores el estadístico de prueba es una t-Student, $t = \frac{Q}{s(Q)} = \frac{Q}{\sqrt{rMSE \sum_{i=1}^{k} c_i^2}}$ donde c_i son los coeficientes para las medias en el **contraste**.

Los contrastes anteriores pueden ser probados empleando SPSS, la secuencia de comandos es:



Una vez que se han seleccionado las columnas que contienen a la variable dependiente y al factor bajo estudio se elije la opción Contrasts y se van declarando los coeficientes del contraste; cada vez que se escribe un coeficiente hay que presionar el botón Add para su registro. Los resultados que reporta SPSS para el caso del primer contraste se dan en la Tabla 6.4. Notar que el supuesto de homogeneidad de varianza puede relajarse cuando se realizan estas comparaciones planeadas. Ya sea que se asuma homogeneidad de varianza o no, la prueba t es no-significante por lo que se concluye que las ventas medias no difieren si los colores usados se asocian al color de la bebida, rosa o verde pálido, respecto a si sugieren otro tipo de concepto, frescura o naturalidad. La hipótesis nula no se rechaza, en este caso, porque el color que sugiere el sabor toronja de la bebida registra ventas altas que compensan las bajas ventas del otro color, mientras que una situación similar se da con el otro par de tratamientos, la bebida transparente registra mejores ventas que la de color azul.

	Valor del contraste	Desviación estándar	Prueba t	gl	Significancia
Ventas Assume equal variances	50.0000	62.81587	.796	12	.442
Does not assume equal variances	50.0000	62.81587	.796	10.731	.442

Tabla 6.4 Prueba de hipótesis para el contraste entre colores asociados a diferentes conceptos

El uso de contrastes ortogonales sobre las pruebas no-planeadas se favorece bajo el argumento de que los contrastes representan un número específico de comparaciones planeadas entre medias no sugeridas por los datos. Sin embargo, según se vio previamente, el uso de una familia de contrastes implica múltiples pruebas y por tanto el riesgo de un incremento en la tasa de error por familia o experimento. En consecuencia se recomienda que cualquier conjunto de hipótesis, ya sea o no planeado, considere el riesgo potencial de un incremento en el riesgo de un error del tipo I como criterio para elegir una prueba de comparaciones múltiples (Bender Lange, 2001).

¿Sabías qué?

Los padres de J. W. Tukey (1915-2000) decidieron educar a su hijo en casa después de reconocer su notable potencial intelectual. Este tipo de educación influyó notablemente en Tukey quien obtuvo los grados de licenciatura y maestría en Química. Mientras realizaba estudios de doctorado en Química en la universidad de Princeton, decidió cambiar de área doctorándose en Matemáticas en 1939. Su gran talento para las matemáticas, y en particular la estadística, se reflejó en contribuciones importantes como técnicas para la estimación de espectros de series de tiempo, algoritmos rápidos para la transformación de Fourier y pruebas para hipótesis simultáneas en un experimento.

Actividad de repaso

6. Análisis de varianza

Calificar cada una de las siguientes declaraciones como falsa o verdadera

_____ La diferencia fundamental entre un modelo de análisis de varianza y uno de regresión es la escala de la variable de respuesta, nominal para el ANOVA y en escala de intervalo o razón para regresión

_____ El uso de t-Student como prueba de comparaciones múltiples tiene el problema de incrementar considerablemente la probabilidad de cometer un error del tipo I

_____ La prueba de Tukey permite realizar comparaciones múltiples entre parejas de tratamientos mientras controla por el error global de un rechazo incorrecto de la familia de hipótesis

_____ Uno de los recursos para controlar el efecto de variables externas en un experimento es manejarlas como covariables

_____ Un experimento factorial es aquel en que se tiene, al menos, una variable externa

_____ Las pruebas planeadas empleando contrastes son recomendadas para factores que tienen más de dos niveles

_____ En un ANOVA los cocientes de sumas de cuadrados siguen la distribución F siempre que se satisfagan los supuestos de independencia, normalidad y varianza constante

_____ Cuando en un experimento se tiene más de una variable de respuesta el análisis estadístico multivariante que se aplica es el MANOVA

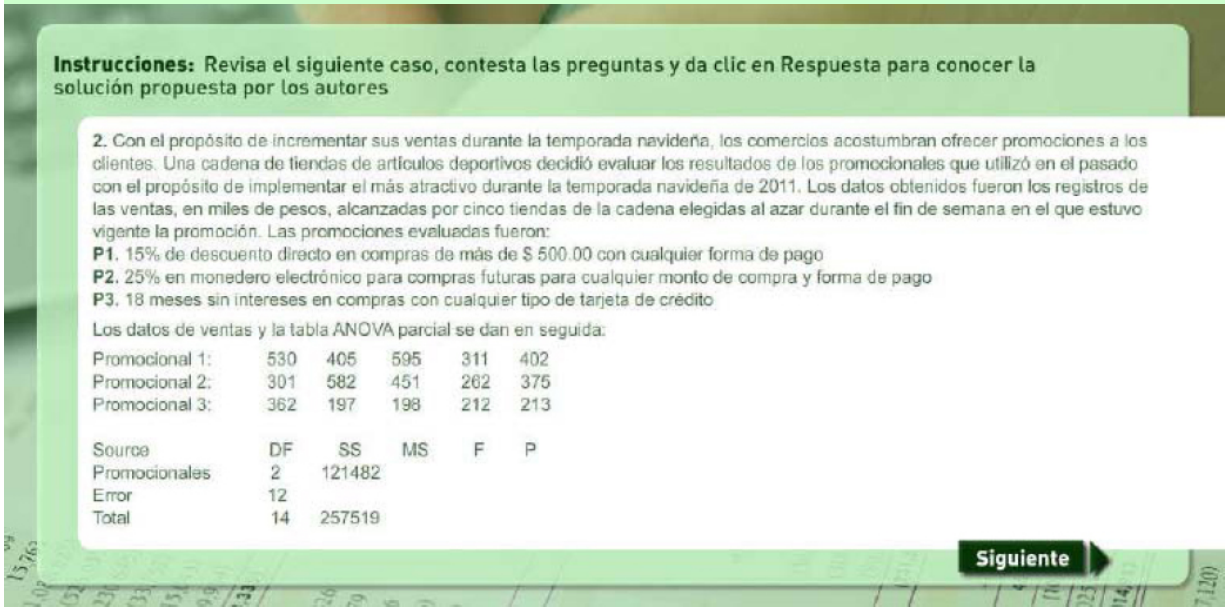
_____ En un diseño totalmente al azar no hay restricciones en cuanto a cómo asignar las unidades experimentales a tratamientos

____ El ANOVA es una técnica estadística multivariable que permite descomponer la variabilidad total observada para una respuesta en sumas de cuadrados asignadas a fuentes de variación bien definidas

Respuestas: 1-F, 2-V, 3-V, 4-V, 5-F, 6-F, 7-V, 8-V, 9-V, 10-V

Actividad de repaso (2)

6. Análisis de varianza



Instrucciones: Revisa el siguiente caso, contesta las preguntas y da clic en Respuesta para conocer la solución propuesta por los autores

2. Con el propósito de incrementar sus ventas durante la temporada navideña, los comercios acostumbran ofrecer promociones a los clientes. Una cadena de tiendas de artículos deportivos decidió evaluar los resultados de los promocionales que utilizó en el pasado con el propósito de implementar el más atractivo durante la temporada navideña de 2011. Los datos obtenidos fueron los registros de las ventas, en miles de pesos, alcanzadas por cinco tiendas de la cadena elegidas al azar durante el fin de semana en el que estuvo vigente la promoción. Las promociones evaluadas fueron:

P1. 15% de descuento directo en compras de más de $ 500.00 con cualquier forma de pago
P2. 25% en monedero electrónico para compras futuras para cualquier monto de compra y forma de pago
P3. 18 meses sin intereses en compras con cualquier tipo de tarjeta de crédito

Los datos de ventas y la tabla ANOVA parcial se dan en seguida:

Promocional 1:	530	405	595	311	402
Promocional 2:	301	582	451	262	375
Promocional 3:	362	197	198	212	213

Source	DF	SS	MS	F	P
Promocionales	2	121482			
Error	12				
Total	14	257519			

Siguiente

6.4 El experimento factorial

Cuando se consideran dos o más factores en un experimento, fijar los niveles de un factor mientras se varían los niveles del otro u otros factores es considerablemente ineficiente. La mejor forma de evaluar el efecto de los factores es moverlos simultáneamente, esto es, probar todas las posibles combinaciones entre ellos. Bajo este esquema se minimiza el total de pruebas a realizar y se generan datos para estimar la interacción entre los factores Batra y Jaggi (s. f.). Esta interacción ocurre cuando el cambio en la respuesta, al variar los niveles de un factor, depende de los niveles de otro u otros factores, en otras palabras hay interacción cuando se espera que ocurra un cambio adicional en la variable dependiente al combinar dos niveles particulares de los factores experimentales. El efecto de la interacción más los efectos directos de cada uno de los factores, denominados efectos principales, conforman la suma de cuadrados entre los grupos que están definidos por las distintas combinaciones de los niveles de los factores.

Para estimar el error experimental o variabilidad dentro de grupos en un factorial se requieren réplicas; una réplica implica repetir todas las posibles combinaciones de los factores. Por tanto la descomposición de la variación de la respuesta para el caso de un experimento con únicamente dos factores es:

$$SST = [SS(A) + SS(B) + SS(A*B)] + SSE$$

El modelo ANOVA para un experimento bifactorial que se muestra en (4) es aditivo-multiplicativo, el término multiplicativo describe el efecto de la interacción:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \qquad (4)$$

donde:

Y_{ijk}: k-ésima observación de la variable dependiente medida bajo los efectos del nivel i del factor 1 y el nivel *j* del factor 2

μ: media general o global de la variable dependiente sobre todas las combinaciones

α_i: es el efecto del nivel i del factor 1 manejado en el experimento.

β_j: representa el efecto del nivel j del factor 2 manejado en el experimento.

$(\alpha\beta)_{ij}$: representa el efecto interacción entre factores

ε_{ijk}: es el término aleatorio o error experimental

Los cálculos requeridos para completar el ANOVA de un experimento bifactorial se muestran en la Tabla 6.5. El término Y_{ij}. es el total para cada combinación de los factores en tanto Y_i y Y_j son los totales a cada nivel de los factores bajo estudio.

Fuente de variación	gl	Sumas de cuadrados	Cuadrados medios	F
Factor A	a-1	$\sum_{i=1}^{a} \frac{Y_{i.}^2}{rb} - FC$	SSA/(a-1)	MSA/MSE
Factor B	b-1	$\sum_{j=1}^{b} \frac{Y_{.j}^2}{ra} - FC$	SSB/(b-1)	MSB/MSE
	(a-1)(b-1)	$\sum_{i=1}^{a}\sum_{j=1}^{b} \frac{Y_{ij}^2}{r} - FC - SSA - SSB$	SS(AxB)/[(a-1)(b-1)]	MS(AxB)/ MSE
Residual	abr(r-1)	Por diferencia	SSE/[ab(r-1)]	
Total	abr-1	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r} Y_{ijk}^2 - FC$		

Tabla 6.5 Cálculos para el ANOVA en un sentido

Como en el caso de otros métodos multivariados, lo recomendable es que los cálculos para el ANOVA se realicen con apoyo de software estadístico. En MINITAB hay una opción específica para efectuar el análisis de varianza en dos sentidos que aplica al caso de dos factores en estudio: ANOVA two-way. La Figura 6.2 muestra todas las opciones que MINITAB ofrece para el ANOVA; la opción Two-way es muy simple de utilizar, basta con declarar en cuáles columnas se encuentran la variable de respuesta y los códigos para los dos factores en estudio.

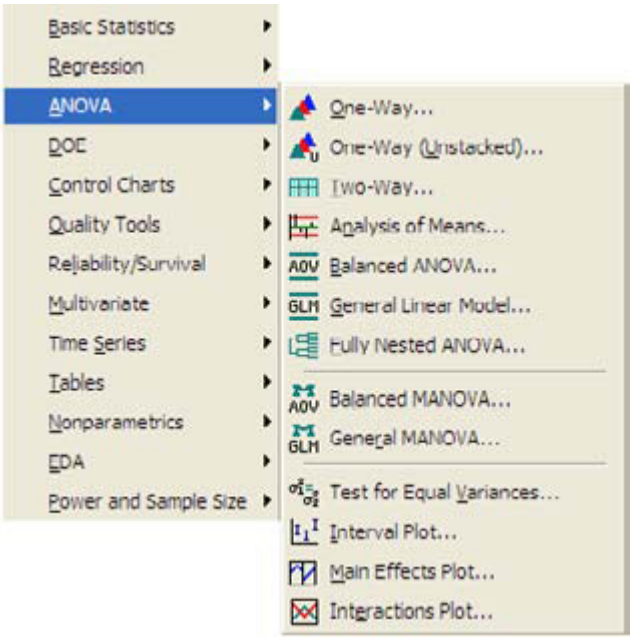


Figura 6.2 Diferentes opciones para el Análisis de Varianza en MINITAB

Ejemplo 6. Para el primer cuarto del 2011, el Internet World Stats estimó aproximadamente 2 mil millones de usuarios de Internet, aproximadamente el 30.2% de la población mundial, lo que representa un incremento de 480.4% respecto al año 2000. El mayor crecimiento en número de usuarios, de un 1,037.4% fue para América Latina y el Caribe (World Usage Patterns & Demographics, diciembre 5, 2011). Estos usuarios resultan ser un mercado potencial sumamente atractivo, sin embargo cuando estos individuos realizan búsqueda de productos y servicios, el 75% de ellos no pasa de la primera página del sitio (Marketshare.Hitslink.com, octubre 2010). Ante esto, empresas como Google han buscado que los sitios de Internet evolucionen hacia sitios de auto-servicio que ofrezcan soluciones valiosas e información completa al consumidor potencial. Para diseñar mejor estos sitios es importante tomar en consideración el perfil demográfico de los consumidores para así proporcionar la información y diseñar los esquemas de presentación que mejor se ajustan al perfil del consumidor potencial. Con este propósito se recolectó información sobre la utilidad percibida para los sitios de búsqueda de hoteles, expresada en una escala de 0-50 puntos donde mayor es mejor, según la apreciación de cuatro segmentos

de consumidores definidos en términos del género (factor A con niveles masculino y femenino) y rango de edad (factor B con tres niveles: 18-25 años, 26-45 años y mayores de 45 años) del usuario. Tres usuarios de cada segmento fueron entrevistados y el ANOVA para los datos de utilidad percibida se da a continuación. Es importante notar que en este caso los datos no son resultado de una experimentación puesto que las características de los usuarios de Internet no son manipulables. Recordar que si bien el ANOVA es particularmente útil en el contexto de diseño experimental, también es aplicable en el caso de datos de encuesta.

Two-way ANOVA: utilidad percibida versus edad, género

Source	DF	SS	MS	F	P
Edad	2	1064.11	532.056	31.50	0.000
Género	1	0.22	0.222	0.01	0.911
Interacción	2	184.11	92.056	5.45	0.021
Error	12	202.67	16.889		
Total	17	1451.11			

Las siguientes tres hipótesis estadísticas son de interés:

H_0: $\alpha_i = 0$ para toda i versus H_1: $\alpha_i \neq 0$ para al menos una $i = 1,2...a$

H_0: $\beta_j = 0$ para toda i versus H_1: $\beta_j \neq 0$ para al menos una $j = 1,2...b$

H_0: $(\alpha\beta)_{ij} = 0$ para toda i,j versus H_1: $(\alpha\beta)_{ij} \neq 0$ para al menos una i,j

Las pruebas *F* asociadas a cada una de las hipótesis anteriores se dan en la penúltima línea del ANOVA. Con base en los valores *P* asociados a estas pruebas se concluye que hay una interacción significante entre los dos factores. Cuando la interacción es significante, no es apropiado comparar niveles de un mismo factor puesto que la media de la respuesta no depende del cambio de un factor sino de cómo se combinan los niveles de los factores. Lo que procede es construir un gráfico que describa cómo interactúan los factores, tal gráfico se reporta en la Figura 6.3.

En el eje horizontal de este diagrama se representan los niveles de uno de los factores, usualmente el que tenga mayor número de niveles que en este caso es rango de edad, y en el eje vertical los correspondientes promedios de la variable de respuesta a cada combinación de los factores. El diagrama muestra dos gráficas, una a cada nivel del género del consumidor que es el otro factor. El gráfico se puede construir con MINITAB si se selecciona la opción *Interactions Plot* dentro del menú ANOVA que se mostró en la Figura 6.2.

A partir del gráfico de la Figura 6.3 se puede concluir que para el grupo de hombres, los que están en los dos menores rangos de edad perciben una utilidad baja para los sitios de búsqueda de hoteles en comparación con los hombres de más de 45 años. En contraste, en el grupo de mujeres son aquellas en el rango de edad intermedio, 26-45 años, las que tienen la menor percepción de utilidad, el grupo de mujeres jóvenes percibe una mayor utilidad de los sitios de búsqueda de hoteles.

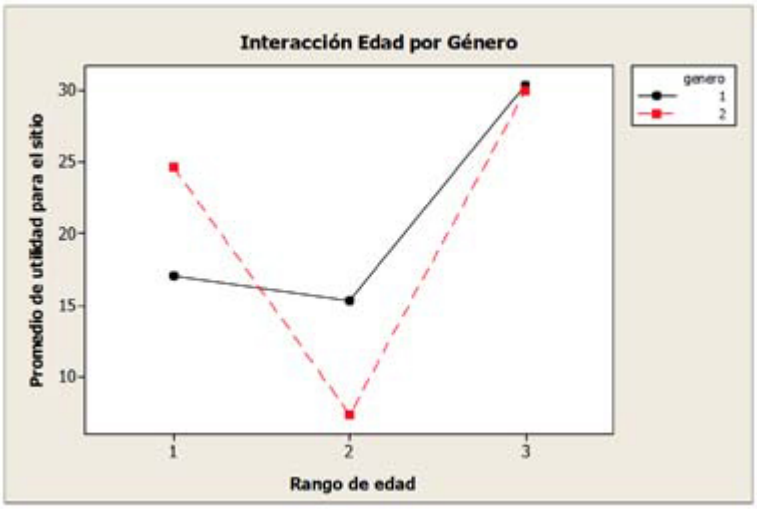


Figura 6.3 Descripción gráfica para la interacción entre dos factores

Cabe aclarar que en aquellos casos en que la interacción resulte nosignificante, el gráfico de interacción se sustituye por un gráfico de efectos principales. También se aplicaron pruebas de comparaciones múltiples para investigar las diferencias específicas

que hay entre los varios niveles de los factores declarados significantes.

En caso de que un experimento incluya más de dos factores, aparte de efectos principales adicionales, también habrá un mayor número de interacciones. Aquellas interacciones entre dos factores o de dos letras se conocen como interacciones de primer orden y son efectos que resulta importante estimar. Por el contrario, las interacciones de mayor orden, tres y más letras, suelen ser poco relevantes y, más que estimarlas sus grados de libertad asociados, se utilizan para obtener un estimado del error experimental o se aprovechan para reducir el número de combinaciones experimentales a probar. El fraccionamiento de factoriales, esto es la reducción en el número de combinaciones de un experimento con un gran número de factores, es uno de los recursos estadísticos más valiosos en experimentación. El fraccionamiento de factoriales permite un ahorro substancial en el número de corridas experimentales sin detrimento en la cantidad de información requerida para elegir aquellos factores o condiciones que resultan en los mejores valores para la respuesta (Montgomery, 2009). En el contexto de análisis conjunto (ver siguiente capítulo) el diseño de nuevos productos se lleva a cabo variando simultáneamente una gran cantidad de atributos lo que implica el uso de fracciones de factoriales tanto regulares como complejas.

Ejemplo 7. Un consorcio de agencias de investigación de mercados evaluó el efecto que tiene el costo del servicio (alto, promedio y bajo), el alcance del servicio (si la agencia realizó únicamente parte del proyecto o bien lo condujo totalmente) y el nivel de participación de la compañía compradora (el cliente participó en la investigación o dejó toda la responsabilidad a la agencia) sobre la evaluación que hace el cliente de la calidad de los proyectos realizados por la agencia de investigación. Esta variable de respuesta se midió como un índice cuya base es 100%, que representa cumplimiento total de expectativas. Valores menores del 100% corresponden a proyectos que no lograron satisfacer las expectativas de los clientes mientras valores sobre 100% significan que se excedieron expectativas. Se entrevistaron a múltiples

clientes hasta obtener dos evaluaciones a cada combinación de los tres factores, lo que dio un total de 24 observaciones (a cada una de las 3x2x3 combinaciones hay dos réplicas). Como en el ejemplo 6, este estudio no es un experimento, los datos fueron obtenidos a partir de una encuesta con clientes de las agencias del consorcio.

Para realizar el ANOVA de un experimento con más de dos factores en MINITAB, se pueden utilizar las opciones de *Balanced* ANOVA o GLM (General Linear Model). Esta última rutina es la opción más general y es equivalente a la opción de análisis *Univariate* GLM de SPSS. GLM permite declarar cualquier tipo de modelo ANOVA, ya sea balanceado o desbalanceado. La rutina en ambos paquetes estadísticos permite considerar **covariables**, realizar pruebas de comparaciones múltiples diversas, y construir gráficos para interacciones y efectos principales. La opción de *Balanced ANOVA* en MINITAB es más restrictiva ya que requiere de un igual número de réplicas a las varias combinaciones de los factores, no permite el uso de covariables ni tampoco llevar a cabo pruebas de comparaciones múltiples. Para utilizar tanto esta opción como GLM es necesario que el usuario declare el modelo estadístico que describe a los datos, esto es los componentes de modelos como los reportados en (1) y (2). Como la media general y el término aleatorio siempre están presentes en un modelo ANOVA, lo que se requiere es indicar cuáles efectos son fuentes de variación relevantes. En un ejemplo como éste en el cual se estudian tres factores y se cuenta con dos réplicas se pueden estimar los efectos principales (A, B y C), las tres interacciones de primer orden o dos letras (A*B, A*C y B*C) y la interacción de tres letras o segundo orden (A*B*C). En la ventana de diálogo de Balanced ANOVA que se muestra en la Figura 6.4. Es posible declarar todos estos efectos en el recuadro Model o bien utilizar una notación simplificada que indica a MINITAB, que todos los efectos del factorial antes mencionados son requeridos, esta notación es la siguiente: A|B|C. Si la interacción de tres letras se considera no-significante, es necesario omitirla del modelo, para simplificar la declaración de este modelo reducido, basta con restar aquellos efectos no deseados del factorial, según se muestra en la ventana de diálogo principal de la

Figura 6.4. A este punto es importante recordar que el ANOVA se sustenta en tres supuestos básicos para los términos de error: independencia, varianza constante y normalidad.

En la Figura 6.4 se muestra la ventana de diálogo secundaria que se activa después de elegir el botón Graphs; los supuestos del ANOVA se validan a través de un análisis gráfico de residuos, similar a como se hizo para el análisis de regresión (ver Capítulo 4).

En la ventana de diálogo principal para esta rutina, aparte del cuadro para declarar el modelo, hay otro cuadro para indicar si alguno de los factores en estudio es aleatorio. Un factor es fijo si sus niveles son los únicos que se van a probar y el factor se denomina aleatorio si los niveles que se prueban son una muestra al azar de una población de posibles valores. En el contexto de mercadotecnia usualmente no se experimenta con factores aleatorios, pero en áreas de ingeniería estos factores son comunes. Ejemplos de factores aleatorios son los lotes de materia prima que se usan en un proceso de manufactura o los operarios a cargo de maquilar un producto. La variabilidad en la calidad de un lote de materia prima o en las habilidades de un operador pueden inducir variabilidad en la calidad del producto que se elabora; por supuesto se usan distintos lotes cada vez que se fabrica el producto por lo cual más que comparar un lote contra otro, lo que equivale a contrastar niveles específicos del factor, lo que es de interés es determinar qué tanto se afecta la calidad del producto si varía la calidad de la materia prima. Cuando hay factores de este tipo en un proyecto, éstos se declaran aparte ya que su análisis requiere de la construcción de pruebas F especiales. El lector interesado puede consultar el capítulo 13 del libro de D. C. Montgomery (2009) para aprender más sobre esta clase de factores.

