

# ESTADÍSTICA MULTIVARIANTE TOMO II

Técnicas Interdependientes  
con SPSS  
en las Ciencias Sociales

Juan Mejía Trejo



**AMIDI**  
Academia Mexicana  
de Investigación y Docencia  
en Innovación



# **ESTADÍSTICA MULTIVARIANTE TOMO II**

**Técnicas  
Interdependientes  
con SPSS  
en las Ciencias  
Sociales**

*Juan Mejía Trejo*



Este libro fue sometido a un proceso de dictamen por pares de acuerdo con las normas establecidas por el Comité Editorial de la Academia Mexicana de Investigación y Docencia en Innovación (AMIDI)

Primera edición, 2023

D.R. © Academia Mexicana de Investigación y Docencia en Innovación SC (AMIDI)  
Av. Paseo de los Virreyes 920.  
Col. Virreyes Residencial  
C.P. 45110, Zapopan, Jalisco, México

**ISBN Tomo II: 978-607-59397-9-7**

**ISBN Obra Completa: 978-607-59397-7-3**

## Contenido

Introducción.....	5
<b>Capítulo 12. Análisis Factorial .....</b>	<b>6</b>
12.1. ¿Qué es el análisis factorial? .....	6
12.2. Análisis factorial y el proceso de decisión .....	9
12.3. Análisis factorial: Objetivos .....	11
12.3.1. La identificación de estructura mediante el resumen de datos .....	11
12.3.2. Reducción de datos.....	11
12.3.3. El uso del análisis factorial con otras técnicas multivariantes.....	12
12.3.4 Selección de variables .....	13
12.4. Análisis factorial: Diseño .....	13
12.4.1. Las correlaciones entre las variables o los encuestados.....	14
12.4.2. La selección de variables y cuestiones de medición .....	15
12.4.3. Tamaño muestra .....	15
12.5. Análisis factorial: Supuestos.....	16
12.6. Análisis factorial: Estimación y Ajuste .....	18
12.6.1. Análisis factorial común vs. Análisis de componentes principales .....	18
12.6.2. Criterios para el cálculo del número de factores a ser extraídos .....	20
12.6.3. Descripción de los criterios para el cálculo del número de factores a ser extraídos.....	20
12.7. Análisis factorial: Interpretación de factores.....	23
12.7.1. Rotación de factores.....	24
12.7.2. Métodos de rotación ortogonal.....	27
12.7.3. Métodos de rotación oblicua .....	28
12.7.4. Selección del método de rotación.....	28
12.7.5. Criterios para la significación de la carga factorial .....	28
12.7.6. Interpretación de la matriz de factores .....	30
12.8. Análisis factorial: Validación .....	33
12.9. Análisis factorial: Usos adicionales de los resultados.....	34
12.9.1. Selección de variables suplentes para el análisis subsiguiente .....	34
12.9.2. Creación de escalas aditivas .....	35
12.9.3. Cálculo de la puntuación factorial.....	39
12.9.4. Selección entre los 3 métodos .....	39
12.10. Análisis factorial: Resumen para aplicar.....	40
12.11. Análisis factorial. Ejemplos .....	42
Referencias .....	71
<b>Capítulo 13. Análisis Multidimensional y de Correspondencias .....</b>	<b>74</b>
13.1.-El análisis multidimensional. ¿Qué es?.....	74
13.2. Análisis multidimensional. Cómo actúa .....	76
13.3. Análisis multidimensional vs. otras técnicas de enfoque interdependiente .....	79
13.4. Análisis multidimensional. Paso 1: Objetivos .....	82
13.5. Análisis multidimensional. Paso 2: Diseño .....	84
13.6. Análisis dimensional. Paso 3: Condiciones de aplicabilidad.....	91
13.7. Análisis multidimensional. Paso 4: Estimación y ajuste.....	91
13.8. Análisis multidimensional. Paso 5: Interpretación.....	102
13.9. Análisis multidimensional. Paso 6: Validación .....	103
13.10. Análisis multidimensional. Ejemplos.....	104
13.11. Análisis de correspondencias.....	118