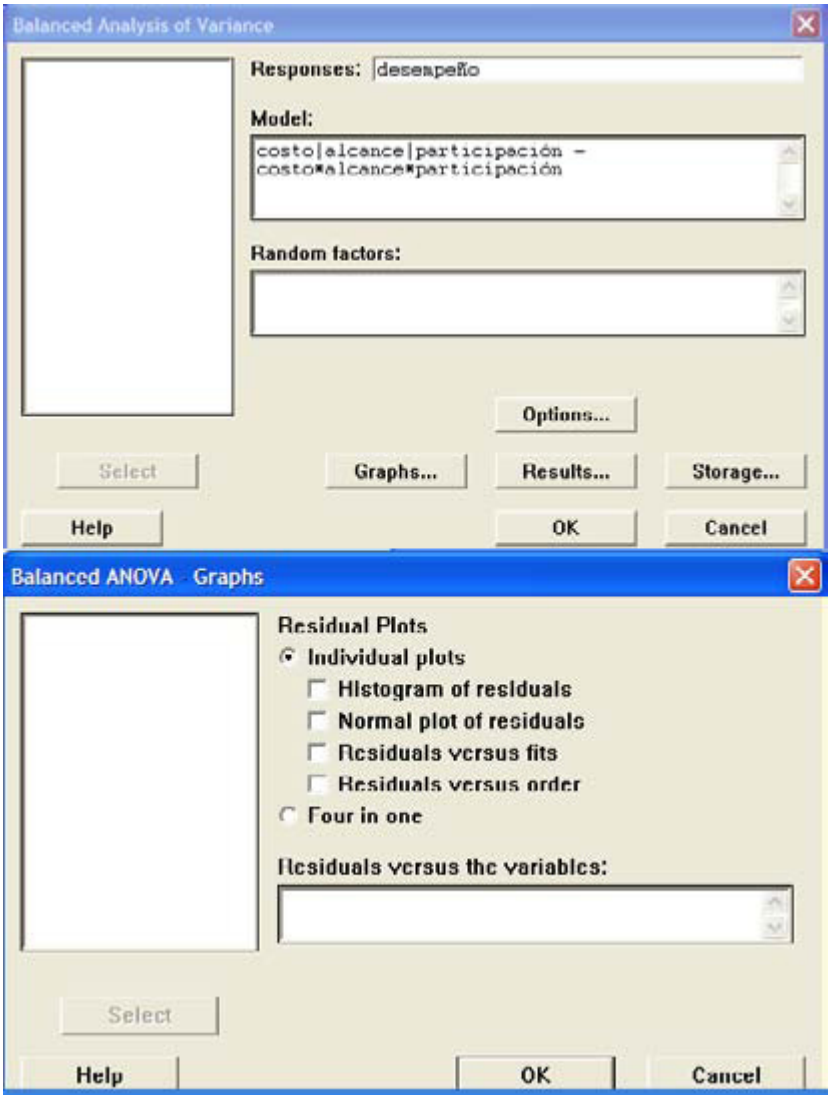


Figura 6.4 Ventanas de diálogo para el caso balanceado del ANOVA

El listado de salida para este ejemplo se reporta a continuación, notar que en la primera parte del listado MINITAB identifica el tipo de factores en estudio, en este caso todos son fijos.

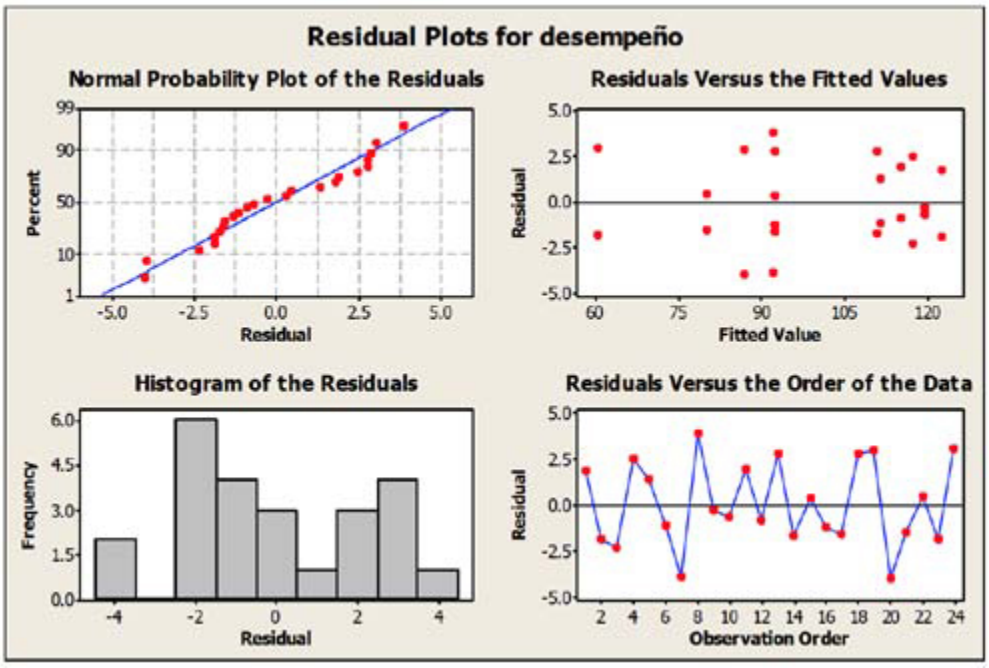
ANOVA: desempeño versus costo, alcance, participación

Factor	Type	Levels	Values
costo	fixed	3	1, 2, 3
alcance	fixed	2	1, 2
participación	fixed	2	1, 2

Analysis of Variance for desempeño

Source	DF	SS	MS	F	P
costo	2	4857.29	2428.64	284.84	0.000
alcance	1	870.01	870.01	102.04	0.000
participación	1	1906.38	1906.38	223.59	0.000
costo*alcance	2	1.54	0.77	0.09	0.914
costo*participación	2	13.56	6.78	0.80	0.471
alcance*participación	1	300.33	300.33	35.22	0.000
Error	14	119.37	8.53		
Total	23	8068.48			

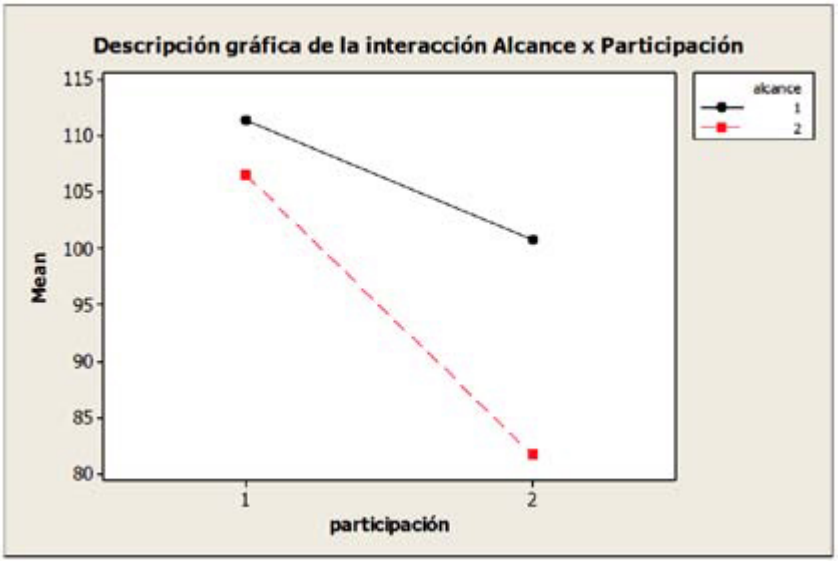
S = 2.91997 R-Sq = 98.52% R-Sq(adj) = 97.57%



El conjunto de gráficos que constituyen el análisis para los residuos del modelo no revela desviaciones a los supuestos del ANOVA ni tampoco la presencia de observaciones inusuales. El gráfico de probabilidades acumuladas versus residuos ordenados, izquierda superior, podría sugerir alguna desviación a la normalidad ya que los puntos parecen desviarse de la línea recta en su parte central, sin embargo la prueba de bondad de ajuste de Anderson-

Darling tiene asociado un valor P = 0.148 por lo que el supuesto no se disconfirma, recordar que para realizar la prueba de bondad de ajuste es necesario guardar los residuos en la hoja de trabajo y elegir la opción Storage en la ventana de diálogo principal.

En cuanto al análisis estadístico para los datos, el coeficiente de determinación *R-Sq* que se reporta el final de la tabla ANOVA permite concluir que los tres factores explican notablemente bien la variabilidad en las evaluaciones de los proyectos de investigación de mercados. Notar que los grados de libertad son 14 = abc (r-1) + (a-1)(b-1)(c-1) = (3x2x2)(2-1) + 2 gl de la interacción de tres letras. Las pruebas F indican que los tres efectos principales (costo del proyecto, alcance y participación del cliente) son altamente significantes (valores P = 0.000) pero también lo es la interacción (Alcance x Participación), por lo que, más que examinar las calificaciones promedio por nivel de costo o participación, lo que procede es interpretar la interacción apoyándose del gráfico correspondiente, el cual se muestra enseguida:

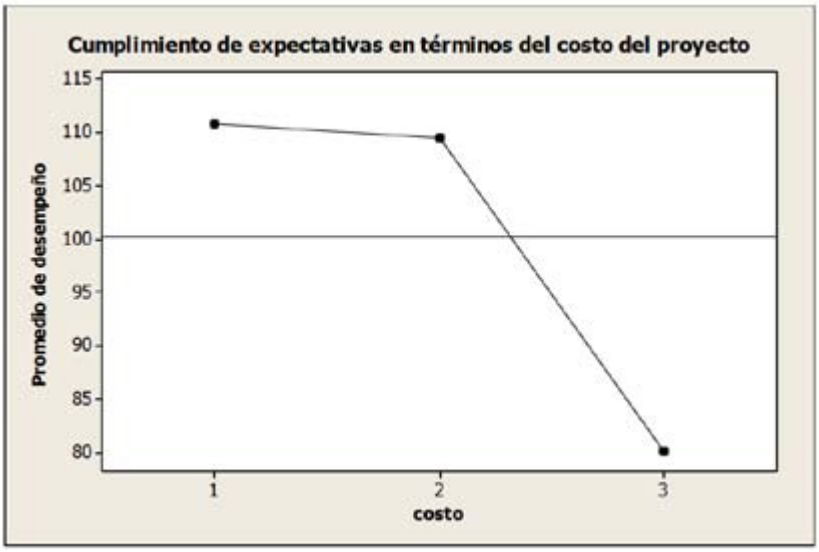


El gráfico anterior permite establecer que cuando el cliente no participa en el proyecto de investigación de mercados, nivel 2 del factor, es más difícil que el proyecto satisfaga sus expectativas.

Cuando además el proyecto es realizado en su totalidad por la agencia de investigación de mercados(alcance 2) su calidad, en

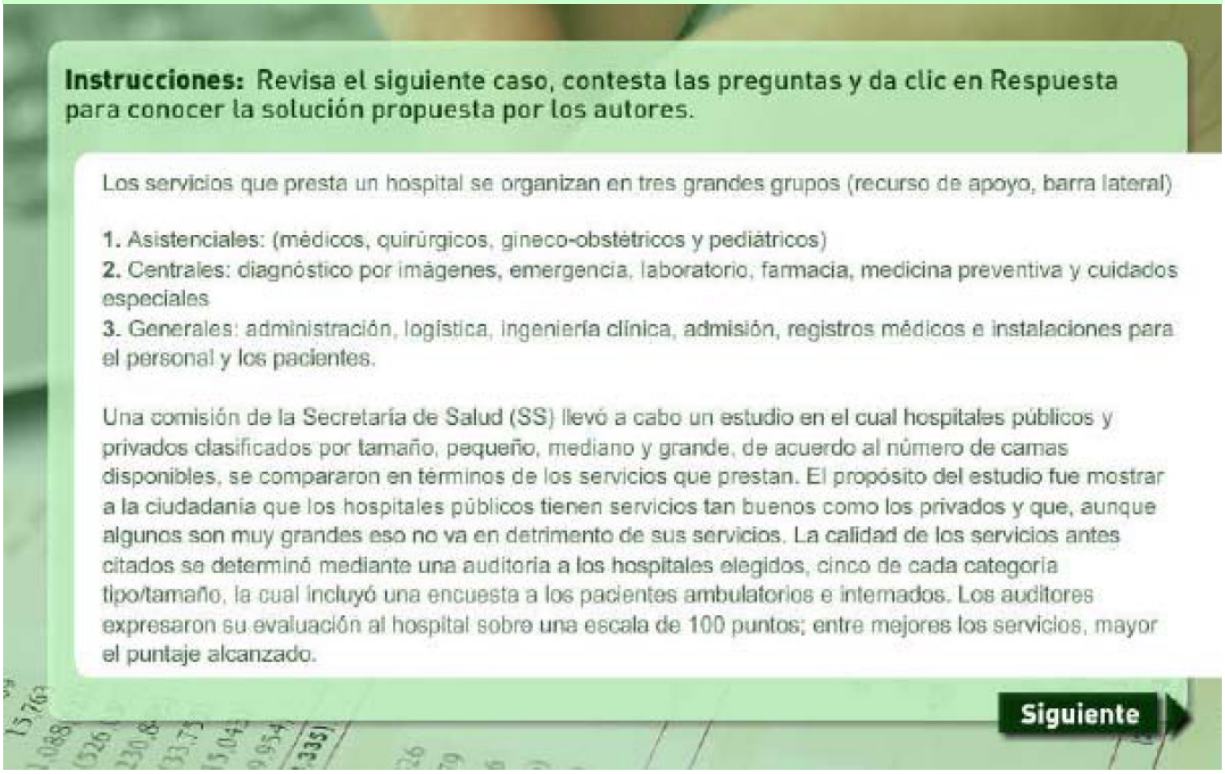
términos de cumplir expectativas, se juzga mucho peor respecto a los proyectos en los cuales la agencia de investigación participó solo con una parte (alcance 1).

Para el factor costo, el cual se declaró significante pero que no interactúa con los otros dos factores, es apropiado explorar diferencias entre niveles a través de un gráfico de efectos principales. Este gráfico muestra el promedio de la respuesta a cada uno de los niveles de costo y se puede elaborar eligiendo la opción Main Effect Plot dentro del menú ANOVA que se muestra en la Figura 6.2. Para este ejemplo, el gráfico revela que cuando el costo del proyecto es muy alto, por arriba del promedio de las cotizaciones presentadas por otras agencias, es difícil cumplir con las expectativas, el promedio de cumplimiento es de 80%; en contraste si el costo es bajo a medio, los proyectos sí satisfacen y aún exceden las expectativas del cliente.



Actividad de repaso

6. Análisis de varianza



6.5 Otros diseños experimentales

En la segunda sección de este capítulo se indicó que entre los propósitos centrales del diseño experimental está el control de variables externas con el propósito de eliminar fuentes de variabilidad adicional y estimar apropiadamente el efecto que los factores en estudio tienen sobre la respuesta. Los experimentos presentados en secciones anteriores no han hecho un control explícito de las variables externas; la elección y asignación aleatoria de unidades a condiciones experimentales ha sido el único recurso para cancelar o atenuar el efecto de estas variables. En esta sección se utiliza la estrategia de bloqueo o agrupamiento de unidades experimentales con el propósito de eliminar posibles fuentes de variación adicionales que se sabe tienen una influencia sobre la variable dependiente o de respuesta.

El llamado **diseño de bloques** consiste en formar grupos de unidades experimentales similares entre sí en relación a una variable externa identificada por el investigador. Por ejemplo, los individuos que evalúan nuevos cereales pueden variar significativamente en cuanto a sus niveles de consumo del producto. Los consumidores ocasionales del producto podrían tender a calificar pobremente los nuevos productos porque los cereales en general no les gustan mucho.

Una forma de corregir o atenuar estas diferencias puede ser simplemente formar al azar grupos de consumidores y a cada integrante de un grupo solicitarle que califique alguno de los nuevos productos; los grupos así formados incluirían individuos cuyo consumo de cereal es variable y por tanto al compararlos entre sí las diferencias entre personas se cancelarían. Sin embargo, un mejor diseño sería agrupar primero a los consumidores en términos de su nivel de consumo de cereales, bajo, regular y alto, y posteriormente pedir a los individuos de cada grupo que evalúen alguno de los nuevos productos elegido al azar. Bajo este esquema, como todos los individuos en un mismo grupo o bloque son similares en cuanto a su consumo de producto, si hay diferencias en sus

evaluaciones éstas se atribuirán al nuevo producto bajo evaluación mas no a sus preferencias individuales en cuanto a los cereales. Si todos los nuevos productos de cereal son evaluados por únicamente uno de los consumidores de cada bloque se tendrá un Diseño de Bloques Completamente al Azar (RCBD, Randomized Complete Block Design, por sus siglas en inglés). Este es el diseño de bloques más simple e implica que el tamaño de los bloques es igual al número de niveles del factor en estudio. Pero no siempre se pueden tener bloques cuyo tamaño coincida con el número de tratamientos.

Así se describe en el siguiente ejemplo de pruebas con champús:

Ejemplo 8. Considere un experimento en el cual se desea evaluar por consumidores la calidad de cinco champús hidratantes. La forma de realizar la prueba fue dividir la cabellera del consumidor en dos, en cada parte de la cabeza se aplicó un producto asignado al azar. Esta prueba de medias cabezas se hace porque las características particulares del cuero cabelludo influyen también en la efectividad del champú, sin embargo dividir la cabellera en cinco partes no es una opción razonable por lo cual el máximo de productos que se puede evaluar por persona es de dos. En este caso el diseño es uno de bloques pero incompletos, si todos los productos se prueban el mismo número de veces el diseño será de Bloques Incompletos Balanceados, cuando algunos champús se prueban más veces que otros el diseño será de Bloques Incompletos Parcialmente Balanceados. La generación y análisis de estos diseños es más elaborada que para el RCBD pero al ser menos restrictivos, permiten el uso de bloques en múltiples aplicaciones (John, 1971).

El modelo estadístico lineal para describir el experimento unifactorial en un diseño de bloques es el siguiente:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}(0, \sigma^2) \qquad (5)$$

Donde α_i es el efecto asociado a los niveles del factor (i = 1,2, …,k) y β_j describe el efecto de los bloques (j = 1,2,…,b).

Note la similitud del modelo (5) con aquél que describe al experimento bifactorial en un diseño totalmente al azar (4) en este caso, el modelo es totalmente aditivo porque se asume que no hay interacción entre la variable de bloqueo y el factor en estudio. Los bloques son un artificio para el control del error experimental por lo que la hipótesis de igualdad en la respuesta media entre bloques no resulta ser estadísticamente correcta puesto que los bloques fueron formados intencionalmente, son una restricción a la aleatoriedad del diseño y no una variable bajo prueba. Puesto que la interacción no es un efecto que se desea estimar, no es necesario contar con réplicas, es decir que cada tratamiento se prueba únicamente una vez por bloque y los grados de libertad $(k-1)(b-1)$ que corresponderían a la interacción son los que tiene asignados el estimado del error experimental.

Ejemplo 9. En el experimento con distintos colores para una nueva bebida energética, es claro que las unidades de supermercado a las que se asignó al azar lotes de bebida de distintos colores para la venta, resultaron diferentes en sus características. Los supermercados son de cadenas con diferente presencia en el mercado, algunos son de reciente creación y ubicados tanto en zonas céntricas como en centros comerciales. Esta heterogeneidad en las unidades experimentales contribuyó de manera importante a la variabilidad en las ventas; es evidente que

un supermercado de cadena líder, ubicado en una zona de alta afluencia reportó ventas fuertes para el nuevo producto aún cuando el color de éste no fuera el más atractivo. Para mejorar el diseño del experimento se propuso repetir la prueba pero considerando el efecto que tiene la cadena de supermercados sobre las ventas registradas para la nueva bebida. Cuatro tiendas de cada una de las cuatro principales cadenas de supermercados con presencia nacional, todas ubicadas dentro de la zona metropolitana de la ciudad de Toluca, fueron elegidas al azar. Las cadenas participantes, en orden según su participación en el mercado mexicano, fueron: Wal-Mart, Comercial Mexicana, Soriana y Chedraui. También al azar, se surtió un lote de bebida de cierto color a cada una de las tiendas de la misma cadena. Después de una semana se registraron las ventas de cajas de producto por supermercado participante.

En este experimento el factor bajo estudio es el color del producto mientras que la cadena de supermercados es la variable externa que se controla al integrar bloques de cuatro tiendas de la misma cadena. Cada tienda prueba únicamente un color del producto por lo que el diseño es uno de bloques totalmente al azar. Para procesar los datos se utiliza la opción Two Way ANOVA de MINITAB. Pero como en este caso el modelo (5) es totalmente aditivo, es necesario declararlo. Para ello, en la ventana de diálogo principal, que se muestra en la Figura 6.5, hay que elegir el recuadro Fit additive model. Los datos se capturan como se muestra en la hoja de trabajo, cada columna representa una variable: la cadena (codificada), el color del producto (codificado) y las ventas (en número de cajas vendidas). MINITAB no distingue entre factor y bloques, esto lo sabe el investigador de mercados, así que hay que declarar a los bloques como si fuera un factor y omitir los detalles de su prueba de hipótesis ya que el propósito del estudio no es demostrar que hay diferencias entre las cadenas de supermercados sino que los colores de la bebida inducen la venta.

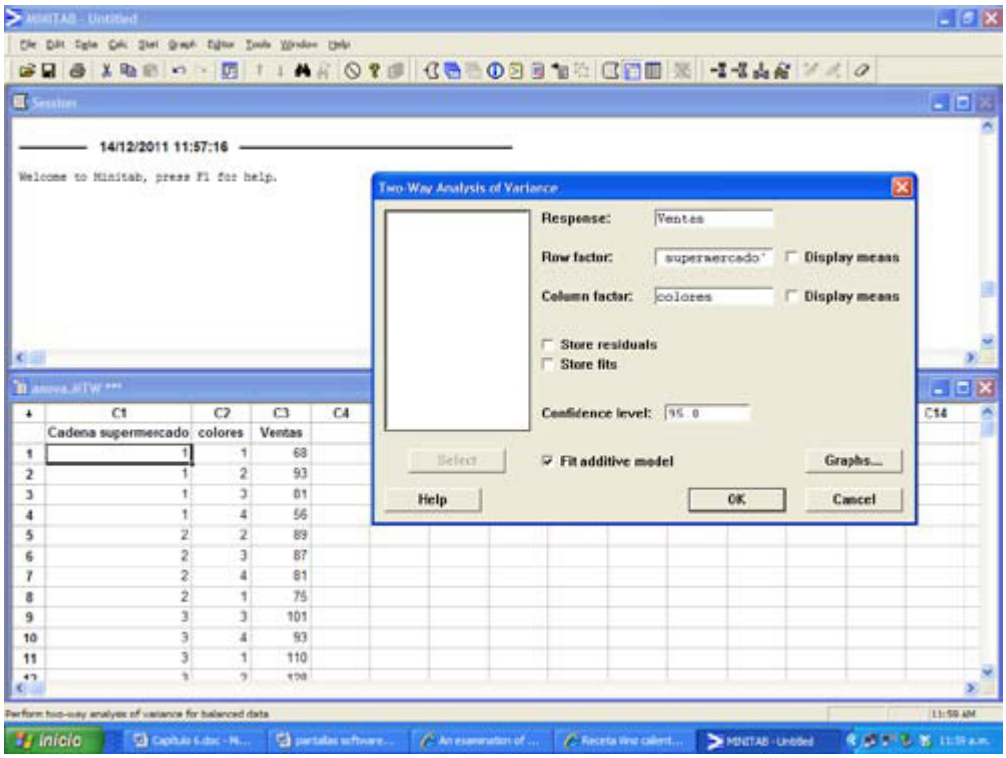


Figura 6.5 Análisis de varianza para el experimento en bloques en MINITAB

El análisis de varianza es el siguiente:

Two-way ANOVA: Ventas versus Cadena supermercado, colores

Source	DF	SS	MS	F	P
Cadena de supermercado	3	30507.7	10169.2	68.30	0.000
Colores	3	3139.7	1046.6	7.03	0.010
Error	9	1340.1	148.9		
Total	15	34987.4			

S = 12.20 R-Sq = 96.17% R-Sq(adj) = 93.62%

La única hipótesis de interés de acuerdo con el modelo (5), es

H_0: $\alpha_i = 0$ para toda $i = 1,2,3,4$ versus H_1: $\alpha_i \neq 0$

La cual se rechaza con un nivel de significancia estimado de 1% (P = 0.010). Si bien la prueba F asociada a los bloques, cadenas de supermercados, no es válida, la magnitud de la suma de cuadrados entre bloques (SSBloques = 30,507.7) muestra que fue un acierto su uso ya que esto reduce considerablemente la variabilidad de la respuesta. Para determinar qué tan importantes fueron los bloques en el control de la variación de las ventas lo apropiado es comparar el error experimental del diseño con y sin bloques. En el ejemplo 3 donde la cadena de supermercados no fue controlada MSE (diseño totalmente al azar) = 3,946 en tanto en este diseño modificado MSE (bloques totalmente al azar) = 148.9. El cociente de estos cuadrados medios es MSE(CRD)/MSE(RCBD) = 26.5 lo que implica que el diseño de bloques es 26.5x100 = 2,650 más eficiente que el diseño totalmente al azar, en otras palabras, 2,650 réplicas en un diseño totalmente al azar darían la misma cantidad de información que 100 bloques o réplicas para un diseño de bloques totalmente al azar (Steel y Torrie, 1980 p. 215).