13.12. Análisis de correspondencias. Ejemplos.....	121
Referencias .....	133
<b>Capítulo 14. Análisis Cluster.....</b>	<b>136</b>
14.1. Análisis cluster ¿qué es?.....	136
14.2. Análisis cluster ¿cómo funciona? .....	138
14.3. Análisis cluster. Medición de la similitud y creación de conglomerados .....	140
14.4. Análisis cluster ¿cuántos conglomerados formar?.....	143
14.5 Análisis cluster. Paso 1: Objetivos .....	147
14.6. Análisis cluster. Paso 2: Diseño .....	148
14.7. Análisis cluster. Paso 3: Condiciones de aplicabilidad.....	157
14.8. Análisis cluster. Paso 4: Ejecución y ajuste del conglomerado .....	158
14.9. Análisis cluster. Paso 5: Interpretación de los conglomerados .....	167
14.10. Análisis cluster. Paso 6: Validación y perfil de grupos .....	168
14.11. Análisis cluster. Resumen .....	169
14.12. Análisis cluster jerárquico. Ejemplos .....	170
14.13. Análisis cluster. Observaciones finales .....	191
Referencias .....	192
<b>Apéndice. Matriz de pruebas estadísticas sugeridas.....</b>	<b>194</b>

## Introducción

Al tener mayores capacidades y disponibilidad de los recursos de cómputo, hoy en día, hace que el análisis multivariante se presente en varias aplicaciones de software por lo que se incrementan las posibilidades de ser usado en diversas disciplinas como las Ciencias de la Administración, siendo el Statistical Package for the Social Sciences (SPSS, de IBM), el Analytics, Business Intelligence and Data Management (SAS, de SAS Institute y/o de World Programming), Statistica (de STATISTICA), el lenguaje R (software libre) sólo por mencionar algunos como de los más utilizados en los campos académico y profesional a nivel mundial.

Así que, no es de extrañar que en las Ciencias Sociales, se observe de manera creciente un repunte en la presentación de reportes, artículos, capítulos de libro o libros que discutan diversos aspectos teórico empíricos y su interpretación basados en dichas aplicaciones de software.

En nuestro caso, adoptamos SPSS 20 de IBM, para el desarrollo de los temas de este libro. Basados en lo anterior, mostramos esta obra que tiene como principales objetivos:

- 1.-Presentar un documento que sirva a propios y extraños al tema, que tengan la necesidad de conocer tanto los conceptos tratados en este tomo, como el de manipular los diversos comandos que ofrece SPSS 20 de IBM al respecto de los casos problema, presentados como ejemplo.

- 2.-Para una mayor comprensión del tratamiento de los casos, se expone la secuencia propuesta por Hair et al. (1999) de los 6 pasos: objetivos, diseño, supuestos, ejecución, interpretación y validación, como el eje de presentación y resolución de dichos caso.

3. El desarrollo básico de las técnicas de : análisis factorial, análisis multidimensional y de correspondencias así como el análisis cluster.

Es deseo del autor, contribuir en el lector en la adquisición de conocimiento que se aplique en el mundo práctico y que ayude a su interpretación teórica. Si no fuere el caso, se espera que al menos sirva como otro peldaño útil a escalar en el logro de su formación académica y/o profesional.



## Capítulo 12. Análisis Factorial



### 12.1. ¿Qué es el análisis factorial?

Esta técnica estadística multivariante ha tenido una creciente utilización desde fines de los 90s en todas las áreas de investigación de las ciencias de la administración de carácter empresarial. A medida que se incrementa el número de variables que intervienen en las técnicas multivariantes, se crea la necesidad también mayor de conocer a profundidad tanto la estructura como las interrelaciones de las variables. El **análisis factorial**, es una técnica especialmente adecuada para analizar las pautas de relaciones complejas y multidimensionales encontradas por los investigadores del campo de las ciencias de la administración. El objetivo en este apartado es el de definir y explicar los aspectos fundamentales de las técnicas de **análisis factorial** en términos conceptuales, lo más amplios posibles. Es posible aplicarlo para examinar las pautas subyacentes o las relaciones para un amplio número de variables y determinar si la información puede ser resumida en una serie de factores o componentes más pequeños, por lo que se necesitará conocer de las directrices básicas para presentar e interpretar los resultados de estas técnicas. Para saber más, ver IBM, 2011a; IBM, 2011b, IBM, 2011c.