El concepto de bloques puede extenderse para considerar más de una variable externa. Cuando dos variables externas se utilizan para agrupar a las unidades experimentales se da lugar al diseño en **cuadro latino**. En este tipo de diseño las unidades experimentales se agrupan en términos de dos variables, esto es por columna y por renglón dentro del cuadro, los niveles de los tratamientos se representan por letras latinas (A, B, C, etc.) de ahí el nombre del diseño el cual es altamente balanceado: el tamaño del cuadro latino debe coincidir con el número de niveles del **factor** que se estudia. Lo anterior impone restricciones al uso del diseño que implica un número razonable de corridas experimentales solo si el número de niveles del factor está entre cinco y ocho; de otra forma se vuelve incosteable el experimento en este diseño.

En un diseño en cuadro latino cada tratamiento ocurre una y solo una vez en cada columna o renglón, el modelo ANOVA correspondiente está dado en (6)

$$Y_{ijk} = \mu + \alpha_i + \rho_j + \varepsilon_{ijk} \quad (6)$$

En el modelo anterior, αi es el efecto del i-ésimo nivel del factor en estudio, ρj representa el efecto de la variable externa asignada a los renglones y γ_k el efecto de la variable externa de las columnas; los términos de error ε_{ijk} se asumen variables independientes, normales y de varianza constante. El modelo (6) es un modelo completamente aditivo ya que se asume de nuevo que las variables externas, tanto de renglones como de columnas, no interactúan con el factor bajo estudio.

Ejemplo 10. La gerencia de nuevos productos de la empresa fabricante de la recién creada bebida energética ha decidido ampliar el mercado de prueba y evaluar los colores del producto en supermercados de varias ciudades. De esta manera se espera incrementar la validez externa del estudio y asegurar que el color de la bebida tendrá aceptación nacional. Se sabe que la preferencia por productos de esta categoría varía por zona, en el centro del país esta categoría de productos tiene altas ventas, pero en el área del Golfo de México las ventas son, más bien, bajas. Dadas estas diferencias, se eligieron cuatro ciudades de prueba, ubicadas en zonas, para las cuales las ventas de bebidas energéticas varían fuertemente: Aguascalientes, zona Pacífico, Veracruz en el Golfo de México, Monterrey, zona Norte, y Querétaro en la zona centro. En cada ciudad se gestionó que un supermercado, de cada una de las cuatro cadenas elegidas, probara durante una semana al azar alguno de los colores de la bebida. El diseño resultante fue uno de cuadro latino, con la cadena de supermercados como variable en las columnas y la ciudad como variable en los renglones.

Los colores de la bebida fueron asignados al azar a las letras del cuadro (A = verde pálido, B = transparente, C = rosa y D = azul), después de una semana se registró el número de paquetes de bebida vendidos, los datos se muestran en la Tabla 6.6.

Ciudad de prueba	Wal-mart	Comercial Mexicana	Soriana	Chedraui
Aguascalientes	A = 150	C = 101	C = 89	A = 68
Veracruz	B = 173	D = 93	D = 87	B = 93
Monterrey	D = 223	B = 110	B = 81	D = 81
Querétaro	D = 195	B = 128	B = 75	D = 56

Tabla 6.6 Diseño en cuadro latino para prueba de colores de una bebida energética

Para analizar los datos anteriores se usa la rutina GLM de MINITAB, de nuevo no hay distinción entre variables externas y factores; la diferencia está en el modelo que especifica el usuario, el cual según se muestra en la Figura 6.6, es totalmente aditivo a diferencia del modelo para un experimento en tres factores como el del Ejemplo 7. Los resultados para el análisis de varianza que reporta MINITAB son:

General Linear Model: ventas bebida versus ciudad, cadena, colores

Factor	Type	Levels	Values
ciudad	fixed	4	1,2,3,4
cadena	fixed	4	1,2,3,4
colores	fixed	4	1,2,3,4

Analysis of Variance for ventas bebida, using Adjusted SS for Tests

Source	DF	Sen SS	Adi SS	Adi MS	F	P
ciudad	3	954.7	954.7	318.2	4.95	0.046
cadena	3	30507.7	30507.7	10169.2	158.33	0.000
colores	3	3139.7	3139.7	11046.6	16.29	0.003
Error	6	385.4	385.4	64.2		
Total	15	34987.4				

S = 8.01431 R-Sq = 98.90% R-Sq(adj) = 97.25%

Como en el caso del diseño de bloques, la única hipótesis estadística de interés es aquella que compara los varios niveles del factor en estudio, en este caso los colores de la bebida. La prueba F asociada es altamente significante (P = 0.003) concluyéndose que hay diferencias en las ventas promedio dependiendo del color que se haya dado a la bebida energética. Si bien las pruebas F para las variables externas Ciudad y Cadena de supermercados no son válidas, la magnitud de las sumas de cuadrados y el cociente F es evidencia de que estas variables contribuyen a la variabilidad de las ventas y que, al considerarlas, se redujo el error experimental.

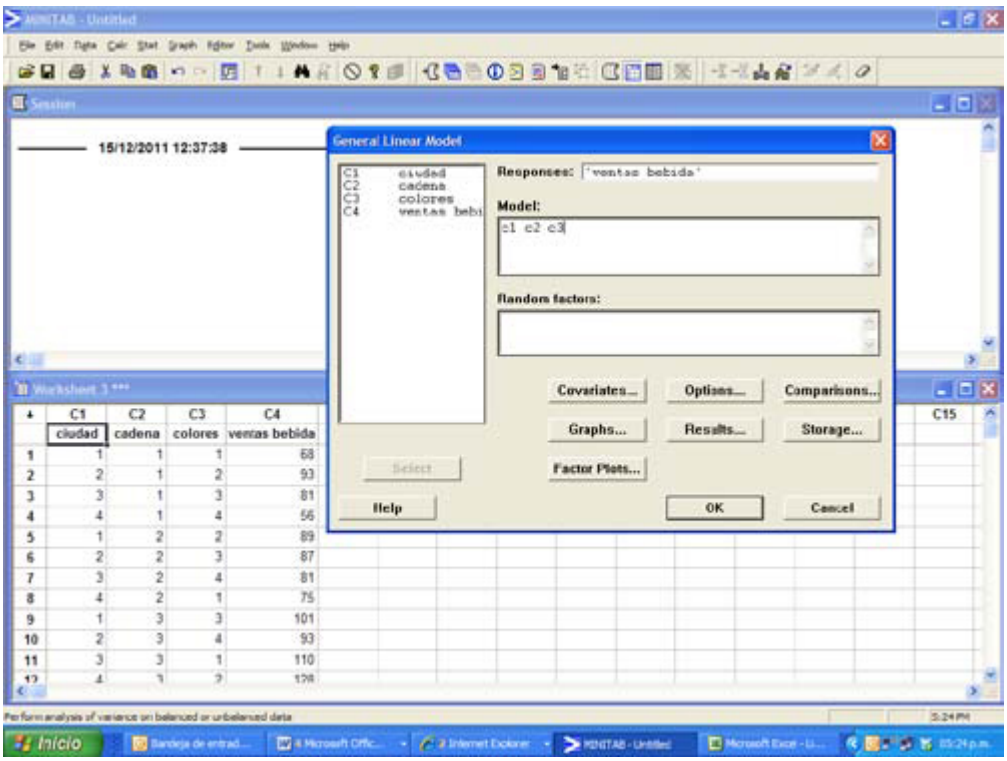


Figura 6.6 Análisis de un experimento en cuadro latino empleando GLM en MINITAB

6.6 Análisis de covarianza (ANCOVA)

Esta técnica se utiliza en casos en los que el investigador desea controlar el efecto de una variable externa que está en una escala métrica, y que recibe el nombre de variable concomitante o covariable. Estas variables externas deben observarse antes de efectuar el estudio, si se miden durante su realización es necesario asegurar que no se verán afectadas por las condiciones experimentales ya que de otra forma el análisis de covarianza eliminará una gran parte del efecto que los tratamientos tienen sobre la respuesta (Neter, Wasserman y Kutner, 1990).

El análisis de covarianza es una mezcla de dos técnicas: análisis de varianza y análisis de regresión. Se lleva a cabo en forma secuencial (Steel y Torrie, 1989), en una primera etapa se realiza un análisis de regresión para estimar la influencia que la covariable tiene sobre la variable dependiente; posteriormente se ajustan las medias de los tratamientos $\mu_i = \mu + \tau_i$ por las diferencias en los

valores de la covariable para compararlas entre sí, $(\overline{X})$ de la covariable. Esta situación se describe gráficamente en el diagrama de la Figura 6.7.

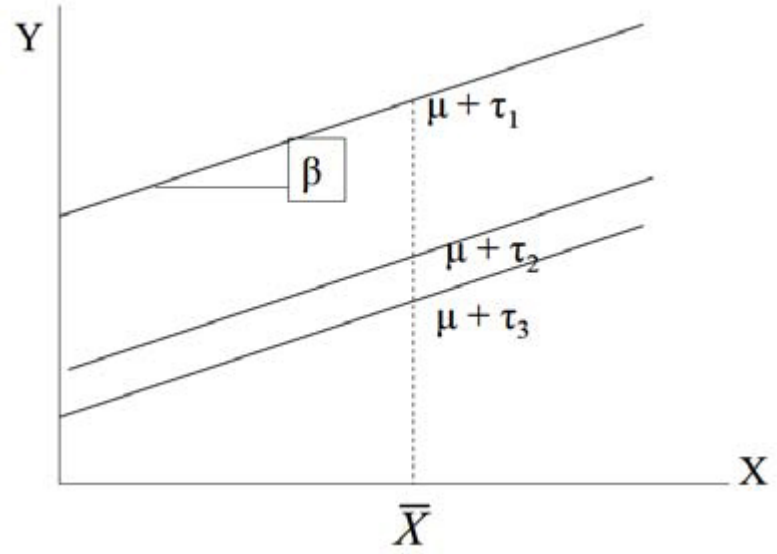


Figura 6.7 Comparación entre tratamientos (k=3) en presencia de una covariable

Los supuestos para el análisis de convarianza combinan aquellos del análisis de regresión y de varianza:

a) La covariable es fija, medida sin error y no interacciona con los niveles del factor, o factores, experimentales

b) La covariable tiene una relación lineal con la variable dependiente

c) Los errores del modelo son independientes, normales y con varianza constante.

Dados estos supuestos, el modelo correspondiente para el caso más simple de análisis de covarianza, un único factor bajo estudio en un diseño totalmente al azar, está dado como en (7)

$$Y_{ij} = \mu + \tau_j + \beta(X_{ij} - \overline{X}) + \varepsilon_{ijk} \qquad (7)$$

Donde β es el coeficiente de la covariable *Xij* que cuantifica la magnitud del cambio, positivo o negativo, que la covariable induce sobre la respuesta

La tabla ANCOVA no es fácil de construir manualmente (ver Steel y Torrie, 1980, para una discusión detallada), pero con el apoyo de software estadístico el análisis es fácil de realizar. Tanto en el caso de SPSS como de MINITAB, el análisis de covarianza está implementado a través del procedimiento GLM. La ventana de diálogo correspondiente en MINITAB se observa en la Figura 6.8, en la pantalla principal hay que elegir el botón Covariates para declarar cuál de las columnas del archivo de datos contiene los valores de la covariable. El modelo que se declara en la figura es el (7), esto es, un modelo lineal con únicamente el efecto de un factor (promoción) y una covariable (valoprom).

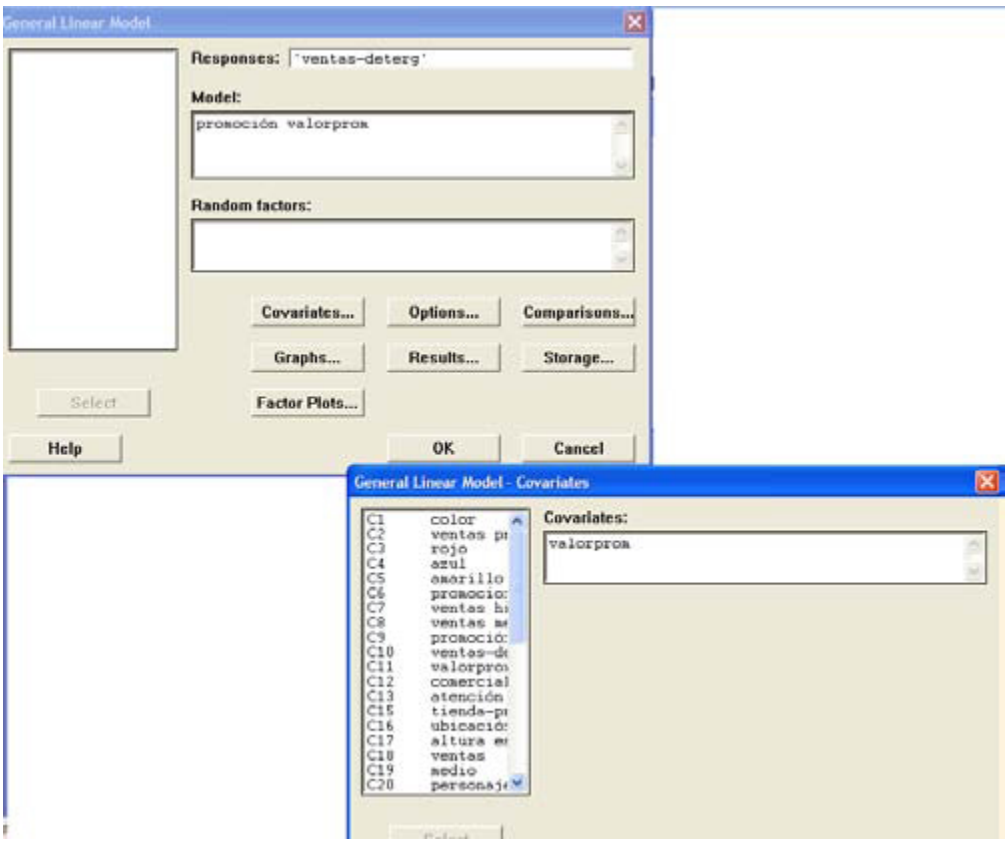


Figura 6.8 Análisis de covarianza empleando GLM de MINITAB

Ejemplo 11. Ante la creciente competencia por la captación de clientes las agencias de automóviles más antiguas, ubicadas en la zona metropolitana de la ciudad de Toluca, decidieron ofrecer promociones en la compra de autos. Tres paquetes promocionales, definidos al azar, fueron ofrecidos por períodos de un mes en doce agencias participantes. La respuesta a la promoción fue el total de ventas registradas por las agencias, en número de autos, durante el mes de vigencia de la promoción. Las agencias no son homogéneas entre sí puesto que distribuyen diferentes marcas de automóvil, están ubicadas en distintos lugares, varían en su tamaño y la variedad de modelos en exhibición, entre otros aspectos.

Para controlar esta variabilidad adicional e incrementar la precisión en las comparaciones entre los tres paquetes promocionales se incluyó como covariable las ventas totales que cada agencia tuvo durante el último año, ya que esta variable concomitante refleja las diferencias en tamaño y posicionamiento tanto de marca de automóvil como de agencia. Las promociones evaluadas fueron:

A = Medio año de seguro de cobertura completa

B = Placas y verificación doble cero incluidas en la compra del automóvil

C = Gratis el juego de accesorios a elección del comprador

El análisis de covarianza arrojó los siguientes resultados:

General Linear Model: Ventas mensuales versus promoción auto

Factor	Type	Levels	Values
promoción	auto	3	1,2,3

Analysis of Variance for Ventas mensuales, using Adjusted SS for Tests

Source	DF	Sen SS	Adi SS	Adi MS	F	P
Promoción	2	368.67	67.74	33.87	8.77	0.023
Ventas anuales	1	56.02	56.02	56.02	14.51	0.013
Error	5	19.31	19.31	3.86		
Total	8	444.00				

S = 1.96523 R-Sq = 95.65% R-Sq(adj) = 93.04%

Term	Coef SE	Coef	T	P
Constant	-7.059	5.916	-1.19	0.286
Ventas anuales	0.10557	0.02772	3.81	0.013

Tukey Simultaneous Tests

Response Variable Ventas mensuales

All Pairwise Comparisons among Levels of Promoción auto

Promoción auto = 1 subtracted from:

promoción Difference SE of Adjusted

auto	of MEANS	Difference	T-Value	P-Value
2	-5.518	1.767	-3.123	0.0576
3	-9.136	4.612	-1.981	0.2119

Promoción auto = 2 subtracted from:

promoción Difference SE of Adjusted

auto	of MEANS	Difference	T-Value	P-Value
3	-3.618	5.311	-0.6811	0.7841

La segunda parte del listado corresponde al análisis de regresión de las ventas mensuales con promoción sobre las ventas anuales de las agencias. El coeficiente de regresión es positivo y altamente significante (P=0.013), por cada aumento en las ventas anuales de un automóvil, se espera que las ventas mensuales se incrementen en 0.10557 automóviles. En cuanto a las promociones, los resultados de la prueba de Tukey permiten concluir que la promoción 1, seguro de auto gratis por seis meses, es la más atractiva, con 5.518 automóviles más vendidos en promedio por agencia cuando se hace esta promoción. Las promociones 2, placas y verificación, y 3,paquete de accesorios gratis, tuvieron resultados comparables e inferiores a los de la primera promoción.

Actividad de repaso

6. Análisis de varianza

1. Diseño experimental en que se controla el efecto de una variable externa agrupando las unidades experimentales con base en los niveles de esa variable.

a. Cuadro latino

b. Factorial

c. Bloques al azar

d. Lineal

2. Supuesto estadístico para el análisis de covarianza:

a. Relación lineal entre la respuesta y la covariable

b. Independencia entre la respuesta y la covariable

c. Respuesta medida sin error

d. Múltiples factores en el experimento

3. Tamaño del cuadro latino requerido cuando un factor tiene seis niveles:

a. 6

b. 36

c. 12

d. 18

4. Procedimiento empleado para acomodar las combinaciones de un factorial en bloques de menor tamaño que el total de combinaciones.

a. Fraccionamiento de factoriales

b. Bloques incompletos

c. Confusión de factoriales

d. Uso de covariables

5. En el análisis de covarianza representa el cambio esperado en la respuesta ante un cambio unitario en el valor de la covariable.

a. Efecto del factor

b. Error experimental

c. Coeficiente de la covariable

d. Media global

6. En un experimento con un factor a 5 niveles, en bloques totalmente al azar, si se realizan 4 réplicas, 4 bloques, los grados de libertad para el error son:

a. 20

b. 12

c. 16

d. 15

7. Diseño que resulta de utilizar bloques de menor tamaño que el número de niveles del factor pero con la condición de que se tenga el mismo número de réplicas por tratamiento.

a. Bloques incompletos balanceados

b. Bloques totalmente al azar

c. Bloques incompletos no-balanceados

d. Cuadro latino

8. Si en un análisis de covarianza el valor P para la prueba F asociada a la covariable es P = 0.33, se concluye que:

a. Hay diferencias significantes en los niveles del factor

b. La covariable y el factor son independientes

c. La covariable no afecta la respuesta

d. La covariable tiene un efecto significante sobre la respuesta

9. Si el cociente MSE (Diseño totalmente al azar) /MSE (Diseño de bloques) = 123% implica que:

a. El diseño de bloques es menos eficiente que el totalmente al azar

b. El diseño de bloques es más eficiente que el totalmente al azar

c. Los dos diseños son equivalentes

d. Se requieren menos réplicas en el diseño totalmente al azar que en el de bloques e.

10. En un diseño totalmente al azar el efecto de posibles variables externas se controla a través de:

a. Formación de bloques

b. Agrupamiento de unidades experimentales en términos de dos variables externas (columna y renglón)

c. Empleo de covariables

d. Asignación aleatoria de unidades experimentales a los tratamientos

RESPUESTAS: c, a, b, c, c, b, a, d, b, d.

6.7 Aplicaciones del ANOVA con apoyo computacional

Los diseños de bloques, cuadro latino y el uso de covariables no están restringidos al caso de los experimentos unifactoriales, pueden utilizarse cuando hay dos o más factores. Pero hay un punto de consideración, cuando el número de combinaciones para los factores es muy grande, resulta difícil y/o costoso acomodar todas estas combinaciones en bloques o cuadros latinos. Por ejemplo si se tiene un experimento con dos factores, cada uno con tres niveles, el total de posibles combinaciones es de nueve. Un cuadro latino 9x9 implica 81 experimentos, tal cantidad resulta prohibitiva en términos de costos y control sobre la toma de datos. En estos casos, la estadística ofrece opciones valiosas para reducir el número de pruebas. En el caso de los bloques, si bien anteriormente se mencionaron los diseños de bloques incompletos, éstos aplican a casos en los que hay un único factor. Cuando se tiene un experimento factorial, la estructura del experimento permite aplicar lo que se conoce como Confusión de Factoriales (Arroyo, 1994; Montgomery, 2009) para distribuir las combinaciones en bloques o cuadros latinos de menor tamaño de una manera eficiente. Así pues, el investigador de mercados cuenta con herramientas estadísticas para realizar experimentos complejos y analizar datos de estudios causales en los que se consideran múltiples variables; un ejemplo de aplicación en el cual los datos se procesaron con SPSS se reporta a continuación:

En el desarrollo de un nuevo producto de confitería, se evaluó el efecto que tres factores de diseño tienen sobre varios atributos sensoriales del producto (sabor en general, intensidad del sabor, apariencia y color) y en la calificación global que los consumidores asignaron al producto. Varios productos fueron elaborados al realizar combinaciones de los siguientes factores:

- la consistencia del producto, espuma de chocolate o chocolate duro,
- el grado de dulzor del producto, bajo y alto,

» el nivel de acidez en el sabor del producto, bajo y alto.