El **análisis factorial** es el nombre genérico de aquellos métodos estadísticos multivariantes que se enfocan en **definir la estructura subyacente en una matriz de datos**. Normalmente, incluye el tratar el problema de **cómo analizar la estructura de las interrelaciones (correlaciones)** entre un gran número de variables (por ejemplo, las puntuaciones de prueba, artículos de prueba, respuestas de cuestionarios) con la definición de una serie de **dimensiones subyacentes comunes, conocidas como factores**. Con el **análisis factorial**, podrá identificar, en principio, las **dimensiones separadas de la estructura** y entonces **determinar el grado en que se justifica cada variable por cada dimensión**. Una vez determinadas las



dimensiones y la explicación de cada variable, se pueden lograr los dos objetivos principales para el análisis factorial: **el resumen y la reducción de datos**. Al momento de resumir los datos, se obtienen las **dimensiones subyacentes** que, interpretadas y comprendidas, **describen los datos con un número de conceptos mucho más reducido que las variables individuales originales**.

La **reducción de datos** se obtiene con el **cálculo de la puntuación para cada dimensión subyacente y sustituirlos por las variables originales**. Esta técnica multivariante, es de las primeras en considerar su capacidad para **hacer acomodos** de las variables múltiples con el objetivo de comprender las relaciones complejas que **no son posibles con los métodos univariantes y bivariantes**.

Al aumentar el número de las variables **también aumenta la posibilidad** de que las **variables estén no correlacionadas y no sean representativas de unos conceptos distintos**. En su lugar, los **grupos de variables pueden estar interrelacionados** en la medida en que son todos representativos de un concepto más general, tendiendo como causales posibles: el diseño, el intento de medir las muchas facetas de personalidad o la imagen del establecimiento, o puede surgir simplemente de la adición de nuevas variables. Usted **tiene que saber cómo se relacionan las variables** para interpretar mejor los resultados. Si el **número de variables es demasiado grande o deba dar una mejor representación a un número de conceptos más pequeño** en vez de las facetas múltiples, ésta técnica ayuda en la **selección de un subgrupo de variables representativo** o incluso **crear nuevas variables** como sustitutas para las variables originales mientras mantengan su carácter original.

El análisis factorial es diferente de las **técnicas de dependencia (regresión múltiple, el análisis discriminante, el análisis multivariante de la varianza o la correlación canónica)**, en el que se consideran una o más variables explícitamente como las variables de criterio o dependientes y todas las demás son las **variables de predicción o independientes**.

El **análisis factorial** es una **técnica de interdependencia** en la que se consideran **todas las variables simultáneamente**, cada una relacionada con todas las demás y empleando todavía el concepto del valor teórico, el compuesto lineal de las variables.

**En el análisis factorial**, los valores teóricos (los factores) se forman para **maximizar** su explicación de la serie de variables entera, **y no para predecir una(s) variable(s) dependiente(s)**. Si hiciéramos una analogía con las técnicas de dependencia, cada una de las variables (originales) observadas sería una variable dependiente, que es una función de una serie de factores (dimensiones) subyacentes y latentes que están compuestas por todas las otras variables. Por tanto, cada variable es predicha por todas las demás. Por el contrario, se puede considerar cada factor (valor teórico) como una variable dependiente que es una función del conjunto entero de las variables observadas.