Los ocho posibles productos de chocolate fueron evaluados por 10 consumidores potenciales, adolescentes entre 14 y 18 años, todos los participantes evaluaron todos los productos para, de esta manera, corregir por las diferencias en las preferencias de un individuo a otro. Se requiere determinar cuáles factores, o combinaciones de ellos, determinan la calificación global que recibió el producto.

En primer lugar es necesario caracterizar el experimento, o estudio causal en caso de que se hayan usado datos de encuesta u observación, esto es, identificar factores, variables externas y de repuesta. En este caso se trata de un experimento factorial con tres factores a dos niveles cada uno (2x2x2 = 8 combinaciones de los tres factores) en un diseño de bloques totalmente al azar. Cada individuo es un bloque puesto que prueba todos los productos y se esperan diferencias relevantes de persona a persona. Con esta información se declara el modelo correspondiente que se requiere

para utilizar GLM en SPSS. La ventana de diálogo principal, que se abre luego de elegir la secuencia de comandos: Analyze > General Linear Model > Univariate (hay una única variable de respuesta), se muestra en la Figura 6.9.

El archivo de datos contiene múltiples columnas y es necesario seleccionar la columna que contiene los datos de la variable dependiente o respuesta así como las columnas en donde se han codificado los niveles de los factores, que pueden ser fijos como, en este caso, o aleatorios. Al igual que con GLM de MINITAB, las covariables se declaran por separado en el recuadro correspondiente.

Una vez registradas las variables hay que declarar el modelo, al presionar el botón Model se abre la ventana de diálogo que se muestra a la derecha en la Figura 6.9. En el caso de SPSS el default es el factorial completo (Full factorial) lo que en este ejemplo implica estimar los efectos principales de consistencia, dulzor, acidez y persona que evalúa, más todas sus interacciones de dos, tres y cuatro letras. Puesto que la variable "persona que evalúa" no es un factor sino que identifica a los bloques, se asume que no hay interacción de esta variable con los tres factores experimentales. Además de esto, la interacción de tres letras consistencia x dulzor x acidez no resulta relevante para estimar, en consecuencia, sólo las interacciones de dos letras son requeridas. Es decir que el modelo ANOVA que se propone es:

$$Y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \rho_l + \varepsilon_{ijkl}$$

Donde α i es el efecto de la consistencia (i = 1,2), βj el efecto del dulzor (j= 1,2), γk el efecto de la acidez (k= 1,2) y ρl representa a los varios consumidores (l = 1,2,…,20) que evaluaron el producto.

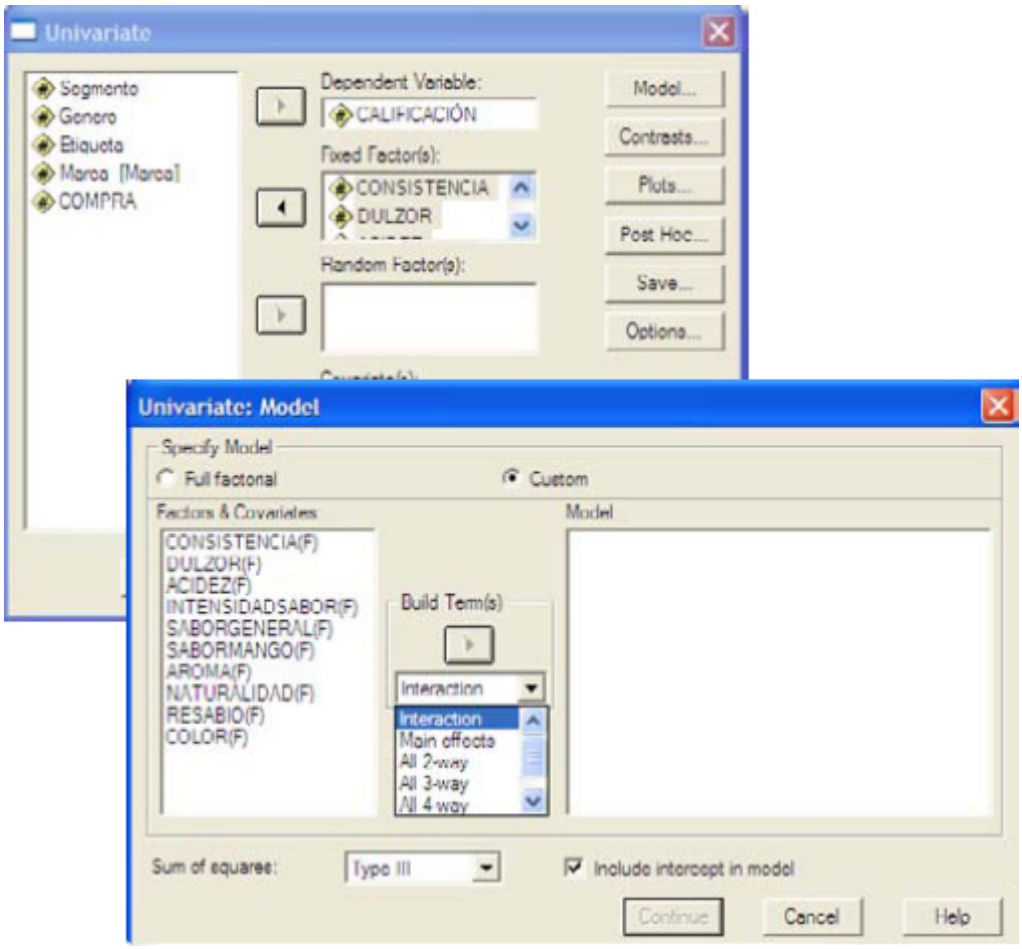


Figura 6.9 Análisis de varianza en SPSS para los datos de experimentación de un nuevo producto

Para declarar el modelo anterior en SPSS hay que seleccionar la opción Custom; primero hay que elegir Main effects y seleccionar aquellos factores y variables de ruido que incluye el modelo para después elegir la opción All 2-way y proceder a seleccionar las parejas de factores cuyas interacciones de dos letras se desea estimar. El ANOVA resultante se muestra en la Tabla 6.7.

Source	Type II Sum of Squares	DF	Mean Square	F	Sig.
Correct Model	916.971(a)	23	39.868	5.261	0.000
Intercept	1394.797	1	1394.797	184.047	0.000
CONSUMIDOR	71.018	9	7.891	1.041	0.418
CONSISTENCIA	62.498	1	62.498	8.247	0.006
DULZOR	51.386	1	51.386	6.780	0.011
ACIDEZ	51.779	1	51.779	6.832	0.011
CONSISTENCIA * DULZOR	72.909	1	72.909	9.621	0.003
CONSISTENCIA * ACIDEZ	87.466	1	87.466	11.541	0.001
DULZOR * ACIDEZ	61.632	1	61.632	8.132	0.006
Error	1085.024	64	7.5785		
Total	12201.995	80			
Corrected Total	943.712	79			

a R Squared = .507 (Adjusted R Squared = .471)

Tabla 6.9 ANOVA para la evaluación de productos de confitería

Del ANOVA anterior se concluye que todos los efectos son significantes y puesto que las interacciones son relevantes es necesario construir los gráficos correspondientes para explicarlas. En cuanto al efecto de los bloques = consumidores, el cociente F indica que no hay diferencias sustanciales entre los consumidores que evaluaron el producto por lo que en experimentos posteriores podría omitirse este efecto.

Los gráficos para las interacciones se dan en la Figura 6.10, para construirlos hay que elegir la opción Plots en la ventana principal e indicar cuáles interacciones desean representarse gráficamente. La opción Separate lines proporciona los gráficos usuales que se han presentado a lo largo de este capítulo, sólo es necesario indicar cuál factor se presentará en el eje horizontal y para cuál otro se hará la representación con líneas diferentes.

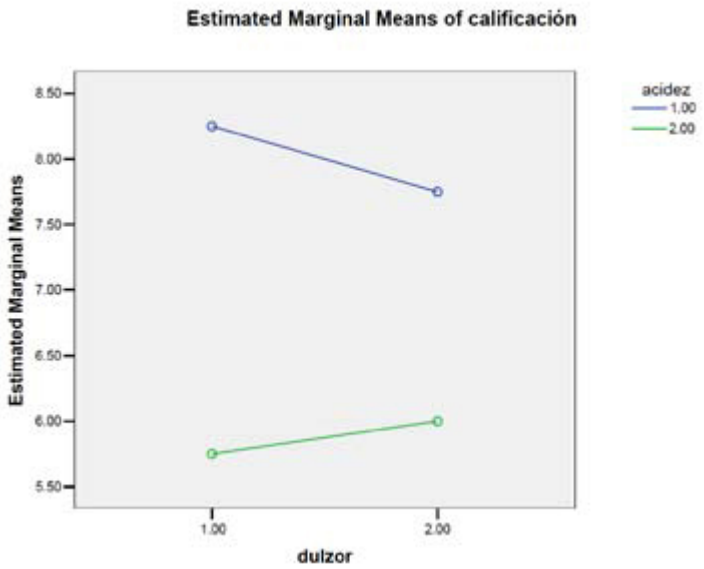


Figura 6.10 a. Interacción dulzor x acidez en la elaboración de un producto de chocolate

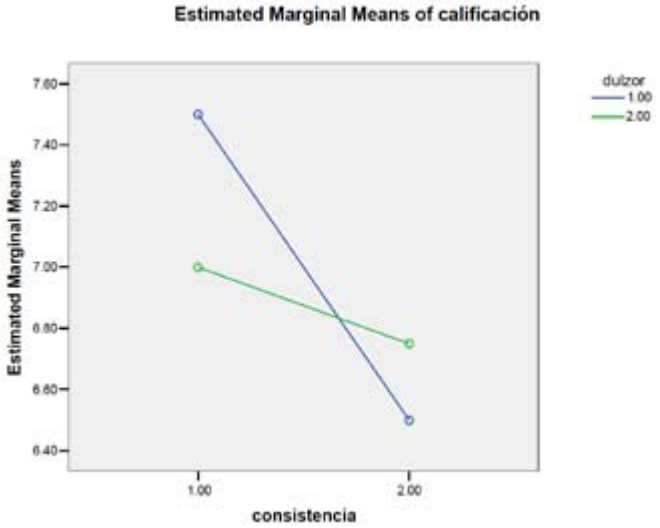


Figura 6.10 b. Interacción consistencia x dulzor en la elaboración de un producto de chocolate

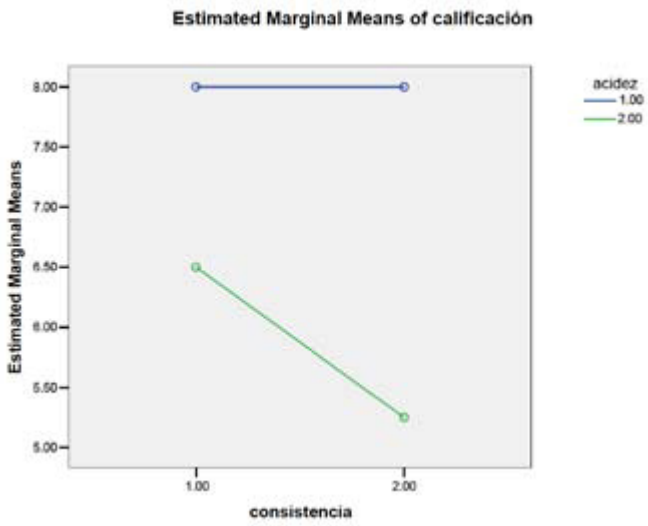


Figura 6.10 b. Interacción consistencia x acidez en la elaboración de un producto de chocolate

La discusión para los gráficos de la Figura 6.10 es como sigue:

Interacción dulzor x acidez. Cuando la acidez es baja, nivel 1, un incremento en el dulzor deteriora la calificación que recibe el producto, lo contrario ocurre cuando la acidez es alta, a mayor dulzor mejor evaluado es el producto.

Interacción dulzor x consistencia. El producto tipo chocolate duro recibe en general menor calificación que el tipo espuma, pero la pérdida en calificación es más severa cuando la dulzura del producto es baja, recta con mayor inclinación para dulzor 1 que para dulzor 2.

Interacción acidez x consistencia. Si la acidez es baja, ambos tipos de producto, espuma y chocolate duro, son igualmente evaluados, pero si la acidez es alta el producto de chocolate duro resulta más mal calificado que el de la espuma de chocolate.

Con base en el análisis anterior, se recomienda a la empresa lanzar al mercado el producto de chocolate tipo espuma, con un dulzor bajo y un grado de acidez en el sabor bajo.

Ejercicio integrador

6. Análisis de varianza

Instrucciones: Revisa el siguiente caso, realiza las operaciones indicadas y responde a lo siguiente.

Caso – Ejercicio integrador

Conoce la propuesta de solución por parte de los autores para este caso. Solución

a) Caracterizar el experimento y escribir el modelo ANOVA correspondiente

b) Analizar los datos y establecer qué efectos son significantes

c) ¿Qué efecto tiene el precio del producto sobre la calificación que recibe?

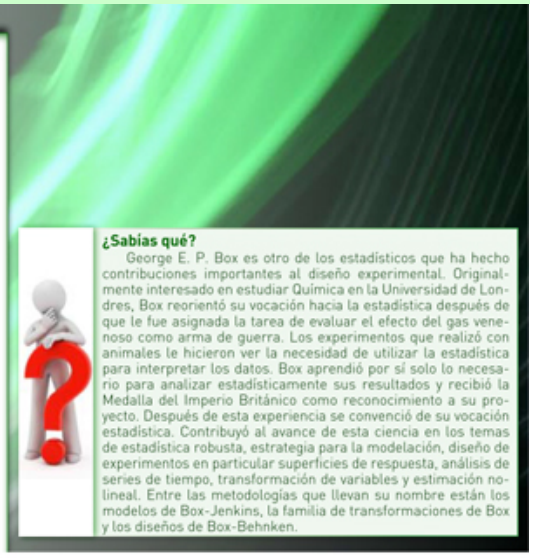
d) ¿Qué producto fue el mejor evaluado?

Conclusión capítulo 6

6. Análisis de varianza

En este capítulo se introdujeron los conceptos básicos de diseño y análisis de experimentos. Si bien el uso de experimentos se asocia a ciencias como la física, química y biología, esta metodología, en estudios causales, se ha extendido a otras disciplinas incluyendo las ciencias económico-administrativas y sociales. La realización de experimentos en laboratorio, donde es más fácil controlar por variables externas y acelerar el tiempo para la obtención de datos, como es mostrar anuncios publicitarios distintos que varían sistemáticamente en sus características o evaluar productos nuevos con distintos atributos son aplicaciones frecuentes del diseño experimental en el contexto de mercadotecnia.

Los estadísticos han propuesto una gran variedad de diseños experimentales y métodos encaminados a reducir el número de pruebas a realizar - lo que reduce los costos del estudio - e incrementar la eficiencia de las pruebas estadísticas que permiten determinar cuáles factores son determinantes para obtener valores deseables de una respuesta. Aparte de los diseños básicos descritos en este capítulo - totalmente al azar, bloques y cuadro latino - hay otros que han surgido para mejorar la problemática de experimentación en áreas específicas pero cuyo uso se ha extendido a otras áreas. En el caso de mercadotecnia, los principios de diseño experimental se han extendido para desarrollar un método de estadística multivariable enfocado al diseño de nuevos productos, el análisis conjunto, tema del último capítulo de este eBook.

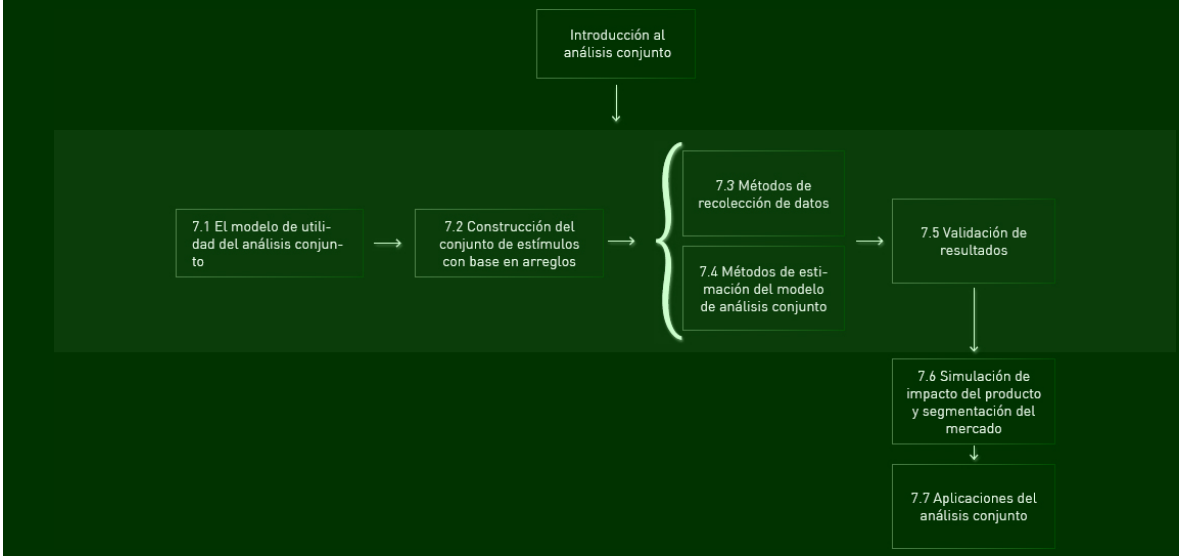


¿Sabías qué?

George E. P. Box es otro de los estadísticos que ha hecho contribuciones importantes al diseño experimental. Originalmente interesado en estudiar Química en la Universidad de Londres, Box reorientó su vocación hacia la estadística después de que le fue asignada la tarea de evaluar el efecto del gas venenoso como arma de guerra. Los experimentos que realizó con animales le hicieron ver la necesidad de utilizar la estadística para interpretar los datos. Box aprendió por sí solo lo necesario para analizar estadísticamente sus resultados y recibió la Medalla del Imperio Británico como reconocimiento a su proyecto. Después de esta experiencia se convenció de su vocación estadística. Contribuyó al avance de esta ciencia en los temas de estadística robusta, estrategia para la modelación, diseño de experimentos en particular superficies de respuesta, análisis de series de tiempo, transformación de variables y estimación no-lineal. Entre las metodologías que llevan su nombre están los modelos de Box-Jenkins, la familia de transformaciones de Box y los diseños de Box-Behnken.

Capítulo 7
Introducción al análisis en conjunto

Organizador temático



7. Introducción al Análisis Conjunto

Introducción

El **análisis conjunto** fue sugerido por Paul Green, experto en el desarrollo de nuevos productos, con el propósito de extender la información que proporciona el Escalamiento Multidimensional, MDS por sus siglas en inglés (MultiDimensional Scaling) en cuanto a las preferencias relativas para un grupo de productos. Empleando como referencia el trabajo desarrollado en el área de psicometría en lo referente a la descomposición de un juicio global, el Dr. Green sugirió descomponer la preferencia global expresada para un producto en **contribuciones parciales**,(*partworths* en inglés) asociadas a los atributos del producto bajo consideración. La idea central del análisis conjunto es identificar y determinar descriptores específicos de los atributos de un producto o servicio a partir de la evaluación o preferencia global que expresa el consumidor, con el propósito de desarrollar nuevos productos o bien hacer extensiones a la línea actual. Una vez que se han estimado las contribuciones parciales de los varios atributos del producto para cada individuo, éstas se utilizan para segmentar el mercado en términos de los beneficios que los consumidores buscan en un

producto. En una última etapa, los resultados del análisis se utilizan para hacer una **simulación** de la introducción del nuevo producto en el mercado con el objetivo de determinar cómo va ganando participación y, en un momento dado, anticipar un potencial canibalismo.

Las ventajas principales del análisis conjunto se pueden resumir como sigue:

1. Simula la situación de elección que hace el consumidor. Cuando el consumidor evalúa un producto no considera sus atributos por separado, o de uno en uno, sino que evalúa al producto en términos del conjunto de atributos que le ofrece. Esta técnica presenta el producto completo o permite al consumidor elegir aquellos atributos que desea ver en el producto

2. Es una técnica descomposicional. Busca comprender qué características, marca, empaque, por mencionar algunas, fueron determinantes en la selección de un producto o en su evaluación global.

3. Permite la identificación de los segmentos. Puesto que no todos los consumidores eligen un producto por la misma razón, entonces aquellos grupos de consumidores que consideraron los mismos criterios para la elección constituyen diferentes segmentos de beneficio. Esta técnica ayuda a decidir cuál o cuáles productos conviene introducir en el mercado.

4. Permite desarrollar modelos de **utilidad** para una combinación de atributos. Con base en estos modelos se puede responder a preguntas como: "¿qué pasa si… otros atributos se incluyen en el producto?", por ejemplo.

El artículo inicial de Green y Rao (1971) sobre cómo el análisis conjunto se podría usar en el estudio de la conducta del consumidor y el artículo clásico de Green y Wind (1975) sobre una aplicación específica captaron el interés de los investigadores de mercados en esta metodología (Hauser y Rao, 2002). A partir de entonces se han alcanzado notables avances para el análisis conjunto y se han realizado múltiples proyectos de aplicación. Entre los más interesantes está el diseño de la cadena de hoteles Courtyard by Marriott (Wind, Green, Shifflet, Shifflet y Scarborough, 1989) el cual se resume en el recuadro siguiente.

A principios de los 80, la cadena de hoteles Marriott decidió desarrollar un nuevo concepto de cadena para atender a aquellos consumidores que no estaban satisfechos con la oferta actual. Aparte de cubrir los objetivos económicos de la empresa, el nuevo concepto debía ser valioso para los

clientes, minimizar el canibalismo hacia los otros hoteles Marriott y ofrecer una ventaja competitiva distintiva. Para completar tan ambicioso proyecto, la gerencia de Marriott contrató como consultores externos a dos académicos, los doctores Wind y Green, para que le asistiera en el diseño del concepto de una nueva cadena de hoteles Courtyard by Marriott. En este proyecto se utilizaron múltiples métodos cuantitativos, el método central fue un análisis conjunto híbrido complementado con el uso de un modelo para evaluar la elasticidad en el precio del servicio de hospedaje. Otros métodos de estadística multivariante usados fueron el escalamiento multidimensional y el análisis de conglomerados.

El diseño para la nueva cadena de hoteles consideró siete factores de diseño o facetas principales:

1. Factores externos, edificio, tipo de alberca, ubicación y tamaño del hotel

2. Habitaciones, tamaño y decorado, distribución física, tipo de baño y entretenimiento

3. Servicio de alimentación, tipo y ubicación del restaurante, servicio a cuartos, máquinas para venta de bebidas y snacks

4. Instalaciones para relajarse, ubicación, atmósfera y clientela de éstas

5. Servicios, reservaciones, registro, centro de mensajes, renta de auto y servicios de mantenimiento

6. Instalaciones para actividades vacacionales, sauna, cuarto de juegos, zona para niños, canchas de tenis y racquetball

7. Seguridad, vigilancia, detectores de humo y cámaras.

En total se consideraron hasta 50 atributos de diseño, cada uno con 2-8 niveles. El estudio incluyó a 346 consumidores que viajan por negocios y a 255 personas que viajan por placer. Para facilitar la tarea de evaluación de conceptos de hotel se utilizó el análisis conjunto híbrido, que requirió que cada participante evaluara la utilidad de los niveles de cada atributo, uno a la vez, considerara qué tan importante le era cada atributo para después evaluar un número reducido de perfiles completos de 8 a 9 elegidos a partir de un diseño maestro complejo que aseguraba la estimación de los efectos de interés. Para evaluar el efecto del precio, una vez que el cliente potencial había elegido aquella combinación de atributos que le resultaba más atractiva en una faceta, debía calcular el precio de los atributos o servicios

elegidos. Si el precio total era superior al que estaba dispuesto a pagar por su estancia en el hotel, debía re-considerar el conjunto de atributos y decidir cuáles prefería que no tuviera el hotel para que el precio fuese aceptable. Los datos de cada faceta se procesaron con un análisis conjunto categórico para definir el conjunto de variables predictoras que explicara la utilidad de cada participante. Se calcularon entonces los parámetros del modelo conjunto híbrido para clusters de clientes potenciales. Finalmente se trabajó en una simulación de elecciones cuyo propósito fue evaluar el atractivo de mercado de varios paquetes de atributos y servicios y estimar su participación de mercado. La validez del estudio se evaluó con el método de uno-fuera pero el impacto en la rentabilidad y crecimiento de la cadena Marriott fue la mejor evidencia de que el nuevo concepto de Courtyard by Marriott aportaba valor a un segmento de consumidores.