Cualquiera de estas analogías ilustra las diferencias de propósito entre **las técnicas de dependencia** (la predicción) y **la interdependencia** (identificación de estructura). Las técnicas analíticas de factores pueden lograr sus propósitos desde una perspectiva **exploratoria** o **confirmatoria**. Muchos investigadores lo consideran solamente **exploratorio**, útil para la búsqueda de una estructura entre una serie de variables o como un **método de reducción de datos**. Desde esta perspectiva, las

técnicas de análisis factorial **“extraen lo que proporcionan los datos”** y no tienen ninguna restricción en la estimación de los componentes o el número de componentes a ser extraído. Para muchas aplicaciones, si no todas, resulta apropiada esta aplicación del análisis factorial. No obstante, en otras situaciones tendrá sus pensamientos preconcebidos sobre la estructura real de los datos, que se basan en un apoyo teórico o investigaciones previas. Es posible que el investigador quiera probar las hipótesis que implican cuestiones tales como qué variables deberían ser agrupadas en un factor o el número exacto de factores. En estos casos, se requiere un **análisis factorial que adopte un enfoque confirmatorio --es decir, valorar hasta qué punto los datos se ajustan a la estructura esperada**. Los métodos tratados en esta sección NO proporcionan la estructura necesaria para la prueba de hipótesis formalizada. Abordamos explícitamente la **perspectiva confirmatoria** del análisis factorial está radicada en las **ecuaciones estructurales**. En este capítulo, sin embargo, observamos las técnicas analíticas de factores principalmente desde un punto de vista exploratorio o no confirmatorio.

### **Problema hipotético:**

Supongamos que mediante una investigación cualitativa una empresa de servicios de mercadotecnia digital ha identificado 100 características diferentes de empresas contratantes de sus servicios (consumidores), que como consumidores han mencionado que afectan su elección de en la planeación de la campaña web. La empresa de mercadotecnia digital quiere entender cómo deciden sus consumidores acerca de sus servicios, pero opina que no puede valorar las 100 características individuales o desarrollar planes de acción para tantas variables, porque son demasiado específicos. En su lugar, **a la empresa de mercadotecnia digital le gustaría saber si sus consumidores piensan en una dimensión determinante más general en vez de únicamente en aspectos específicos**. Para identificar estas **dimensiones**, la empresa de mercadotecnia digital deberá:

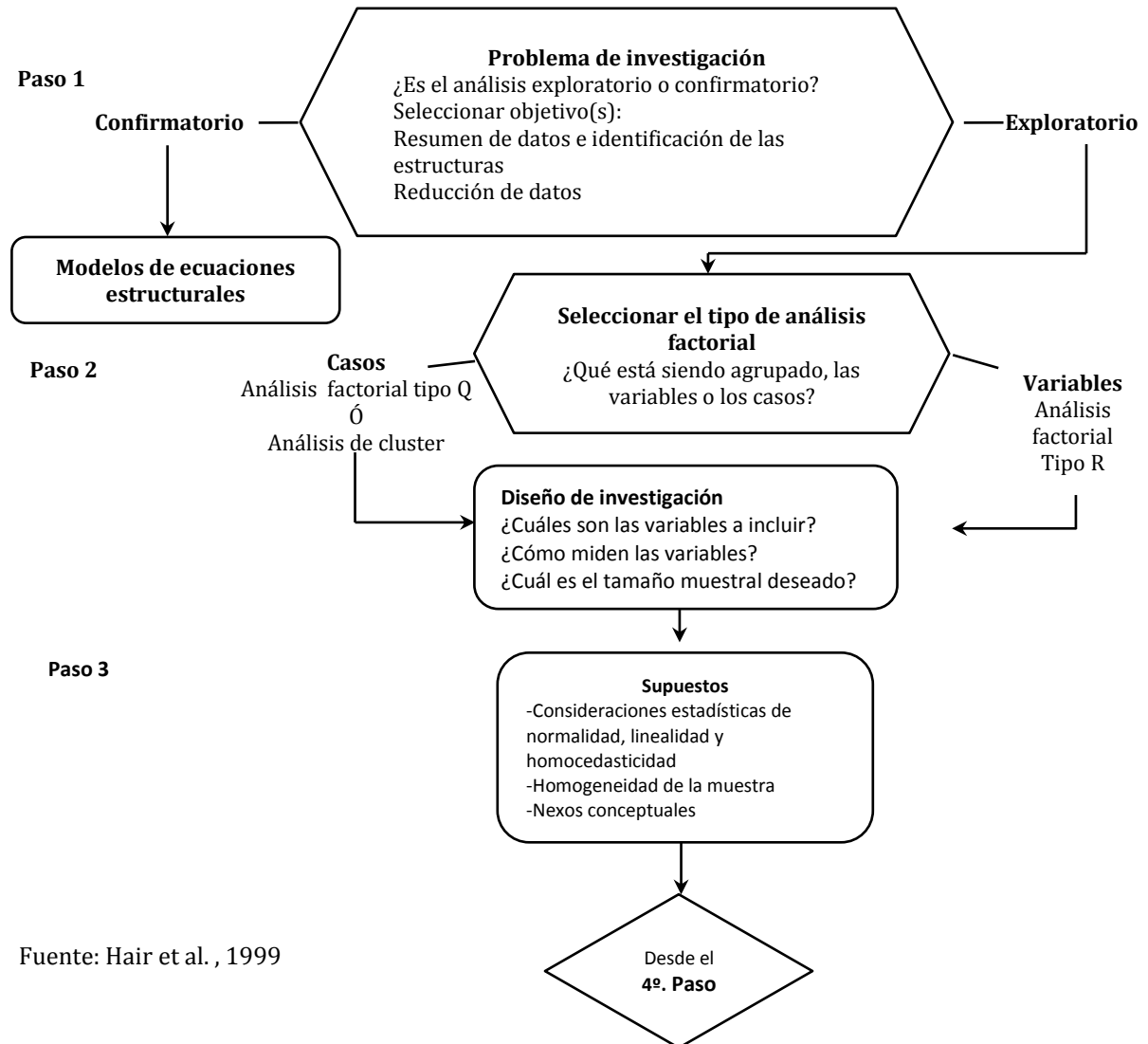
1. Realizar una encuesta solicitando valoraciones de consumidores sobre cada uno de estos aspectos específicos.
2. Emplear el **análisis factorial para identificar las dimensiones determinantes subyacentes**. Se considera que los aspectos específicos que se correlacionan en gran medida forman parte de una dimensión más amplia.
3. **Convertir éstas dimensiones en compuestos de las variables específicas**, que a su vez permitan a las dimensiones ser interpretadas y descritas. En nuestro ejemplo, el análisis factorial podría identificar **dimensiones como planeación, estrategia, implementación, tecnología, experiencia de usuario** como las dimensiones determinantes utilizadas por los encuestados.

Cada una de estas dimensiones contiene aspectos específicos que son una faceta de la dimensión determinante más amplia. A raíz de estos resultados la empresa de servicios de mercadotecnia digital puede usar estas dimensiones (factores) para definir áreas generales para la planificación y actuación.