¿Sabías qué?

El artículo seminal en el cual se inspiraron los investigadores de mercadotecnia para desarrollar el análisis conjunto fue escrito por un psicólogo-matemático y un estadístico. En este trabajo Luce y Tukey (1962) estaban interesados en investigar las condiciones bajo las cuales se podían derivar escalas de medición tanto para variables dependientes como independientes únicamente a partir de información ordinal del efecto conjunto de las variables independientes y de un regla hipotética de composición. Luce y Tukey denominaron medición conjunta, conjoint measurement, a su estrategia.

7.1 El modelo de utilidad del análisis conjunto

El análisis conjunto presupone que la evaluación global que un consumidor hace de un producto o servicio es resultado de la utilidad que percibe en él, donde esta utilidad es una función de las características o atributos que pueden tener diferentes productos. Cada uno es, por tanto, una alternativa de elección que se representa por un vector x que tiene *m* entradas; cada entrada corresponde a un nivel o valor particular de los atributos del producto/servicio, es decir

$$U(x) = f[U_1(x_1), U_2(x_2),..., U_m(x_m)]$$

donde la función de utilidad $U_j(x_j)$ es, a su vez, una función de las utilidades parciales asociadas a cada atributo y *f*[.] es la función a través de la cual se agregan estas utilidades. Los modelos de análisis conjunto son de descomposición puesto que se busca determinar las preferencias específicas para cada atributo a partir de:

1) la información sobre la preferencia global declarada para el objeto y

2) el conjunto de sus características o atributos.

La elección del modelo o función de descomposición es una de las varias decisiones críticas en análisis conjuntos. La naturaleza de la función de descomposición especifica cómo las contribuciones parciales de los atributos integran la calificación global que recibe un objeto. La forma de la función de descomposición también determina el número de estímulos que se requiere

evaluar para ajustar el modelo de utilidad, por ejemplo cuando la función de utilidad incluye términos de interacción, ver capítulo 6, entonces el número de estímulos requeridos aumenta. Los modelos de utilidad más utilizados son:

1) Modelo aditivo, en este caso se requiere estimar una única contribución parcial (w_j) para cada nivel del factor, esto es $U(x) = \sum w_j U_j(x_j)$

2) Modelo con interacciones de primer orden, en este caso se asume que hay combinaciones particulares (x_j, x_k) de los factores que contribuyen a la utilidad global percibida para el producto, en este caso $U(x) = \sum w_j U_j(x_j) + \sum w_{jk} U_j(x_j) U_k(x_k)$

3) Componentes o contribuciones parciales, *part-worhts*, este es el caso más general ya que busca obtener estimaciones específicas para la contribución de cada nivel del factor.

El diagrama de la Figura 7.1 describe cada una de las funciones de utilidad anteriores.

Un elemento central de la descomposición de la preferencia global es

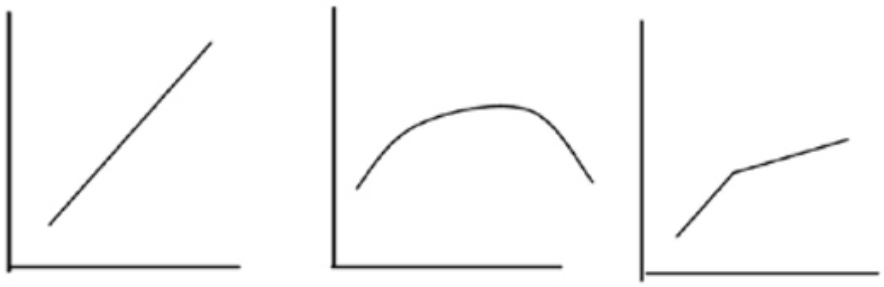


Figura 7.1 Los tres tipos de modelos de descomposición, utilidad, más usados

que los atributos del producto sean factibles, claros de comprender, separables y útiles para el equipo de diseño de nuevos productos. En la selección de atributos es importante la propiedad de independencia preferencial, esto es, dos atributos x_j y x_k, por ejemplo tipo de restaurante que tiene el hotel y disponibilidad de servicio a cuarto, son preferencialmente independientes de otros atributos si los intercambios o sustituciones entre ellos trade-offs no dependen de otros atributos. Cuando esta propiedad se sostiene, es válido utilizar modelos de descomposición aditivos.

7.2 Construcción del conjunto de estímulos con base en arreglos ortogonales

Es evidente que el diseño de un nuevo producto/servicio involucra una gran cantidad de atributos, considere el caso de la nueva bebida energética que se

citó como ejemplo en el capítulo anterior. El perfil de este producto requiere además de especificar el color de la bebida, 4 opciones, la elección de posibles sabores, por ejemplo limón, naranja y toronja, el tipo de envase, plástico o vidrio, su presentación, 300 ó 350 mililitros y contenido calórico, 100 ó 250 kilocalorías. Si todos estos atributos se tomaran en cuenta, el total de posibles nuevos productos es de 4x3x2x2x2 = 96. Aparte del alto costo de elaborar todas las posibles bebidas, debe considerarse la dificultad que representaría solicitar a un consumidor que evalúe todas las bebidas. Ya no se trata únicamente de reducir el número de corridas experimentales y los costos del proyecto, sino también de obtener datos confiables y evitar agotar a los consumidores con las tareas de evaluación.

En el capítulo anterior se comentó que la estadística ofrece opciones para fraccionar experimentos factoriales, estas ideas son ampliamente aplicadas en el caso del análisis conjunto para obtener diseños ortogonales. Estos diseños son muy eficientes para estimar las contribuciones parciales de los atributos pero tienen un costo ya que algunos de ellos únicamente permiten la estimación de los efectos principales. Esto es equivalente a asumir que no hay interacción entre los dos atributos; en términos de la función de utilidad que presupone el análisis conjunto esto implica que si se tienen dos atributos y cada uno puede asumir dos condiciones o niveles diferentes, entonces la contribución parcial de los atributos A y B cuando ambos están en su nivel alto es igual a la suma de las contribuciones parciales del nivel alto de A más la contribución parcial del nivel alto de B (Hauser, s.f.).

Los **diseños ortogonales** que se utilizan en análisis conjunto no corresponden únicamente a fracciones regulares de experimentos factoriales, algunos de estos diseños son mucho más complicados ya que la cantidad de factores es muy grande y los niveles de los atributos pueden no ser iguales cabe recordar que 2 ó 3 son los casos básicos que se manejan en el área estadística del diseño experimental.

Ejemplo 1. Suponga el caso de una nueva botana salada en cuyo diseño se consideran los siguientes atributos: A = base de la botana, papa o tortilla, B =

tipo de chile, chipotle o jalapeño y C = intensidad del picante, bajo o alto. El total de posibles combinaciones de estos tres factores son ocho, lo que equivale a ocho productos con distintos perfiles que se pueden representar (Montgomery, 2009) como:

(1) = todos los atributos en su nivel bajo, esto es un producto de papa, condimentado con chile chipotle y una baja intensidad de picante

a = un producto en el cual el atributo A está en su nivel alto, botana de papa, y el resto de los atributos en su nivel bajo,

b = atributo B en su nivel alto, chile jalapeño

c = atributo C en su nivel alto, alta intensidad del picante.

El resto de las combinaciones son: ab, ac, bc y abc. Si la letra figura en la combinación significa que se utilizó el nivel alto del atributo, de otra forma se trabajó con el nivel bajo del atributo.

En la evaluación de estímulos, el área de nuevos productos recomendó que un consumidor no probara más de cuatro productos ya que como el sabor de éstos es muy intenso, las evaluaciones para una mayor cantidad de botanas ya no resultarían confiables puesto que el consumidor dejaría de apreciar el sabor del producto. Las preguntas que surgen son: ¿Cuáles combinaciones debería evaluar un consumidor? de no elaborarse y evaluarse todos los productos ¿podrían estimarse las interacciones entre los atributos?

Para responder a estas preguntas debe considerarse la información necesaria para estimar los efectos de interés cuando se han probado los ocho productos respecto a cuando únicamente hay información para cuatro productos. Para calcular la contribución parcial de los atributos del producto, sus efectos principales, se requiere determinar el cambio en la preferencia de aquellos productos elaborados al nivel alto del atributo respecto de los elaborados al nivel bajo. Si los ocho productos fueran evaluados por un mismo consumidor, el cálculo requerido para estimar los efectos principales de acuerdo con la notación anterior es:

$$A = (a+ab+ac+abc)-[b+c+bc+(1)]$$

$$B = (b+ab+cb+abc)-[a+c+ac+(1)]$$

$$C = (c+ac+cb+abc)-[a+b+ab+(1)]$$

Para estimar las interacciones de primer orden, esto es la contribución de una combinación específica de dos factores, por ejemplo bc (ambos atributos están en su nivel alto), se determina qué tanto cambia la calificación global al variar los niveles del atributo B cuando el atributo C está en su nivel alto respecto a cuando está en su nivel bajo, es decir:

$$BxC = [abc+bc]-[ac+c]-[ab+b-\{a+(1)\}]$$

$$Bx\ C = [B(alto) - B\ (bajo);\ C\ alto] - [B(alto) - B\ (bajo)];\ C\ bajo]$$

En el supuesto de que únicamente las siguientes cuatro combinaciones van a ser preparadas: (1), ab, bc y ac, aún es posible estimar las contribuciones parciales de los tres atributos a partir de las evaluaciones del consumidor a estas cuatro botanas:

$$A = (ab+ac)-[bc+(1)] \quad B = (ab+bc)-[ac+(1)] \quad C = (ac+bc)-[ab+(1)]$$

Con solo cuatro combinaciones, la interacción entre los atributos tipo de chile e intensidad del picante se calcularía como: BxC = (bc-ac)–[ab–(1)], pero esta cantidad es exactamente igual–A = -[ab + ac] + bc + (1). De donde, si únicamente se tienen las evaluaciones de cuatro productos, no es posible utilizar **más que el modelo de descomposición aditivo**, ya que las utilidades asociadas con las interacciones o combinaciones particulares de los factores no son estimables.

Ejemplo 2. En el ejemplo 1 hay dos atributos más a considerar: D = presentación, 100 ó 150 gramos, y E = cantidad de grasa en el producto, regular o reducida en grasa. En este caso el total de **perfiles** o productos es de 32. Para determinar una cantidad de estímulos que resulte razonable se utilizan los siguientes diseños ortogonales:

2^{5-1} = 16 productos. Como cada consumidor únicamente puede evaluar con objetividad cuatro productos, resulta conveniente distribuir los 16 perfiles entre

cuatro distintos participantes. El total de las evaluaciones permitirá estimar un modelo de utilidad con algunas interacciones (AxD, AxE, BxD, BxE, CxD y CxE) o bien el modelo aditivo.

2^{5-2} = 8 productos. De nuevo habría que dividir el total de productos entre dos evaluadores, en este caso las 8 evaluaciones sólo permitirían estimar el modelo de utilidad aditivo, no es posible calcular la contribución de combinaciones específicas de los factores.

Los diseños requeridos para aplicar el análisis conjunto se generan utilizando software especializado; sin embargo los diseños anteriores, al ser fracciones regulares de la serie 2^k, pueden obtenerse empleando la aplicación Design of Experiments (DOE) en MINITAB. La secuencia de comandos es la siguiente: Stat > DOE > Factorial > Create Factorial Design. La ventana de diálogo correspondiente se muestra en la Figura 7.2.

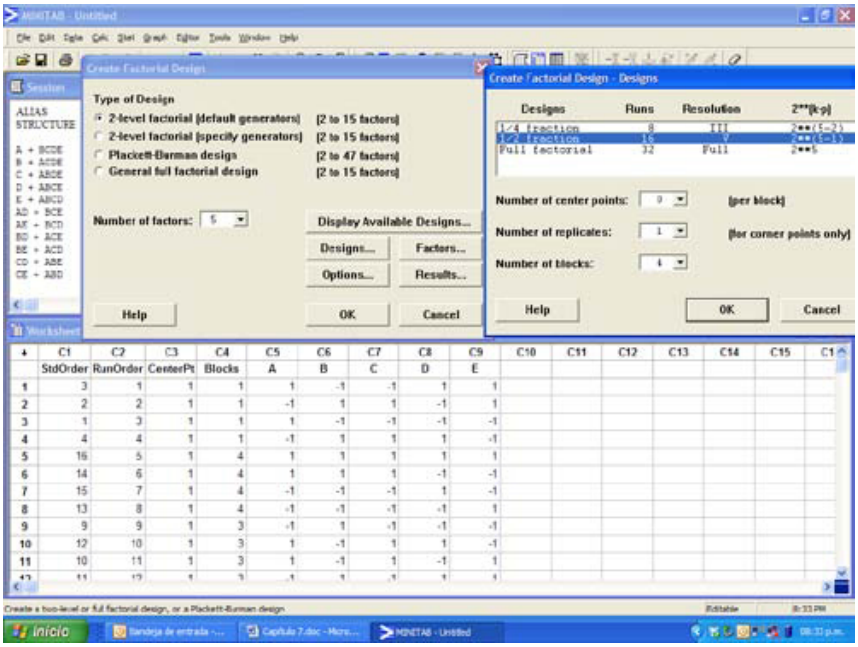


Figura 7.2 Construcción de diseños ortogonales para atributos con dos niveles

En la ventana de diálogo principal se requiere indicar el número de atributos o factores en estudio; de acuerdo a la descripción del ejemplo 2, hay 5 factores. Al elegir el botón Designs se despliega una ventana secundaria en donde se elige el número de combinaciones deseadas para el diseño. La opción Number of blocks permite dividir estas combinaciones en subgrupos de menor tamaño, 4 para el ejemplo 2, para así asignar un número conveniente de perfiles a los respondientes. Después de presionar OK, el diseño ortogonal especificado se desplegará en la hoja de trabajo de MINITAB según se muestra en la Figura 7.2. En la cuarta columna (C4) titulada Blocks se muestran las cuatro combinaciones que habría que asignar a cada evaluador. Las columnas C5-C9 indican las

combinaciones de atributos a evaluar. El código (-1) representa el nivel bajo del factor y el (1) el nivel alto. Por ejemplo la primera combinación (1, -1, -1, 1, 1) es una botana de tortilla, sazonada con chipotle, con baja intensidad de picante, en presentación de 150 grs. y reducida en grasa. En la pantalla del listado, visible en la esquina superior izquierda, se reporta la ESTRUCTURA DE ALIAS que indica al usuario cuáles efectos es posible estimar a partir de las evaluaciones para únicamente 8 ó 16 productos. En este caso los efectos que se pueden estimar son las contribuciones parciales para cada uno de los cinco atributos del producto (factores A-E) más las de cinco interacciones entre parejas de Figura 7.2 Construcción de diseños ortogonales para atributos con dos niveles atributos.

7.3 Métodos de recolección de datos

La elaboración de los perfiles de posibles productos, que constituyen los estímulos que se presentan a los consumidores potenciales, es otra de las decisiones críticas en análisis conjunto.

Los perfiles deben estar bien conformados y sus beneficios han de ser claros para el evaluador, los siguientes puntos son importantes de tomar en cuenta en la elaboración de estímulos:

1. Correlación inter-atributos. Se refiere a evitar proponer productos que involucren combinaciones de atributos que no resultan creíbles. Por ejemplo: alto rendimiento y alta potencia en un nuevo automóvil o bajo precio y alta calidad en una prenda de vestir.

2. La descripción del perfil. Debe incluir todos los atributos relevantes, es incorrecto usar únicamente atributos de valor o que contribuyan a incrementar la preferencia, también es importante considerar aquellos atributos que pueden afectar negativamente la preferencia para así tratar de contrarrestar su efecto al combinarlos con atributos valiosos.

3. La precisión de los atributos. Los niveles o condiciones que de los atributos tienen que ser precisos y distinguibles para el evaluador. Los niveles deben poder comunicarse verbal o gráficamente o bien crear realmente el objeto para que el consumidor potencial pueda distinguir claramente con qué atributos cuenta.

Una vez elaborados los estímulos, los esquemas principales que se utilizan para hacer su presentación y evaluación son:

1. Selección. A cada participante se le muestra un conjunto de alternativas de producto y se le pide que seleccione aquél que le interesaría comprar o que percibe como más valioso. Bajo este formato los datos de preferencia global resultan no-métricos, se compra o no se compra el producto, y requieren de un Análisis Conjunto Categórico o Análisis Conjunto Basado en Selecciones (CBC, Choice-Based Conjoint Analysis) que es más complejo que el análisis aplicado a datos obtenidos con otros esquemas de presentación de estímulos. Este tipo de método se recomienda para el caso de productos del tipo “me too”.

2. Perfil completo. Bajo este esquema, cada participante recibe una descripción, tarjeta, foto o imagen virtual que refiere el perfil completo para varios productos. El consumidor debe evaluar, en una escala de preferencia o intención de compra, cada perfil completo, tal y como lo haría en el caso de elegir nuevos productos que se introducen en el mercado. Los datos generados bajo este esquema son, con frecuencia, métricos ya que el participante expresa su preferencia sobre una escala de rating, por ejemplo un termómetro para intención de compra o una escala de preferencia con 5-7 categorías. Los datos obtenidos bajo este esquema de presentación son apropiados para un Análisis Conjunto Métrico. Pero los datos de preferencia pueden también ser ordinales cuando el participante ordena el total de posibles productos de acuerdo con su preferencia. Este tipo de método se recomienda para productos que son extensiones de una línea ya existente o para productos que son innovaciones verdaderas.

3. Evaluaciones de perfiles parciales. Si el supuesto de independencia preferencial se satisface, es razonable que los encuestados evalúen perfiles completos en los cuales algunos de los atributos están fijos. En una primera etapa se le hacen al participante una serie de preguntas auto-explicadas para definir aquellos atributos que percibe como más valiosos. En etapas sucesivas, el participante evalúa perfiles parciales que se determinan en función de su repuesta ante conjuntos de perfiles parciales previos. Esta selección adaptiva de perfiles se conoce como Análisis Conjunto Adaptivo (ACA). La estimación de las contribuciones parciales de los atributos se hace utilizando técnicas estadísticas más sofisticadas como el **método Bayesiano Jerárquico** (HB = Hierarchical Bayes) y los métodos poliédricos adaptivos cuyo propósito es filtrar un gran número de atributos para identificar y medir únicamente aquéllos que son los más valiosos para el consumidor (Hauser y Rao, 2002).

4. Preferencias declaradas. En esta opción los participantes no califican o establecen un orden de preferencia para las alternativas que evalúan sino que eligen un determinado perfil a partir de un conjunto de opciones que se les presentan. Si el total de posibles opciones o combinaciones que definen el producto, no es muy grande un mismo participante puede evaluarlas todas y elegir aquellas que podría comprar. También se presenta a cada participante un

sub-conjunto de opciones y se le pide que elija un perfil, el total de perfiles es entonces evaluado por un grupo de consumidores. Para datos de este tipo, las contribuciones parciales se estiman empleando modelos de selección aleatoria (logit y probit) como los mencionados en la última sección del Capítulo 5. Recordar que en estos modelos la utilidad global que un producto representa para el consumidor se define como una combinación de los atributos del producto más un término aleatorio.

5. Híbrido. En este caso, a cada participante se le solicita realizar primero una evaluación auto-descriptiva de la importancia que le representan los diferentes atributos de un producto/servicio, posteriormente el participante evaluará un grupo reducido de perfiles completos sobre una escala métrica u ordinal; los perfiles son construidos a partir de los atributos que fueron declarados como más valiosos. La función de utilidad correspondiente a este Análisis Conjunto Híbrido se obtiene al combinar los resultados de ambas tareas. Este esquema resulta apropiado cuando hay una gran cantidad de atributos bajo evaluación.

El tipo de escala en que se desea obtener la evaluación o calificación global, lo que a su vez se vincula con el tipo de análisis conjunto que se considera aplicar.

El nivel de utilidad particular que se proponga para cada atributo; el modelo de utilidad lineal es el más simple y el que requiere menos cantidad de datos para estimarse, mientras que en el modelo de utilidad por partes, como se requiere estimar la contribución parcial para cada valor de un atributo, requiere de mayor cantidad de información puesto que hay un mayor número de parámetros a estimar. En este caso es conveniente presentar múltiples perfiles completos a un mismo evaluador.

El análisis se desea realizar a nivel individual, por participante, o agregado sobre el grupo de participantes que hace la evaluación y que representa a un segmento de mercado.

Elaborados los estímulos hay que hacer una presentación de éstos a los participantes, para que tengan una idea clara de las cualidades del producto que se les presenta. Para ello se pueden usar tarjetas descriptivas, como en el caso Courtyard by Marriott, fotografías, prototipos del producto o imágenes virtuales y recursos multi-media. De acuerdo con Hauser y Rao (2002) “en los pasados treinta años, las representaciones de estímulos han estado limitadas únicamente por la imaginación de los investigadores.” Entre las aplicaciones más notables para presentación de estímulos está el proyecto de investigación multidisciplinaria MIT Sloan Virtual Costumer (2006).

El propósito de este proyecto es utilizar herramientas tecnológicas, Webbased, efectivas y simples para crear un ambiente virtual que facilite la coinnovación, esto es el diseño de nuevos productos a partir de información

relevante que proporcionan los consumidores a las empresas. El uso de estas herramientas reduce considerablemente el tiempo necesario, de días a semanas, para completar un proyecto ya que el consumidor proporciona la información en línea a través de sitios Web interactivos y amigables. Entre los métodos desarrollados para apoyar este proyecto están:

• Métodos de análisis conjunto rápidos y eficientes para datos métricos y datos basados en la selección de una alternativa

• Métodos de optimización para el diseño de preguntas y perfiles parciales requeridos en ACA

• Métodos nuevos para identificar atributos “obligados” o que debe tener un producto/servicio para que el consumidor lo considere aceptable, conocidos como Gards. Estos métodos buscan identificar estrategias lexicográficas para elaborar descripciones concretas sobre los atributos requeridos en un producto.

• Tarjetas para la elaboración de perfiles completos de producto de tal forma que se reduzca el número de evaluaciones para el encuestado.

Dahan y Hauser (2002) discuten a detalle los siguientes seis métodos que aprovechan la Web para obtener y procesar rápida y efectivamente la información que proporciona el consumidor cuando se le presentan nuevos productos:

1. Análisis conjunto basado en la Web (WCA, Web Conjoint Analysis). Los estímulos que se presentan al consumidor se elijen seleccionando, a partir de tarjetas, conjuntos de atributos relevantes del producto, por ejemplo calidad de la imagen y el sistema de apertura de la lente en el caso de una cámara fotográfica. En este método el objetivo principal no es la selección de estímulos sino la presentación para que el consumidor tenga una idea clara de los beneficios que ofrece el producto. Para el caso de la cámara fotográfica, el consumidor puede elegir alguno de los productos y acceder a un demo sobre cómo se utiliza y qué fotografías produce. Las comparaciones entre parejas de atributos son el método de generación de datos preferido cuando se utiliza este recurso ya que de esa manera se reduce el número de estímulos que el consumidor ve en la pantalla de la computadora.

2. Estimación poliédrica adaptiva rápida (FP, Fast Polyhedral). Este método deriva de la preocupación de los investigadores por reducir el número de estímulos o perfiles que el consumidor tiene que evaluar. A medida que el número de estímulos se incrementa, la información que proporciona el encuestado es menos confiable debido al cansancio y tedio que representa la tarea, para contrarrestar esto lo apropiado es un **análisis conjunto adaptivo** que amolda las preguntas que se hacen al consumidor dependiendo de sus respuestas previas. Cada consumidor se describe como un vector de

importancias relativas asignadas a cada uno de los M atributos de un producto. Si estas importancias se escalan entre 0 y 100, el conjunto factible de posibles importancias relativas es un hipercubo en M dimensiones. Gracias al uso de algoritmos de puntos interiores, una vez que se ha respondido a q preguntas, es posible estimar las importancias que el sujeto asigna a los atributos en el conjunto factible remanente.

3. Diseño del usuario (UD, User Design). Esta metodología se utiliza para determinar cuáles son los atributos más deseables en un producto, qué atributos interactúan entre sí y qué combinaciones corresponden al producto ideal. Específicamente se utilizan interfaces para que el consumidor seleccione interactivamente los atributos que desea en un producto, similar a cuando un comprador entra al sitio Web de Dell y configura la *laptop* que desea adquirir. La aplicación UD integra visualmente los atributos para mostrar un producto virtual que se modifica interactivamente al indicar qué atributos se desean comprar respecto a qué atributos tiene el producto. Las compensaciones entre atributos (tradeoffs) que realiza el consumidor se registran para determinar sus utilidades relativas lo que permite evaluar rápidamente un gran número de atributos que interactúan entre sí. Esta aplicación se ha utilizado en el diseño de cámaras fotográficas, bolsas para laptops, juguetes y productos de lavandería; los usuarios la han encontrado fácil de utilizar, rápida y divertida.

4. Prueba de concepto virtual (VCT, Virtual Concept Testing). A través de esta aplicación se presentan nuevos conceptos de producto a los consumidores quienes expresan sus preferencias comprando, a diferentes precios de acuerdo al perfil de atributos, aquellos que les resultan más atractivos. Esta aplicación sustituye a las pruebas de concepto convencionales, presenta al consumidor productos virtuales gracias a los avances en multi-media y el Internet de banda ancha.

5. Securities Trading of Concepts (STOC, Securities Trading Of Concepts). Aprovechando la capacidad de los servidores Web esta aplicación busca monitorear las interacciones entre consumidores para descubrir sus preferencias.

6. Bomba de información (IP, Information Pump). Este método complementa al anterior al analizar el lenguaje que usan los consumidores cuando evalúan nuevos conceptos.

Los métodos del proyecto Consumidor Virtual se utilizan en cualquier etapa de desarrollo del producto, desde la prueba de concepto hasta el refinamiento de atributos de acuerdo con el precio que el consumidor está dispuesto a pagar, y se espera que evolucionen a medida que avancen las tecnologías de información y comunicación (TIC). Entre las ventajas que ofrecen estos métodos sobre los métodos tradicionales de diseño del producto están:

» el uso de prototipos virtuales, lo que elimina la necesidad de elaborar prototipos físicos

» la rapidez con que se obtiene y procesa la información

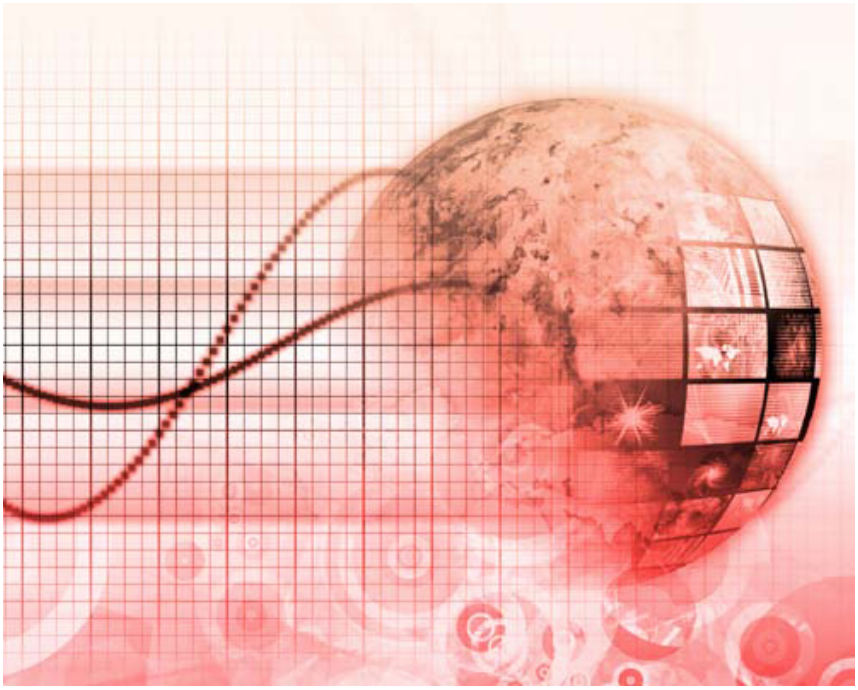
» la posibilidad de obtener información de consumidores con distintos perfiles y que residen en diferentes zonas geográficas.

Sin embargo, también tienen desventajas, entre ellas:

» requieren cierta programación para adecuarlos a cada aplicación,

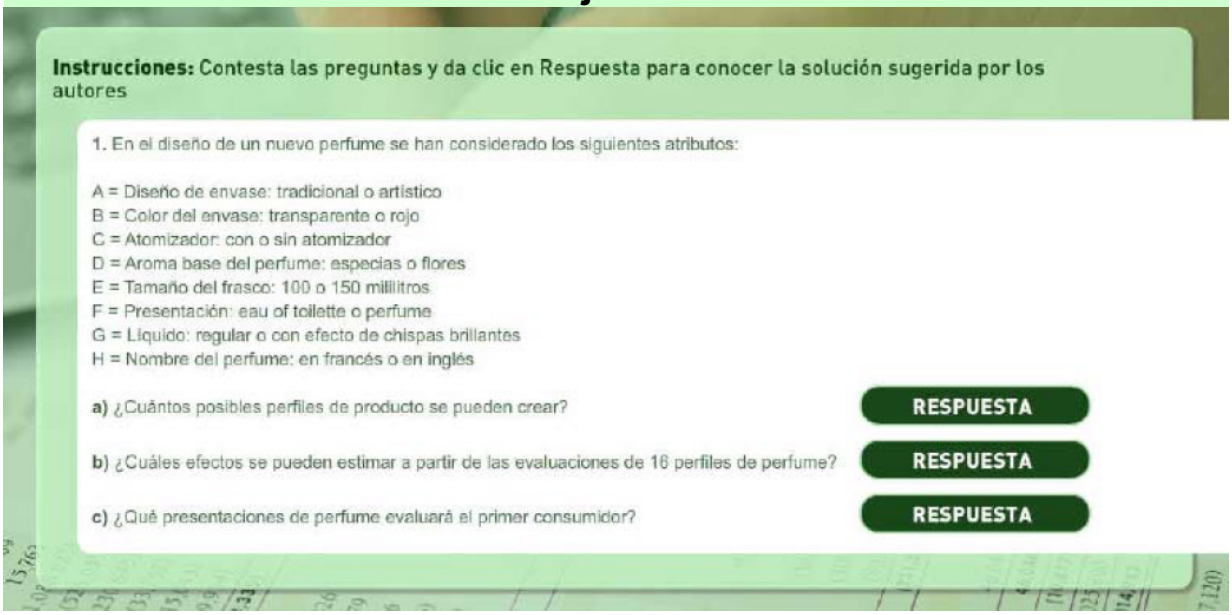
» algunos de los atributos son difíciles de presentar en forma virtual, por ejemplo el olor y sabor del producto

» los sujetos que los utilizan requieren de tener cierta habilidad computacional lo que limita su difusión en ciertos segmentos socioeconómicos, sobre todo en países en desarrollo como México.



Actividad de repaso

7. Introducción al Análisis Conjunto



Actividad de repaso (2)

7. Introducción al Análisis Conjunto

7.4 Métodos de estimación del modelo de análisis conjunto

Otro de los problemas críticos en análisis conjunto es la estimación de las contribuciones parciales de los atributos. La escala en la cual estén expresadas las preferencias o evaluaciones de los consumidores potenciales es determinante para la elección del método de estimación, los métodos disponibles incluyen:

1. MONANOVA (Monotonic ANOVA). Esta variante del Análisis de Varianza se utiliza cuando las evaluaciones de los consumidores corresponden a rangos de preferencia y el análisis conjunto se realiza a nivel individual.

2. Regresión. Cuando las evaluaciones están en una escala de intervalo, los atributos del producto se representan como variables dummy, especialmente cuando cada atributo puede tomar solo dos niveles (ver el Capítulo 4, Análisis de Regresión Lineal) por lo cual es posible utilizar mínimos cuadrados ordinarios para estimar modelos de descomposición lineales o con interacciones. La principal ventaja de esta forma de estimación es su simplicidad y que los coeficientes de regresión se escalan para medir la utilidad de un cambio unitario en los atributos cuantitativos. Este método también se usa cuando el análisis conjunto se realiza para grupos de participantes y se tienen datos de selección de alternativas. En esta situación, las elecciones del grupo de participantes se

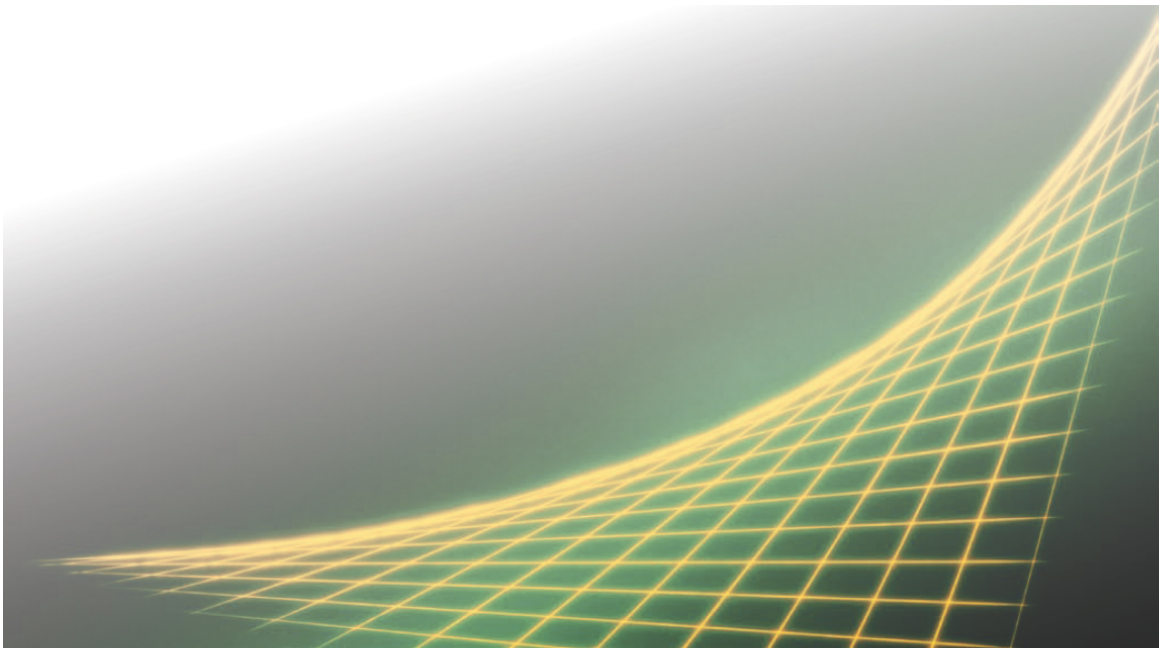
agregan para calcular porcentajes de individuos que eligieron un producto, estos porcentajes son la variable de respuesta para la regresión

3. ANOVA. Se usa como método de estimación para un modelo de contribuciones parciales siempre que las evaluaciones estén sobre una escala de intervalo o bien cuando se agregan los rangos o selecciones sobre grupos de participantes.

4. Modelos de selección aleatoria. De entre los modelos en esta categoría, el más utilizado es el llamado **modelo logit**; en este modelo las contribuciones parciales se estiman a partir de las preferencias o elecciones (1 = elige o compra, 0 = no elige o no-compra) declaradas por los participantes para un conjunto de alternativas. El problema principal con este método es que se requiere que los participantes califiquen muchas alternativas.

5. Estimación Bayesiana jerárquica. Este método permite la estimación de las contribuciones parciales aún a partir de pocos datos de evaluación. La idea básica de este método es que la incertidumbre en las contribuciones parciales o partworths de cada participante tiene una distribución de probabilidad conocida y que los parámetros de esta distribución, a su vez, se distribuyen a través de la población. El método utiliza la información y creencias de los respondientes para revisar los partworths que resultan ser bastante exactos aún cuando se tenga una cantidad limitada de datos (Hauser y Rao, 2002).

6. Métodos de estimación poliédrica. Estos métodos se han diseñado para aquellos casos en los que hay menos respuestas que *partworths* para estimar; las respuestas se manejan como restricciones para el espacio de soluciones lo que resulta en poliedro multidimensional que corresponde a la región de *partworths* factibles. Este es un método avanzado que se ha desarrollado en conjunto con el proyecto de Consumidor Virtual (Hauser y Rao, 2002).



7.5 Validación de resultados

Para evaluar la calidad de los modelos de análisis conjunto se utilizan las siguientes estrategias:

1. Predecir la utilidad o preferencia global para otros perfiles, combinación de atributos, que no fueron evaluados ni utilizados en la estimación de los partworths. Las preferencias globales que predice el modelo de análisis conjunto deben guardar cercana correspondencia con las preferencias expresadas por otros participantes a quienes se les presentan las nuevas alternativas. Tal correspondencia se evalúa a través del cálculo de coeficientes de correlación, ya sea la ρ Spearman si la preferencia está en escala ordinal o el coeficiente de correlación de Pearson cuando está en escala de intervalo.

2. Holdout. Divide los datos y sigue un proceso en tres etapas: i) una primera parte de los datos se usa para estimar los modelos de descomposición de análisis conjunto, ii) los modelos estimados se utilizan para pronosticar las preferencias de los perfiles de los productos que constituyen la otra parte de los datos y iii) se realiza la comparación entre las predicciones y las verdaderas evaluaciones de los datos de reserva. Esta estrategia, muy utilizada en análisis multivariante, ya se describió anteriormente.

7.6 Simulación de impacto del producto y segmentación del mercado

El análisis conjunto se complementa con:

a. la partición de los consumidores en grupos homogéneos en términos de sus preferencias para los productos y

b. la simulación de la participación en el mercado para el nuevo o los nuevos productos.

La segmentación de los consumidores se realiza de diferentes formas, Picón-Prado y Varela-Mallou (2000) describen y contrastan las siguientes opciones:

1. Aplicar un análisis conjunto a las preferencias de cada consumidor para después utilizar las contribuciones parciales individuales como variables de entrada en un análisis de agrupamiento de los sujetos (ver Capítulo 3). Los clusters resultantes corresponden a segmentos de consumidores que asignan el mismo valor o importancia (partworths similares) a los atributos de un producto. Esta estrategia de segmentación a posteriori es equivalente a una segmentación por beneficios en la cual los grupos formados reflejan directamente la importancia o valor que el segmento asigna a cada atributo del producto.

2. Realizar una segmentación a priori con base en variables sociodemográficas definidas por el investigador de mercado. Por ejemplo para las botanas saladas el agrupamiento inicial puede realizarse en función de la edad del participante. A continuación se aplica un análisis conjunto a los datos

por segmento y se proceden a analizar las diferencias que hay en las utilidades parciales entre los segmentos, de esta forma se tienen grupos a priori a los que se asocia una demanda por productos con diferentes perfiles.

3. Realizar una segmentación a posteriori con base en variables sociodemográficas, para posteriormente aplicar un análisis conjunto a los datos de cada uno de los clusters resultantes. La ventaja de esta opción respecto a la anterior, aparte del empleo de un mayor número de variables sociodemográficas, es que los grupos formados son homogéneos dentro de sí y heterogéneos entre sí, sin embargo esto no garantiza que difieran significativamente en cuanto a la importancia y utilidad que tienen los atributos de los productos.

El primer método descrito tiene la gran ventaja de generar segmentos definidos directamente por los beneficios que se busca en un producto; los otros dos métodos se han sugerido porque proporcionan estimaciones más confiables de las preferencias ya que el análisis conjunto se realiza sobre grupos y no a nivel individual. Además, cuando los datos de preferencias son de tipo categórico, compra o no-compra, éstos se agregan para calcular porcentajes lo que simplifica la realización del análisis conjunto.

Los modelos de simulación se utilizan para:

» responder preguntas importantes sobre las preferencias hacia varias alternativas bajo diferentes escenarios definidos, por ejemplo, en términos del número y tamaño de varios segmentos de consumidores o de la variedad de productos nuevos que se introducen

» predecir la participación en el mercado del nuevo o los nuevos productos

» analizar el efecto que la introducción del producto tiene sobre otros ya disponibles en el mercado

» estudiar cómo otros productos se eliminan del mercado a medida que el nuevo gana participación.

La forma más sencilla de realizar una simulación es asumir un modelo de primera elección, esto es presuponer que los consumidores elegirán el mejor producto, es decir, aquel que tiene la mayor utilidad total que se determina sumando las utilidades o contribuciones parciales para los niveles específicos de los atributos de cada alternativa de producto. Otras opciones más avanzadas emplean modelos de selección aleatoria como los modelos logit o probit, que permiten calcular la probabilidad de que se elija cada uno de los productos lo que resulta más realista para un mercado en el que varios productos compiten entre sí. Un modelo de selección aleatoria propone que la probabilidad de que un individuo exhiba la conducta de interés depende de la utilidad percibida en la

decisión, esto es, la probabilidad de que un consumidor cualquiera seleccione el producto o marca "i" (Y = 1) se calcula utilizando la regla de utilidad máxima BTL (Bradley-Terry-Luce):

E(Y) = 1 [Pr(Y = 1)] + 0 [Pr(Y = 0)]= Pr(Ui>0), si Y = 1 significa la compra del producto, entonces

$$\Pr(\text{compra}) = \Pr(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_M X_M + \epsilon i > 0) = \exp(U_i) / [1 + \exp(U_i)] =$$

Pr (compra) = Pr (comprar la alternativa i) / Pr (no comprar alternativa i)

donde U_i es la utilidad global para la alternativa i la cual es una función de los M atributos (X's) que definen el perfil del producto.

Si hay más de dos alternativas (j = 1,2,…i,i+1,…k), se tiene el modelo logit multinomial que propone que la probabilidad de que un individuo elija al producto i es una función de la utilidad percibida para esta alternativa con respecto a la suma de las utilidades para todos los productos:

$$\Pr(\text{elegir el producto i}) = \pi_i = \exp(U_i) / \sum_{j=1}^{k} \exp(U_j)$$

donde

$$U_i^a = \beta_0 + \sum_{j=1}^{M} \beta_{i,j} X_{i,j}$$

a nivel mercado

$$PM(i) = \sum_{k=1}^{K} W^{(k)} \pi_{k,i}$$

Los parámetros del modelo anterior (β's) se estiman con el método de máxima verosimilitud. Conocidos estos valores es posible simular la participación de mercado de cada producto competidor a partir de la utilidad que cada producto representa. La utilidad se calcula sustituyendo aquellos valores específicos de sus atributos (X's). Una vez calculadas las probabilidades de compra individuales, éstas se combinan para determinar la participación en el mercado para el grupo de compradores PM(i), ponderando cada una de las probabilidades individuales π_i, k por W(k) donde k =1,2,…,K se refiere al total de sujetos que reportaron sus preferencias. Los pesos se definen en términos de la frecuencia y monto de las compras de cada sujeto. A partir de esta idea básica se han definido reglas "logit" que permiten establecer los cambios en la estructura del mercado ante la introducción o eliminación de nuevos productos, el modelo SIMOPT (Simulation and Optimization model) considera este tipo de reglas para reportar participaciones de mercado y rentabilidad de los productos.

Al emplear este modelo, el usuario puede realizar los siguientes análisis (Green y Krieger, 1993):

1. Un análisis de sensibilidad para visualizar cuánto cambian las participaciones de mercado de los productos cuando se modifica la importancia de los atributos

2. Un análisis de canibalismo para el producto a través del cual el usuario puede determinar el perfil óptimo que maximiza la participación de mercado para una línea de productos.

3. Un análisis para la frontera Pareto-óptima. A través de este análisis se consideran otros criterios aparte de la participación en el mercado para definir cuál es el mejor producto, por ejemplo rentabilidad.

Ejemplo 3. Considere el caso del diseño de la bebida energética al que se hizo referencia en la sección de ANOVA en un sentido del capítulo anterior. La bebida puede ser de diferentes colores y cada uno de ellos tiene una utilidad distinta para el consumidor. Las contribuciones parciales correspondientes a cada color se reportan en la Tabla 7.1, a partir de estas cantidades se calcula la probabilidad de que un consumidor cualquiera elija cada producto. Cuando estas probabilidades se extienden al conjunto de posibles compradores del producto, lo que se tiene son las participaciones de mercado de cada alternativa.

Color	Utilidad	exp(utilidad)	Probabilidad de compra
Rosas	0.2	1.221	1.221/4.107 = .297
Verdes	-0.7	0.496	0.496/4.107 = .121
Azul	-0.3	0.741	0.180
Transparente	0.5	1.649	0.402
Total		*4.107*	*1.000*

Figura 7.2 Construcción de diseños ortogonales para atributos con dos niveles

Un modelo de primera elección indicará que la bebida energética transparente es la elección única del consumidor; mientras que el modelo logit, estimado a partir de los datos de intención de compra de un grupo de compradores potenciales, revela que si todas las variedades de color del producto se colocan en el mercado, algunas serán más favorecidas que otras (bebida transparente > rosa > azul > verde).

El uso de simulación en análisis conjunto es relevante para la toma de decisiones estratégicas (Orme, 2010), entre las que están:

1. Explorar posibles patrones de sustitución, como canibalismo y efectos de elasticidad cruzada, para marcas y productos con características afines al que se introduce al mercado. Esto requiere de predecir si la mejora de un producto incrementará su preferencia a costa de la reducción en la participación de mercado de marcas propias o de la competencia.

2. Identificar aquellos atributos que conviene mejorar en un producto para incrementar no únicamente su participación en el mercado sino su rentabilidad. Esto significa considerar el costo de mejorar o introducir determinado nivel de un atributo y analizar si el costo se compensa con el aumento en participación de mercado.

3. Analizar el efecto de cambios de precio del producto. Para lograrlo el precio de los otros productos se mantiene constante y se cambia únicamente el precio del producto de interés, simulándose su participación relativa para varios cambios en el precio.

4. Evaluar portafolios de productos destinados a satisfacer a varios segmentos de consumidores para investigar cuáles productos satisfacen a cuáles segmentos y establecer si el portafolio completo genera utilidades atractivas dependiendo del tamaño de los varios segmentos y el precio de los productos.

Actividad de repaso (3)

7. Introducción al Análisis Conjunto

Revisa el siguiente caso, analiza la información y responde a las preguntas.

Se diseñó una nueva galleta salada baja en sodio, con 3 cantidades diferentes, y alta en fibra, con 3 cantidades diferentes. Se aplicó un análisis conjunto a las preferencias de dos grupos de consumidores divididos por su rango de edad: 18-25 años y 26-40 años. Las siguientes funciones de utilidad lineal fueron estimadas para cada grupo:

18-25 años: 100.45 – 1.7 [cantidad de sodio] + 0.2 [cantidad de fibra]

26-40 años: 90.18 – 1.8 [cantidad de sodio] + 0.5 [cantidad de fibra]

a) De acuerdo con los partworths ¿cuál de los dos atributos es más importante para el segmento de 18-25 años?

b) ¿Difieren los dos segmentos en cuanto a la importancia que asignan a los dos atributos que definen el perfil del producto?

c) Determinar la participación que alcanzarían los siguientes dos productos en el segmento de 26-40 años:

Producto 1: 50 mg. de sodio y 3 grs. de fibra

Producto 2: 75 mg. de sodio y 5 grs. de fibra

Asuma que Σ exp(utilidades) para los nueves productos evaluados (3 cantidades de sodio x 3 cantidades de fibra) es de 19.25 x 10-17.