## 12.2. Análisis factorial y el proceso de decisión

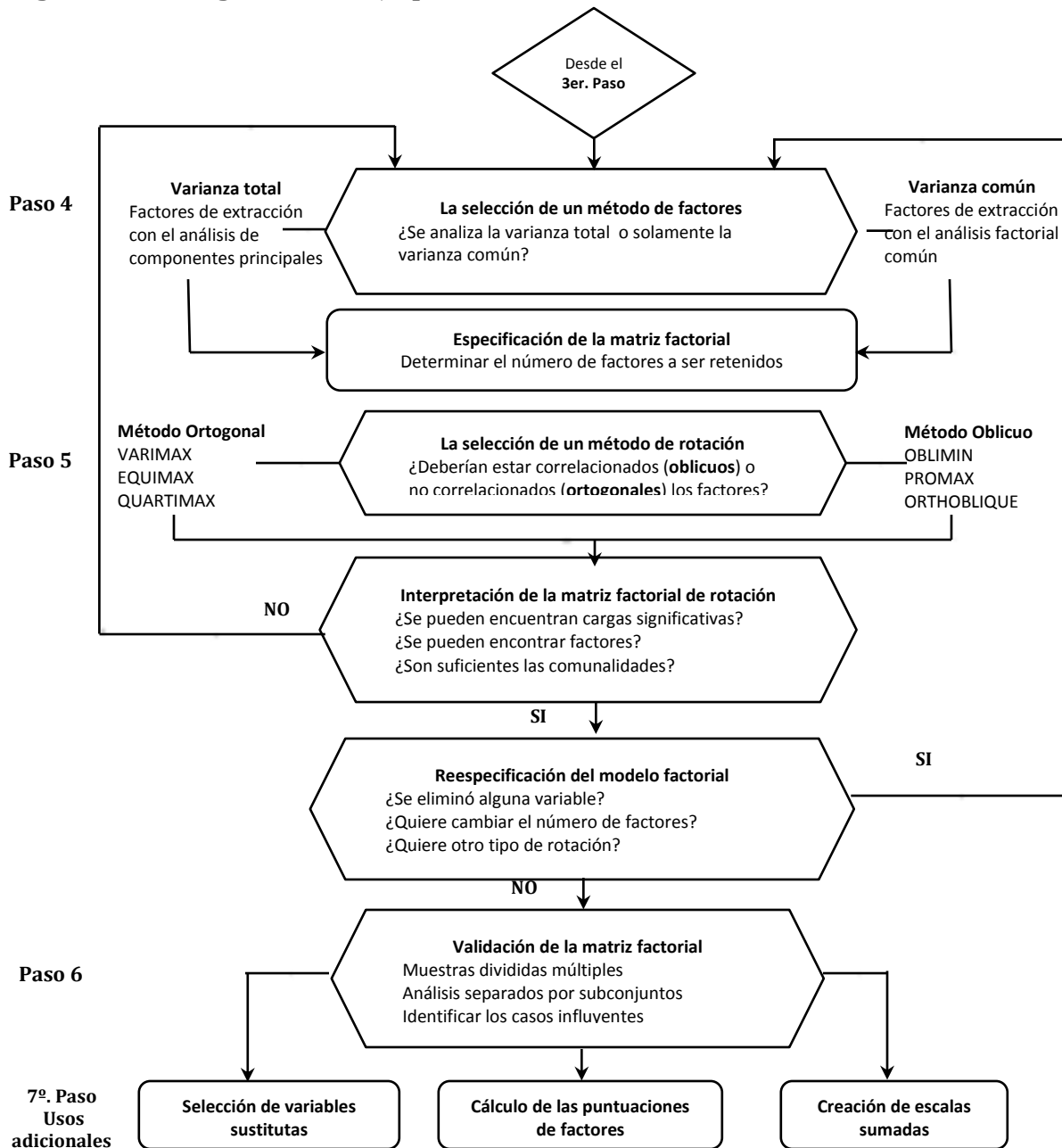
De acuerdo al modelo de seis pasos de Hair et al. (1999) que se introdujo en el **Capítulo 2**, la **Figura 12.1** muestra tres pasos iniciales de la aproximación estructurada para la construcción de modelos multivariantes.

**Figura 12.1 . Diagrama de flujo pasos 1-3 del análisis factorial**



Por otro lado, la **Figura 12.2** muestra en detalle los últimos tres pasos, y uno adicional (el séptimo paso) más allá de la estimación, la interpretación y la validación de los modelos factoriales, que ayuda a la selección de las variables sustitutas, las puntuaciones de factores o la creación de las escalas aditivas para la utilización en otras técnicas multivariantes. A continuación se presenta un análisis de cada paso

**Figura 12.2. Diagrama de flujo pasos 4-6 del análisis factorial**



Fuente: Hair et al., 1999

### 12.3. Análisis factorial: Objetivos

#### Paso 1: establecimiento de objetivos

Como con otras técnicas estadísticas, inicia con el **problema sujeto de investigación**. El propósito general de las técnicas analíticas de factores es **encontrar una manera de resumir** la información contenida a partir de una serie de variables originales **en una serie más pequeña de dimensiones compuestas o valores teóricos (factores) nuevos** aceptando una mínima pérdida de información (en otras palabras, buscar y definir las construcciones fundamentales o dimensiones que se supone sirven de base para las variables originales [Gorsuch, 1983, Rurnnel, 1970].

Las técnicas del análisis factorial deben satisfacer cualquiera de estos dos objetivos:

1. **La identificación de estructura mediante el resumen de datos, o bien**
2. **La reducción de datos.**

#### 12.3.1. La identificación de estructura mediante el resumen de datos

El análisis factorial puede identificar la estructura de las relaciones entre los encuestados o las variables mediante el **análisis de las correlaciones** entre encuestados y/o variables. Por ejemplo, suponga que se tienen datos sobre 100 encuestados basados en 10 características. Si el objetivo de la investigación fuera:

- a) Aplicar el análisis factorial a la matriz de **correlación de los encuestados individuales basada en sus características**, se le denomina el **análisis factorial Q**, siendo un método para combinar o condensar grandes grupos de personas en grupos claramente diferentes dentro de una población mayor, no se usa la aproximación del **análisis factorial Q** con mucha frecuencia. En su lugar, la mayoría de los investigadores utilizan algún tipo de **análisis cluster** para agrupar los encuestados individuales. También véase [Stewart 1981] para otras combinaciones posibles de grupos y tipos de variables
2. El resumen de las características, se aplicaría el análisis factorial a una matriz de **correlación de las variables**. **Éste es el tipo de análisis factorial más común**, y se denomina el **análisis factorial R**, el cual analiza las variables para identificar las **dimensiones que son latentes** (las que no son fácilmente observadas).

#### 12.3.2. Reducción de datos

El análisis factorial también puede:

1. Identificar **las variables suplentes** de una serie de variables más grande para su utilización en análisis multivariantes posteriores, o
2. Crear una serie de **variables completamente nueva, mucho más pequeña en número**, para reemplazar parcial o completamente la serie original de variables para su inclusión en técnicas posteriores.

En ambos casos, el propósito es retener la naturaleza y el carácter de las variables originales reduciendo su número para simplificar el análisis multivariante posterior. Aunque las técnicas multivariantes se han desarrollado para utilizar múltiples variables, **Usted debe siempre buscar la serie de variables más reducida para incluirla en el análisis**. Las cuestiones conceptuales y las empíricas deben respaldar la creación de medidas compuestas. El **análisis factorial proporciona la base empírica** para valorar la estructura de las variables y la potencial para crear estas

medidas compuestas o seleccionar una subserie de variables suplentes para el análisis posterior. **El resumen de datos** hace que la identificación de las dimensiones subyacentes o los factores sean fines de por sí; las estimaciones de los factores y las contribuciones de cada variable a los factores (**denominadas cargas de los factores**) constituyen todo lo que se necesita para el análisis. La reducción de datos depende también de las cargas de los factores; no obstante, las utiliza como la base para identificar las variables para su análisis subsiguiente con otras técnicas o bien para hacer estimaciones de los factores mismos (puntuaciones de factores o escalas aditivas), que a su vez reemplazan las variables originales en análisis subsiguientes.

### **12.3.3. El uso del análisis factorial con otras técnicas multivariantes**

El análisis factorial **proporciona una visión directa de las interrelaciones entre las variables o los encuestados** y un apoyo empírico para abordar las cuestiones conceptuales que tienen relación con la **estructura subyacente de los datos**. Es un complementario importante con otras técnicas multivariantes mediante **el resumen y la reducción de datos**.