RESPUESTAS

a) La cantidad de sodio puesto que su partworth |1.7| es mayor que el del atributo cantidad de fibra = 0.2.

b) Sí, la mayor diferencia es para el atributo [cantidad de fibra] el cual es más apreciado por el segmento de mayor edad (0.5 > 0.2)

c) Primero hay que calcular la utilidad de cada producto usando las funciones estimadas, esto es:

U(producto 1) = 100.45 – 1.8 (50) + 0.5 (3) = 8.11689E-17

U(producto 2) = 100.45 – 1.8 (75) + 0.5 (5) = 6.31586E-36

La participación de mercado (producto 1) = 42.17% y participación de mercado (producto 2) = 0%.

7.7 Aplicaciones del análisis conjunto

El análisis conjunto es una metodología muy utilizada en investigación de mercados y para la cual se han reportado numerosas aplicaciones. En esta sección se discuten únicamente algunos proyectos de análisis conjunto realizados en distintos contextos y empleando diferentes recursos para la presentación de estímulos. A través de estas aplicaciones se podrá apreciar la relevancia de esta técnica para el área de mercadotecnia, en particular en la actividad Desarrollo de Productos.

1. Encuestas generadas por computadora para el análisis conjunto de las preferencias de ciudadanos en relación a proyectos de transporte (Hunt, Abraham y Patterson, 1993).

Objetivo: Demostrar la aplicabilidad del análisis conjunto para medir las actitudes de los residentes de Calgary, Canadá en relación a proyectos de transporte urbano. El software INVIEW fue específicamente desarrollado para generar las alternativas de proyectos, recolectar y analizar la información de las preferencias de los entrevistados.

Metodología: Los respondientes proporcionaron un ranking para un conjunto de alternativas hipotéticas de perfiles completos de proyectos de transporte urbano. Cada proyecto es una combinación aleatoria de once atributos distintos agrupados en las siguientes cuatro categorías de primer nivel:

» Movilidad. Medido como tiempo que toma hacer el viaje para dos modos, auto y transporte público. Los tiempos en auto fueron de 5-60 minutos con incrementos de cinco minutos y para el transporte público entre 25-90 minutos también con incrementos de cinco minutos.

» Tipo de construcción. Medido como clase de casas-habitación que se construirían una vez abierta la ruta, los niveles de este atributo fueron casas-habitación solas y edificios de departamentos.

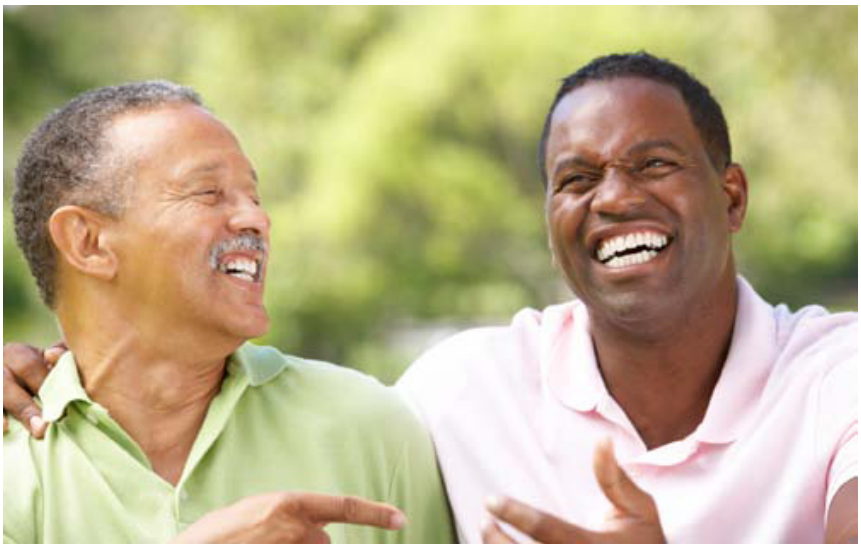
» Impacto sobre el ambiente. Este atributo se definió en términos de: la frecuencia de una mala calidad del aire según el Departamento del Ambiente de Alberta; la proximidad, cerca o lejos, de la ruta de transporte al valle de un río de acuerdo con las reglas del Departamento de Recreación y Parques de Calgary y la pérdida de espacios abiertos. Para que el respondiente tuviera una mejor idea de los niveles de este atributo se mostraron fotografías de varias rutas.

» Costos e impuestos del viaje al trabajo ya sea por auto o transporte público. Los costos de uso del auto se variaron entre 1-5 dólares canadienses con incrementos de 0.25 centavos. Mientras el costo del

transporte público se definió con base a las tarifas vigentes y varía entre 0.5-2.5 dólares canadienses. Los impuestos municipales para las residencias en la ruta se variaron entre 600-1800 dólares canadienses con incrementos de 300 dólares.

Un total de 4,450 jefes de familia fueron elegidos al azar para participar en la encuesta, 961 aportaron la información requerida. Las contribuciones parciales para cada uno de los once atributos que conforman los varios proyectos alternos se estimaron empleando un modelo logit en el cual la utilidad global de una alternativa fue una función lineal de los atributos.

Resultados: El programa de cómputo que se diseñó permitió generar entrevistas individualizadas y facilitó la obtención de datos sobre un amplio rango de tradeoffs para once diferentes atributos de los proyectos de transporte urbano. Por ejemplo la apertura de una avenida que afecta el medio ambiente de un área tuvo el mismo impacto en la utilidad que un incremento en el tiempo de viaje de 8.6 minutos; mientras que la compensación entre tiempo de viaje y costo del viaje indicó que el valor monetario de un viaje al trabajo fue de 11.06 dólares por hora.



2. La aplicación del co-diseño en las innovaciones de nuevos brassieres (Liao y Lee, 2010).

Objetivo: Analizar cómo los fabricantes de sostenes o brassieres desarrollan nuevos diseños del producto a través de co-diseñarlos con los consumidores.

Metodología: El diseño de nuevos productos utiliza los recursos del análisis conjunto para representar los perfiles multidimensionales de los potenciales productos. Esta técnica estadística facilita el co-diseño al adaptar las características del producto a las necesidades del cliente lo que resulta en un diseño más efectivo de brassieres.

Los atributos y niveles correspondientes que se consideraron en el diseño del producto fueron los siguientes:

- Material (Algodón, Mezcla de algodón, seda, Nylon, Spandex)
- Corte y cosido de copas: costura de dos piezas, costura de tres piezas, sin costuras
- Tirantes removibles: presente, ausente
- Color: negro, lavanda, rosa
- Estilo de los tirantes: dos tirantes, halter, strapless
- Relleno: presente, ausente
- Encaje: detalles de encaje en la V del brassiere, sobrepuestos a la copa, ausentes
- Diseño del escote: escotado, forma de corazón

Resultados: Los resultados aportan información relevante para la toma de decisiones para los gerentes de producto; a través del co-diseño se logran productos que resultan más atractivos a los consumidores disminuyéndose así el riesgo de fracaso en el diseño del brassiere. La combinación de atributos o perfil de producto preferido por los encuestados es un brassiere de algodón o de un material con una mezcla de algodón, sin costuras, color lavanda, un estilo de dos tirantes, con detalles de encaje y escotado. La participación en el diseño de los productos de la empresa es también una forma de promover la lealtad y satisfacción del consumidor hacia el producto.

3. Aplicación del análisis conjunto en la formación continua de un servicio de farmacia (Arias-Rico, 2010).

Objetivo: Definir las características más adecuadas para las sesiones clínicas que se imparten en el Servicio de Farmacia de un hospital con el propósito de que facultativos y residentes obtengan el mayor provecho formativo de ellas.

Metodología: El número de atributos considerados fue de cinco para hacer más claras las descripciones y facilitar la evaluación de las sesiones de formación clínica. Los atributos y niveles especificados fueron:

» Duración de las sesiones: menos de 20 min., entre 20-30 min. y más de 30 min.

» Disponibilidad de referencias bibliográficas: sí o no

» Archivo con contenidos de la sesión disponible en dispositivo de almacenamiento: sí o no

» Estructura de los contenidos docentes: solo contenido central, introducción + contenido central, introducción + contenido central + casos clínicos

» Contenidos multimedia: escasos, medios o abundantes.

Dada la gran cantidad de alternativas (144) se generó un diseño ortogonal o fracción de factorial que incluyó únicamente 16 combinaciones. Estos 16 distintos formatos de sesión de formación en servicios de farmacia fueron evaluados por 14 clientes potenciales: siete facultativos especialistas y siete residentes del Servicio de Farmacia del Hospital Universitario Reina Sofía en España quienes expresaron su preferencia para cada perfil completo de sesión sobre una escala de 1-16 en donde 1 = perfil más deseado a 16 = perfil menos deseado.

Resultados: Se utilizaron mínimos cuadrados para estimar las utilidades parciales de los diferentes atributos; los atributos más valorados por ambos segmentos de clientes fueron: la duración de las sesiones, la estructura de los contenidos y la disponibilidad de referencias bibliográficas. A partir de esta

información, la sesión con máxima utilidad y en consecuencia la de mayor preferencia se definió como una de duración menor a 20 min., con disponibilidad de referencias bibliográficas y archivo disponible en dispositivo de almacenamiento, contenidos amplios (introducción+contenido central+ casos clínicos) y contenidos multimedia en abundancia.

4. Uso del análisis conjunto para diseñar un nuevo sitio Web (Feldman, s.f.)

Objetivo: Diseñar un sitio Web de servicios financieros considerando principalmente atributos estéticos pero también un atributo determinante en la percepción sobre la efectividad del sitio, definido por el tiempo de carga.

Metodología: Se eligieron los siguientes cinco atributos, todos a dos niveles, para diseñar perfiles completos para un sitio Web:

- Color tema: A y B
- Gráfico primario: constante o rotatorio, total de tres imágenes a la vez y cambio de imagen cada 20 seg.
- Tamaño del gráfico primario: grande o pequeño
- Orientación del menú de opciones: horizontal (arriba) o vertical (a la izquierda)
- Tipo de caracteres: Serif o sans-serif
- Tiempo promedio para que se cargue el sitio: menos de 10 ó de 15-30 segundos.

Únicamente 16 perfiles completos para el sitio fueron descritos en tarjetas descriptivas que se usaron como estímulos; el diseño ortogonal se definió en términos de los efectos que era de interés estimar: la importancia de los cinco atributos más dos interacciones (tiempo de descarga x gráfico primario y orientación del menú x tamaño del gráfico primario). Los estímulos fueron

evaluados por diez usuarios potenciales durante una sesión de grupo de enfoque.

Resultados: Los rankings de preferencia fueron los datos de entrada para un análisis conjunto para el grupo de usuarios a partir del cual se determinó lo siguiente:

» Sorprendentemente, el tiempo de descarga no tuvo una contribución significante a la utilidad global percibida para el sitio Web lo que representa una oportunidad para reducir el costo del sitio.

» El diseño con un gráfico primario de gran tamaño y orientación horizontal del menú de opciones fue el más preferido (mayor utilidad)

» Siempre que se utilice el formato horizontal del menú, el tamaño del gráfico primario (interacción formato del menú x tamaño del gráfico) no decrece la utilidad global para el sitio Web.

Definidos los atributos con mayor contribución parcial el grupo de diseño del sitio Web siguió un proceso iterativo, aquellos atributos con baja contribución fueron fijados en sus valores más convenientes en términos de costo de diseño/operación del sitio y nuevos atributos, relacionados con la utilidad percibida del sitio, fueron considerados.

5. Una investigación empírica de las preferencias del consumidor en cuanto a servicios móviles (Tripathi y Siddiqui, 2010).

Objetivo: Estimar aquellos factores que inciden en las preferencias de los consumidores hacia los servicios de telefonía móvil para asegurar la retención y lealtad de los usuarios y fortalecer la posición competitiva de la empresa.

Metodología: A través de un estudio exploratorio se identificaron veintiséis determinantes relacionados con el proceso de elección de servicios de telefonía móvil. Mediante un análisis de factores, se definieron seis atributos a partir de

los 26 determinantes; los niveles de los atributos se especificaron a partir de entrevistas con expertos. Los atributos y niveles fueron los siguientes:

- Conectividad de la red: baja caída de llamadas, amplia cobertura, baja congestión
- Servicio al cliente: resolución de consultas, información adaptada al cliente, gestión de quejas
- Tarifa del servicio móvil:Tarifa de llamadas, variedad en la tarifa de los planes, denominación de los cupones de recarga
- Variedad del plan: post pago, tiempo del plan, prepago
- Valor agregado en los servicios: tonos de llamada/llamada, servicios como bromas o astrología, actualización diaria de noticias, deportes, entre otros.
- Tecnología utilizada en la red: CDMA, GSM, Ambas

Para diseñar los estímulos se generó un arreglo ortogonal que incluyó un total de 22 perfiles completos.

Resultados: El ambiente de negocios de los servicios telefónicos móviles es altamente competitivo, lo que hace indispensable interpretar los deseos de los clientes. Como no es posible que la empresa satisfaga todos los requerimientos del consumidor, el análisis conjunto resultó ser un recurso analítico valioso para identificar aquellos criterios que predicen la elección del consumidor ante servicios de telefonía competidores.

Los clientes otorgan gran importancia a la conectividad de la red, el siguiente atributo en importancia fue el servicio al cliente seguido de la tarifa de los servicios móviles. Los atributos de menor importancia fueron la variedad de planes, el valor agregado en los servicios y finalmente la tecnología utilizada en la red.

¿Sabías qué?

El Dr. Paul E. Green, autor del análisis conjunto y profesor emérito de la Wharton School de la Universidad de Pensilvania, ha publicado numerosos artículos académicos sobre avances y aplicaciones del análisis conjunto. Tan solo en las bases de datos de la biblioteca digital del sistema Tec de Monterrey están registrados 56 artículos de su autoría sobre el tema. También ha publicado 16 libros relacionados con el análisis multivariante y sus aplicaciones en mercadotecnia.

Se tiene acceso a algunas de sus publicaciones en la dirección http://marketing.wharton.upenn.edu/people/faculty/green/green_archive_list.cfm.

Ejercicio integrador

7. Introducción al Análisis Conjunto

Instrucciones: Revisa el siguiente caso. Contesta las preguntas y conoce la solución propuesta por los autores.

Considere el proyecto de diseño para una nueva botana salada, los atributos para el nuevo producto son:

A = base de la botana (papa o tortilla)

B = tipo de chile (chipotle o jalapeño)

C = intensidad del picante (bajo o alto)

D = presentación (100 ó 150 gramos)

E = cantidad de grasa (regular, reducción del 20%)

De los 32 posibles perfiles, sólo se elaboraron prototipos para 16 productos (2^{5-1}). Únicamente cuatro productos fueron evaluados por cuatro consumidores regulares de botanas (consumen un paquete al menos dos veces por semana). Los cuatro perfiles a evaluar por participante fueron seleccionados dividiendo la fracción factorial en bloques de tamaño 4. Cada participante expresó su intención de comprar cada producto en una escala termómetro que iba de 0-100, las evaluaciones asignadas a los productos se muestran en la siguiente tabla.

La primera columna indica cuál de los participantes realizó la evolución del producto, las siguientes cinco columnas indican cuál condición o nivel del atributo se utilizó. Si la columna tiene un “1” significa que el atributo se fijó en su nivel “alto” mientras que un “0” implica que el atributo estuvo en el nivel “bajo”. Así la primera combinación corresponde a una botana de papa (A alto), con base de picante de chipotle (B bajo), alta intensidad de picante (C alto), presentación de 100 g. (D bajo) y con una reducción del 20% de grasa (E alto). Este esquema de codificación es equivalente a representar cada atributo con una variable “dummy” para la cual el código 1 = nivel alto del atributo y 0 = nivel bajo del atributo.

Participante	A	B	C	D	E	Pr(compra)
1	1	0	1	0	1	69
1	0	1	0	0	0	61
1	0	1	0	1	1	82
1	1	0	1	1	0	71
2	0	0	1	0	0	55
2	0	0	1	1	1	90
2	1	1	0	1	0	72
2	1	1	0	0	1	68
3	0	0	0	0	1	73
3	0	0	0	1	0	66
3	1	1	1	1	1	85
3	1	1	1	0	0	57
4	0	1	1	0	1	68
4	0	1	1	1	0	73
4	1	0	0	1	1	80
4	1	0	0	0	0	69

Ejercicio integrador

7. Introducción al Análisis Conjunto

a) Estimar los partworths o contribuciones parciales de los atributos empleando regresión. Asumir un modelo aditivo.

b) Identificar aquellos atributos que tienen una contribución parcial relevante para el consumidor.

c) ¿Cuál es la utilidad del atributo cantidad de grasa en el producto?

d) ¿Cuál es el mejor producto?

Conclusión capítulo 7

7. Introducción al Análisis Conjunto

La globalización ha presionado fuertemente a las empresas para sostener y mantener su participación de mercado. El diseño de productos centrado en las necesidades del cliente resulta esencial para lograrlo. La participación activa del cliente en el diseño de nuevos productos no solo incrementa la probabilidad de éxito del producto en el mercado sino que también contribuye a la retención y lealtad del cliente. Desde su introducción al inicio de los años 70, el análisis conjunto se ha convertido en una herramienta de análisis multivariante de gran utilidad para realizar la tarea del investigador de mercados en lo referente al diseño de productos que satisfagan las necesidades del consumidor (Green y Srinivasan 1978, 1990). La técnica utiliza las preferencias globales que los consumidores declaran para varias configuraciones de un producto para estimar utilidades o contribuciones métricas, part-worths, para cada característica o nivel específico de un atributo del producto. De esta forma se logra comprender la estructura de las preferencias de los consumidores al asumir que éstas se relacionan directamente con los beneficios percibidos para los varios atributos que definen el perfil de un producto. La selección de atributos y sus niveles o valores específicos es una de las tareas centrales en el análisis ya que sus combinaciones deben corresponder a un subconjunto representativo del universo de posibles productos.

El análisis conjunto, igual que los otros métodos multivariantes discutidos en este eBook, es una técnica que requiere de cálculos complejos por lo que es necesario contar con apoyo computacional. Puesto que se trata de un método altamente especializado en sus propósitos, los paquetes de aplicaciones estadísticas como MINITAB o SPSS no tienen disponible la opción para realizar análisis conjunto, sin embargo, hay opciones para obtener software especializado en análisis de conjunto.

Glosario general

A B C D E F G H I J K L M N Ñ O P Q R S T U V W X Y Z

A

Análisis de conglomerados

Técnica multivariante que se basa en utilizar la información de múltiples variables que describen a los objetos para dividirlos en grupos conocidos como clusters o conglomerados.

Análisis conjunto

Metodología que se apoya en los principios de utilidad percibida para un producto para estimar un modelo de descomposición de una preferencia global en el cual se determina la contribución parcial que cada atributo del producto hace a la utilidad global.

Análisis conjunto adaptativo (Adaptive Conjoint Analysis, ACA)

Metodología de análisis conjunto que utiliza la información aportada por los encuestados en etapas previas, por ejemplo importancia global asignada a los atributos generales de un producto, para determinar nuevos perfiles parciales de tal forma que se reduzca la cantidad de evaluaciones que hace el respondiente.

Análisis conjunto de selección discreta (Choice Based Conjoint, CBC)

Tipo de análisis conjunto que utiliza como entradas las selecciones que hicieron los consumidores para un conjunto pequeño de alternativas. Al realizar la selección de alternativas los participantes hacen compensaciones implícitas entre los atributos en forma similar a cuando se adquiere realmente un producto.

Análisis conjunto híbrido

Tipo de análisis conjunto en el cual primero se valoran los atributos para definir su utilidad relativa y después se evalúa un grupo de perfiles completos.

Análisis de componentes principales

Método para la estimación de la matriz factorial que consiste en calcular aquellas combinaciones lineales de las variables observables de máxima varianza. En este análisis se utiliza la matriz de correlación o la de varianza-covarianza asumiéndose que la varianza total de cada indicador es básicamente varianza común o compartida.

Análisis factorial confirmatorio

Tipo de análisis multivariable que se utiliza para confirmar las relaciones pre-definidas entre un número específico de variables latentes y un conjunto de variables observables.

Análisis factorial exploratorio

Tipo de análisis multivariable que se utiliza para explorar y comprender el porqué de las asociaciones entre un conjunto de variables tangibles. Las asociaciones se atribuyen a un conjunto desconocido de variables latentes o factores que se infieren a través de los indicadores tangibles.

Análisis factorial tipo R

En este análisis se busca identificar grupos de variables que forman dimensiones latentes o factores a partir de las

asociaciones observadas entre ellos.

ANOVA en un sentido

Se refiere al caso en que el análisis de varianza se utiliza para determinar el efecto que única variable no métrica, factor, tiene sobre una variable dependiente, métrica, o respuesta.

C

Cargas de los factores

Son medidas para el grado de correlación entre las variables originales y los factores latentes, que se utilizan para formar grupos de variables e identificar los factores. Las cargas elevadas al cuadrado indican el porcentaje de la varianza de una variable que se atribuye a un factor específico.

Centroide

Es el vector cuyas entradas son los promedios de un conjunto de variables.

Centroide

Es la media o el valor medio de los objetos incluidos en el conglomerado para cada variable.

Coeficiente de correlación (r)

Indica la fuerza de asociación entre las variables. El signo (+ o -) indica la dirección de la relación. Puede tomar valores entre -1 y +1, con +1 se indica una relación positiva perfecta , 0 la ausencia de relación y -1 una relación inversa o negativa perfecta.

Coeficiente de correlación

Es una medida del grado de asociación lineal entre dos variables estandarizadas.

Coeficiente de correlación parcial

Medidas de la fuerza de la relación entre variable criterio y una variable predictora, donde los efectos de la otras variables predictoras del modelo se mantienen constantes.

Colinealidad

La multicolinealidad aparece cuando una única variable predictora está altamente correlacionada con un conjunto de otras variables Independientes.

Comunalidad

Es aquella porción de la varianza de una variable tangible que es compartida con las otras variables debido a que todas son indicadores de un mismo conjunto de factores comunes. Son iguales a la suma de los cuadrados de las cargas por renglón, esto es sobre todos los factores extraídos.

Constructo

Concepto con un alto grado de abstracción lo que dificulta el medirlo directamente ya que su medición no está libre de errores sistemáticos de magnitud considerable. Hay conceptos con distintos grados de complejidad, aquellos que involucran varios componentes o dimensiones latentes pueden ser mejor especificados a través del análisis factorial.

Contribuciones parciales

Son las aportaciones de valor que cada atributo representa para el consumidor y que convenientemente agregadas reconstruyen la calificación global que se le asignó a un producto o servicio.

Covarianza

Medida estadística que cuantifica la intensidad de la asociación lineal que hay entre dos variables.

D

Datos heterocedásticos

Cuando el término de error tiene una varianza en aumento u ondulante.

Dendograma

Representación gráfica de los resultados de un procedimiento jerárquico. Muestra cómo se agrupan los conglomerados en cada paso hasta que se encuentran todos los objetos comprendidos en un conglomerado mayor.

Diagrama de puntos

Gráfica que resume a un conjunto relativamente pequeño de datos, cada observación se representa como un punto arriba del lugar que tiene en una escala de medición, si hay valores repetidos se apilan verticalmente.

Diagrama de vías

Representación gráfica de las relaciones entre múltiples variables tanto manifiestas como latentes.

Diseño ortogonal

Se refiere a la selección de una fracción de posibles perfiles o combinaciones de atributos que represente una cantidad razonable de estímulos para que un respondiente evalúe.

Distancia de Mahalanobis

Medida estadística para determinar la proximidad o distancia entre datos multivariables. A diferencia de la distancia euclideana, ésta toma en cuenta las correlaciones entre las variables.

E

Ecuaciones estructurales

Conjunto de ecuaciones que describe las relaciones que hay entre las variables latentes y entre éstas y los indicados tangibles.

Eigen-valor

También conocido como valor característico o valor propio. En el contexto de análisis factorial es una medida de la varianza de la combinación lineal de las variables observadas (componente principal). Para calcularlo se suman los cuadrados de las cargas por cada factor.

Eigen-vector

También denominado vector propio o característico. Se refiere a aquellos vectores que al ser multiplicados por una matriz no cambian de dirección sino solo de magnitud. En el contexto de análisis factorial, si se usa componentes principales como método de estimación de la matriz factorial, cada componente principal está asociado a un eigenvector de la matriz de correlación de varianza-covarianza.

Eliminación progresiva (Backward)

Método de selección de variables para incluirlas en el análisis de regresión que comienza con todas las variables independientes en el modelo y posteriormente va eliminando aquellas variables que no suponen una contribución significativa a la predicción.

Encadenamiento completo

Algoritmo de aglomeración en el que se representa la similitud como la distancia entre los miembros más diferentes de cada conglomerado.

Encadenamiento medio

Algoritmo de aglomeración en el que se representa la similitud como la distancia media de todos los objetos dentro de un conglomerado a todos los objetos de otro.

Encadenamiento simple

Procedimiento de aglomeración en el que se representa la similitud como la distancia mínima entre los objetos más cercanos de dos conglomerados.

Error de predicción

Diferencia entre los valores reales y de predicción de la variable criterio para cada observación en la muestra.

Error estándar de la estimación (SEE)

Medida de la variación en el valor previsto que puede usarse para desarrollar intervalos de confianza alrededor de cualquier variable a predecir.

Escala

Recurso de medición que establece un conjunto de normas para determinar la cantidad o magnitud de un atributo que posee un objeto cualquiera, asignarle un símbolo (usualmente un número) y así poder compararlo con otros objetos.

Estimación por pasos en la regresión múltiple

Método de selección de variables para su inclusión en el modelo de regresión que comienza con la selección del mejor predicción de la variable criterio. Las variables independientes adicionales se seleccionan en términos del aumento del poder explicativo que puede añadirse al modelo de regresión. Las variables independientes se añaden a medida que sus coeficientes de correlación parcial son estadísticamente significativos.

G

Gráfica de probabilidad

Diagrama que muestra en el eje horizontal los percentiles muestrales y en el eje vertical los percentiles de una distribución de probabilidad específica. La correspondencia entre ambos percentiles indica hasta dónde los datos se ajustan a una distribución de probabilidad determinada.

Gráfico de distribución normal

La distribución normal se representa por una línea recta con un ángulo de 45 grados. La distribución a comparar se dibuja contra esta línea y cualquier diferencia se muestra como desviación de la línea recta.

H

Histograma

Gráfica de barras en la cual el ancho de la barra es la longitud de un intervalo y la altura la frecuencia de datos en el intervalo. Es una representación empírica de la función de probabilidades o de densidad de los datos.

Homocedasticidad

Descripción de datos en los que la varianza del término de error (E) aparece constante sobre un rango de variables independiente.

I

Indicador manifiesto

Aquellas variables que se pueden observar directamente o medirse relativamente sin error.

Índices de ajuste

En el caso del análisis factorial exploratorio estos índices son medidas que evalúan hasta donde la matriz de covarianzas o correlaciones observadas se reproduce o predice a partir del modelo estimado. Las medidas de calidad del ajuste del modelo estimado incluyen índices de ajuste absoluto, de mejora incremental y de parsimonia o simplicidad del modelo estimado.

Intervalo

Tipo de escala invariante a cualquier transformación lineal, las diferencias en los números permiten una comparación de

elementos, el cero de la escala es arbitrario.

L

LISREL

Acrónimo de Linear Structural Relations un software especial para realizar análisis multivariable latente.

M

Matriz de correlaciones

Tabla que indica la correlación entre todas las variables, tanto independientes como la dependiente.

Matriz de patrones

Esta matriz se genera después de una rotación oblicua, sus entradas son las llamadas “cargas beta” que son equivalentes a coeficientes estandarizados entre cada indicador manifiesto y un determinado factor latente, estas cargas permiten reproducir los valores de las variables a partir de los puntajes factoriales.

Matriz de varianza-covarianza

Es una matriz cuadrada que en su diagonal contiene a las varianzas de un conjunto de variables en tanto el resto de sus entradas son las covarianzas entre parejas de variables.

Matriz factorial

Cada columna de esta matriz se asigna a un factor latente, las cargas de cada variable manifiesta para cada factor son las entradas de la matriz.

Matriz rotada

Se genera después de aplicar una rotación ya sea ortogonal u oblicua, sus entradas son las cargas de cada variable manifiesta en cada uno de los factores extraídos. Estas cargas

se distribuyen mejor entre los factores debido a la rotación aplicada a la matriz de cargas estimada originalmente.

Máxima verosimilitud

Método estadístico de estimación que propone como estimadores de los parámetros de interés a aquellas funciones de los datos que maximizan la llamada función de verosimilitud cuya forma se asume conocida. Los estimadores de MV son por tanto los valores más probables de los parámetros de la función de probabilidad que generó los datos observados.

Método Bayesiano Jerárquico

Método complejo de estimación de los partworths que se utiliza cuando para el análisis conjunto a nivel individual y se aplica aun cuando se tiene una pequeña cantidad de evaluaciones.

Método de Centroide

Algoritmo de aglomeración en el que la similitud entre los agregados se mide como la distancia entre los centroides de los grupos.

Método de dependencia

Método en el cual algunas de las variables explican o se identifican como causantes de la ocurrencia de otras variables.

Método de interdependencia

Son aquellos métodos en los cuales las variables bajo estudio guardan entre sí relaciones de asociación equivalentes; es decir, que no hay variables que se consideren resultado o efecto de otras variables.

Método de Ward

Procedimiento de análisis conglomerados jerárquicos, en el cual la similitud utilizada para unir conglomerados se calcula como la suma de los cuadrados entre los conglomerados para todas las variables.

Método o procedimiento de aglomeración

Los objetos que son más parecidos o de mayor similitud se combinan para construir un nuevo conglomerado agregado. Este proceso se continúa hasta que todos los objetos se combinan finalmente en un conglomerado.

Mínimos cuadrados

Procedimiento de estimación utilizado en la regresión simple y múltiple por lo que se estiman los coeficientes de regresión para reducir la suma total de los residuos cuadrados.

Modelo de descomposición aditivo

En este modelo la función de utilidad global a la que se asocia la preferencia completa es la suma de las contribuciones de cada uno de los atributos para el producto o servicio.

Modelo de descomposición con interacciones

En este tipo de modelo se considera la contribución a la utilidad que tienen combinaciones específicas (interacciones) de los niveles de una pareja de atributos.

Modelo logit

Tipo de modelo de selección aleatoria que presupone que la probabilidad de éxito, la selección o compra, de una alternativa depende de la utilidad percibida para el producto. Esta función de utilidad es una función de los atributos del producto más un componente aleatorio Єi, el modelo logit asume que estos errores son variables aleatorias independientes con una distribución Gumbel.

Multidimensionalidad

Se refiere a conceptos constituidos por varios componentes relativamente independientes entre sí pero que en conjunto cubren los múltiples aspectos que forman el dominio del concepto.

Multiescalas

Se refiere a aquellas escalas que utilizan múltiples reactivos (ítems) con el mismo formato para evaluar una característica

compleja o concepto.

N

Nominal

Son aquellas escalas invariantes a cualquier transformación uno a uno y en las cuales los números en la escala sólo identifican o clasifican objetos.

O

Ordinal

Escala invariante a cualquier transformación monotónica, los números indican una posición relativa pero no la magnitud de la diferencia entre ellos.

P

Perfilado

Procedimiento para definir y caracterizar un conglomerado o segmento. El nombre del segmento se asigna de acuerdo a las características dominantes de objetos que lo integran.

Perfiles (estímulos)

Son aquellas combinaciones de atributos específicos que resultan en productos completos que se solicita evaluar al participante.

Procedimientos no jerárquicos

Se utilizan las semillas del conglomerado para agrupar objetos dentro de una distancia previamente especificada. Este método optimiza la asignación de los objetos al conglomerado más cercano.

Puntajes factoriales

Combinaciones lineales de las variables observables que permiten estimar a los factores latentes. Para su cálculo es necesario estimar los coeficientes o pesos que se asignan a cada indicador. Los puntajes factoriales son distintos de los componentes principales pero el valor de los coeficientes o pesos es mayor cuando una variable tiene una carga grande en cierto factor.

R

Razón

Escala invariante a una transformación multiplicativa, el cero es fijo y los cocientes de la escala indican diferencias relativas entre elementos.

Residuo

La diferencia entre los valores reales y predichos de la variable dependiente.

Residuo estandarizado

Aquellos residuos que una media de 0 y una desviación estándar de 1.

Rotación

Procedimiento mediante el cual se rotan o giran los ejes de referencia de la solución factorial con el propósito de facilitar la identificación de grupos de variables y por ende de los factores latentes. Dependiendo de si la rotación preserva o no la no-correlación entre los factores, las rotaciones se clasifican como ortogonales o latentes.

Rotación Varimax

Tipo de rotación ortogonal muy utilizada que busca maximizar la variabilidad de las cargas en cada columna de la matriz factorial, lo que resulta en una mayor separación entre los

factores y una solución más fácil de interpretar que cuando se aplican otros tipos de rotación.

S

Semilla

En los algoritmos de partición se refiere a aquel dato que se utiliza como generador de un cluster.

Simulación

Consiste en analizar el efecto que la introducción que el o los nuevos productos tendrán en el mercado. La simulación se realiza empleando modelos de selección aleatoria que permiten predecir la participación en el mercado que alcanzaría un producto dependiendo de su perfil de atributos.

Suma del cuadrado de la regresión (SSR)

Representa la cantidad de mejora en la explicación de la variable independiente atribuible a las variables independientes.

Suma del cuadrado de los errores (SSE)

Se usa para denominar a la varianza en las variables dependientes, que aún no han sido tomadas, en cuenta en el modelo de regresión.

Suma total de lo cuadrados (TSS)

Cantidad total de variación, se explica por las variables independientes.

T

Transformación

Crea una nueva variable y elimina su característica indeseable lo que permite una mejor medida de la relación.

U

Utilidad

Función que relaciona el perfil de un producto con el beneficio global que el consumidor percibe para el conjunto de atributos que integran el perfil.

V

Variable latente

Se refiere a aquellas variables que sólo pueden ser inferidas a partir de indicadores tangibles.

Varianza

Medida estadística que cuantifica el grado de heterogeneidad de un grupo de datos.

Referencias

Capìtulo 1. Introducción al análisis multivariante

» Arroyo-López, P.E. y Borja-Medina, J.C. (2011). Pronósticos para la toma de decisiones: Aplicaciones en el contexto de negocios mexicano. Recuperado de: http://www.lulu.com/browse/search.php?search_forum=1&search_cat=2&show_results=topics&return_chars=200&search_keywords=&keys=&header_search=true&search=&locale=&sitesearch=lulu.com&q=&fListingClass=0&fSearch=Arroyo+y+pron%C3%B3sticos&fSubmitSearch.x=10&fSubmitSearch.y=7

» Bishop, Y.M., Fienberg, S.E. y Holland, P.W. (2007). *Discrete multivariate analysis: Theory and practice.* Nueva York, NY: Springer.

» Burns, W.C. (1997). *Spurious correlations.* Recuperado de: http://www.burns.com/wcbspurcorl.htm.

» Devore, J.L. (2008). *Probabilidad y estadística para ingeniería y ciencias* (7ª ed.). D.F., México: Thomson Editores.

» Do Valle, P.O., Rebelo, E., Reis, E. y Menezes, J. (2005). Combining behavioral theories to predict recycling involvement. *Environment and Behavior*, 37, 364-396.

» *Examples of spurious correlations* (s.f.). Recuperado de: http://www.southalabama.edu/coe/bset/johnson/oh_master/Ch1

1/Tab11-02.pdf.

» Hair, J.F. Jr., Anderson, R.E., Tatham, R.L. y Black W.C. (2002). *Análisis multivariante* (5a ed.). Madrid, España: Prentice Hall.

» Iglesias-Antelo, S. y Sulé-Alonso, M.A. (2003). *Introducción al análisis multivariable*. En J. P. Lévy-Mangin y J. Varela-Mallou (eds.) Análisis multivariable para las ciencias sociales. D.F., México: Pearson-Prentice Hall.

» Johnson, R.A. y Wichern, D.W. *Applied multivariate statistical analysis* (4a ed.). Upper Saddle River, NJ: Prentice Hall.

» Jöreskog, K. G. y Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago: Scientific Software International.

» Jöreskog, K.G. y Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.

» Nuval: índice de valor nutricional de los alimentos (23 de octubre de 2003). *Alimentación y salud*. Recuperado de http://alimentacion-salud.euroresidentes.com/2008/10/nuval-ndice-de-valor-nutricional-de-los.html

» Parasuraman, A., Zeithaml, V.A. y Berry, L.L. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing*, 29, 41-50.

» Rosenberg, Morris J. (1968). *The logic of survey analysis.* Nueva York, NY: Basic Books.

» Tonglet, M., Phillips, P.S. y Readc, A.D. (2004). Using the theory of planned behaviour to investigate the determinants of recycling behaviour: A case study from Brixworth, UK. *Resources, Conservation and Reciclying*, 41, 191-214.

Capìtulo 2. Análisis de factores

» Arroyo-López, P. E. y Sánchez-Maldonado, R. (2009). Programas de desarrollo de proveedores como estrategia para la competitividad empresarial. Memorias del 4º. Congreso Internacional de Sistemas de Innovación para la Competitividad. Universidad Iberoamericana. León, Gto. Agosto 2009.

» Abdi, H. (2003). Factor Rotations in Factor Analysis. En Lewis-Beck, M. Bryman, A. y Futing, T. (eds.) *Encyclopedia of Social Sciences Research Methods*. Thousand Oaks, CA: Sage.

» Catena, A., Ramos. M. y Trujillo, H. (2003). *Análisis multivariado, un manual para investigadores*. Madrid: Editorial Biblioteca Nueva.

» Costello, A. B. y Osborne, J. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assesment Research & Evaluation*, 10 (7). Recuperado de: http://pareonline.net/getvn.asp?v=10&n=7.

» DiStefano, C., Zhu, M. y Mindrila, D. (2009). Understanding and using factor scores: considerations for the applied researcher. *Practical Assesment Research & Evaluation*, 14 (20). Recuperado de: http://pareonline.net/pdf/v14n20.pdf

» Fabrigar, L. R., Wegener, D. T., MacCalum, R. C. y Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4* (3), 272-299.

» Fung, W. K. y Kwan, C. W. (1995). Sensitivity analysis in factor analysis: difference between using covarianza and correlation matrices. *Psychometrika*, 60, 607-614.

» Hair, J., Anderson, R., Tatham, R. y Black, W. (1999). *Análisis multivariante*, 5a. ed. México: Prentice Hall.

» Juwaheer, T. D. (2004). Exploring international tourists' perceptions of hotel operations by using a modified

SERVQUAL approach–a case study of Mauritius. *Managing Service Quality, 14* (5), 350-364.

» Kim, J. y Mueller, C. W. (1978). Factor analysis, statistical methods and practical issues. Sage university paper series on Quantitative Applications in the Social Sciences. Newbury Park, CA: Sage.

» MINITAB 16.0. (2003). User's Manual. MINITAB Inc.

» Pardo-Merino, A. y Ruíz-Díaz, M. A. (2005). Análisis de datos con SPSS 13 Base. Madrid: McGraw Hill.

» Stevens, J. (1986). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

» Stewart, D. W. (1981). The application and misapplication of factor analysis in marketing research. *Journal of Marketing Research, 18*, 51-62.

» Vicente y Oliva, M. A. y Manera-Bassa, J. (2003). El análisis factorial y por componentes principales. En Lévy-Mangin y Varela-Mallou (Eds.) *Análisis Multivariable para las Ciencias Sociales*. Madrid: Pearson & Prentice-Hall.

Capìtulo 3. Segmentación de consumidores a través del análisis de conglomerados

» Aldenderfer, M. S. y Blashfield, R. K. (1984). Cluster Analysis. Sage University Paper series on Quantitative applications in the Social Science. Newbury Park, California: Sage Publications.

» Arnold, S. J. (1986). A test for clusters. *Journal of Marketing Research*, 16 (November), 545-551.

» Buchta, C., Dolnicar, S. y Reutterer, T. (2002). A nonparametric approach to perceptions based marketing segmentation: Applications. Austria: SpringerWienNewYork.

» Cruz-Hernández, G. M. (2010). Un enfoque multicriterio para la asignación de embarques a esquemas de seguridad para el transporte de carga. Tesis de maestría en ingeniería industrial no publicada. Tecnológico de Monterrey campus Toluca. Toluca, Méx.

» Fernández, E. (2010). Handling multicriteria preferentes in cluster analysis. European Journal of Operational Research, 202 (3), 819-827.

» Finch, H. (2005). Comparison of distance measures in cluster analysis with dichotomous data. *Journal of Data Science*, 3, 85-100.

» Campo, E. P. y Del Campo, P. C. (2007). Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete FactoClass. *Revista Colombiana de Estadística*, 30 (2), 231-245.

» Jain, A. K. (2009). Data clustering: 50 years beyond k-means. Pattern Recognition Letters. Recuperado de: http://biometrics.cse.msu.edu/Publications/GeneralPRIP/JainDataClustering_PRL09.pdf

» Nemery de Bellevaux, P. (2006). Multicriteria clustering. Université Libre de Bruxelles. Faculté des Sciences Appliquées. Tesis de maestría no publicada. Bruselas, Bélgica.

» Pérez-López C. (2004). *Técnicas de Análisis Multivariante de Datos: Aplicaciones con SPSS*. Madrid: Pearson Educación.

» Picón-Prado, E., Varela-Mallou, J. y Lévy-Mangín, J. P. (2004). *Segmentación de mercados: aspectos estratégicos y metodológicos*. Madrid: Prentice Hall.

» Tryon, R. C. (1939). *Cluster analysis.* New York: McGraw-Hill.

» Vicente, P. y Reisa, E. (2005). Segmenting households according to recycling attitudes in a Portuguese urban area. *Resources, Conservation and Recycling*, 52, 1-12

Capítulo 4. Regresión lineal múltiple

» Belsay, D.A.,Kuh, E. y Welsch, R. E. (1980). Regression Diagnostic:Identifying Influencial Data and Sources of Collinearity. New York: John Wiley & Sons.

» Chen, C. (s.f.) Robust Regression and Outlier Detection with the ROBUSTREG Procedure. Extraido 21 de noviembre de:http://www2.sas.com/proceedings/sugi27/p265-27.pdf

» Draper, N. R. y Smith, H.(1998). Applied Regression Analysis. 3a. edición. New York: Wiley & Sons.

» Fox, J. (2002). Nonparametric Regression. Extraido 23 de noviembre de 2011 de: http://cran.r-project.org/doc/contrib/Fox-Companion/appendixnonparametric-regression.pdf.

» Hair, J., Anderson, R., Tatham, R. y Black, W. (1999). *Análisis multivariante*, 5a. ed. México: Prentice Hall.

» Hanke, J. E. (2006). *Pronósticos en los negocios*, 7a. ed. México: Pearson Educación.

» Kimes, S. E. y Fitzsimmons, J. A. (1990). Selecting profitable hotel sites at La Quinta Motor Inns, *Interfaces*, 20 (2), 12-20.

» MINITAB 16.0. User's Manual. State College, PA: MINITAB Inc.

» Montgomery, D.C.,Peck, E.A.,Vining, G.G.(2001).Introduction to linear regression analysis. Third edition. New York: Wiley & Sons.

» Neter,J,Kutmer,M.,Nathtsheim,C.J.,Wasserman, W. (1996).Applied Linear Statistical Models.Fourth edition.USA: Mc Graw Hill.

» Searle, S. R. (1997). Lineal Models. Wiley Classics Library. New York: Wiley & Sons.

» Tukey, J. W. (1977). Exploratory Data Analysis. Reading, MA: Addison-Wesley

» Weisberg, E. (1985). Applied Linear Regression. 2nd edition. New York: Wiley & Sons.

» Willan, A. W. & Watts, D. G. (1978). Meaningful Multicollinearity Measures. Technometrics, 20, pp 407-412.

Capítulo 5. Análisis discriminante

» Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. New York: Wiley & Sons.

» Arroyo López, P.E. (1997). *Measuring the Extent of Overlap of Brand Densities by Using MDA*. Tesis de doctorado no publicada. Doctorado en Administración, Escuela de Graduados en Administración, Tecnológico de Monterrey campus Cd. de México. Capítulo 3.

» Calvo-Silvosa, A. y Rodríguez-López, M. (2003). Análisis discriminante múltiple, en Lévy-Mangin, J. P. y Varela-Mallou, J. (Eds.) *Análisis Multivariable para las Ciencias Sociales*. Madrid: Pearson-Prentice Hall.

» Erosa-Martín, V. E. y Arroyo-López, P. E. (2011). Reacciones del consumidor mexiquense hacia las faltantes en anaquel. *Contaduría y Administración*. 233 (enero-abril), 33-53.

» Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179-188.

» Guerrero-Dávalos, C. y Terceño-Gómez, A. (en prensa). Cómo seleccionar y contratar empresas en el outsourcing utilizando la metodología de números borrosos. *Contaduría y Administración*. Recuperado de: http://www.contaduriayadministracionunam.com.mx/userFiles/app/pp_03022011.pdf.

» Johnson, R. A. y Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis*, Upper Saddle River, N. J.: Prentice-Hall. Capítulo 11.

» Lachenbruch, P. A. y Goldstein, M. (1979). Discriminant analysis, *Biometrics*, *35* (Marzo), 69-85.

» Malhotra, N. K. (1984). The use of linear logit models in marketing research. Journal of *Marketing Research, 2*, (February), 20-31.

» Malhotra, N. K. (2008). *Investigación de mercados* (5a. ed.). México, D.F.: Pearson-Prentice Hall.

» McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. New York: Wiley & Sons.

» Perreault, W. D. Jr., Behrman, D. N y Armstrong, G. M. (1979). Alternative approaches for interpretation of multiple discriminant analysis in marketing research*. Journal of Business Research, 7*, 151-173.

» Rao, C. (1948). The Utilization of Multiple Measurements in Problems of Biological Classification. *Journal of the Royal Statistical Society: Series B*, 10, 159–193.

» Robertson, T. S. y Kennedy, J. N. (1968). Prediction of consumer innovators: application of multiple discriminant analysis. *Journal of Marketing Research*, *5*, 64-69.

» (El) Semanario. (Enero 19, 2009). Crece 15% demanda global de lácteos 2003-2008. Extraido 25 de octubre, de: http://www.elsemanario.com.mx/news/news_display.php?story_id=14692

» Swets J. A. y Pickett R. M. (1982). *Evaluation of diagnostic systems: methods from signal detection theory*. Nueva York: Academic Press.

» Welch, B. (1939). Note on Discriminant Functions. *Biometrika*, 31, 218–220.

» West, P. M., Brockett, P. L. y Golden, L. L. (1997). A comparative analysis of neural networks and statistical

methods for predicting consumer choice. *Marketing Science*, 16 (4), 370-391.

Aviso legal

Arroyo López, María del Pilar Ester; Borja Medina, Julio César

Análisis multivariante para la inteligencia de mercados

p. 331 cm.

Análisis multivariante para la inteligencia de mercados/Pilar Ester Arroyo López y Julio César Borja Medina

LC: QA287 Dewey: 519.535

eBook editado, diseñado, publicado y distribuido por el Instituto Tecnológico y de Estudios Superiores de Monterrey.

Ave. Eugenio Garza Sada 2501 Sur Col. Tecnológico C.P. 64849 | Monterrey, Nuevo León | México.

Primera edición: enero 2013

ISBN 978-607-501-239-1