Como **resumen de datos**, el análisis factorial proporciona una comprensión clara de cuáles de las variables podrían actuar juntas y cuántas de las variables realmente se puede esperar que tengan un impacto en el análisis. **Por ejemplo**, se esperaría que las variables altamente correlacionadas y miembros del mismo factor tuvieran perfiles similares de diferencia a través de los grupos en el **análisis multivariante de la varianza** o en el **análisis discriminante**. Los procedimientos que muestran el impacto de las variables correlacionadas son los basados en etapas (*stepwise*) de la **regresión múltiple o el análisis discriminante**. Estas técnicas **introducen las variables secuencialmente**, basadas en su **capacidad adicional de predicción** sobre las variables en el modelo. Conforme entra la variable de un factor, **es menos probable que variables adicionales del mismo factor sean también incluidas**, porque están altamente correlacionadas y potencialmente tienen **menos capacidad de predicción adicional**, que las variables que no estén en ese factor. Esto no significa que las otras variables del factor sean menos importantes o tengan menos impacto, sino que **su efecto ya ha sido representado por la variable incluida en ese factor**. Esta visión puede ser incorporada directamente a otras técnicas multivariantes mediante cualquiera de las **técnicas de reducción de datos**.

El análisis factorial proporciona la base para **crear una nueva serie de variables** que incorporan el carácter y naturaleza de las variables originales en una **cantidad de nuevas variables más reducida**, sea con la utilización de **variables suplentes**, sea con la **puntuación de factores o las escalas aditivas**.

De esta manera, se pueden reducir los problemas que se asocian con las grandes cantidades de variables o intercorrelaciones altas entre las variables con la sustitución de las nuevas variables. Así la investigación se beneficia de las relaciones y la visión detallada de la base conceptual y la interpretación de los resultados.

#### 12.3.4 Selección de variables

La **reducción y el resumen de datos** pueden ser llevados a cabo tanto con una serie de **variables preexistentes** como con las **variables creadas** por la nueva investigación. Cuando plantea el uso de una nueva serie, debe realizar una **aproximación conceptual** para determinar qué variables conviene incluir en el análisis. El uso del análisis factorial para la **reducción de datos** es particularmente crítico cuando se requiere la comparabilidad a lo largo de un período de tiempo o en situaciones múltiples. Cuando se usa en una nueva investigación, el análisis factorial puede **determinar también la estructura y/o crear nuevas puntuaciones compuestas** a partir de las variables originales. **Por ejemplo**, uno de los primeros pasos en la construcción de la escala aditiva es valorar la naturaleza de su dimensión y la conveniencia de las variables seleccionadas mediante el análisis factorial.

Por tanto, **aunque no es verdaderamente confirmatorio, el análisis factorial se puede utilizar para valorar la naturaleza de una dimensión propuesta.**

Una vez que se especifica el propósito del análisis factorial, Usted tiene que definir la **serie de variables a examinar**. Por lo que se refiere tanto al análisis factorial **tipo R** o **tipo Q**, Usted especifica implícitamente las dimensiones potenciales que se pueden identificar mediante el carácter y la naturaleza de las variables sujetas al análisis factorial. **Por ejemplo**, en la valoración de las dimensiones de la **experiencia del usuario**, el análisis factorial no podría identificar esta dimensión si no han sido incluidas preguntas sobre el **acceso, la usabilidad y la visibilidad a la página web**. Debe recordar también que el análisis factorial siempre **producirá factores**. Por tanto, el análisis factorial es siempre un candidato potencial para el fenómeno **"basura dentro, basura fuera"**. **Si el investigador incluye indiscriminadamente grandes cantidades de variables y espera que el análisis factorial "lo solucione", entonces la posibilidad de obtener malos resultados es alta.** La calidad y el significado de los factores derivados reflejan un acercamiento conceptual a las variables incluidas en el análisis. El uso del análisis factorial como una técnica de resumen de datos **no excluye la necesidad de una base conceptual para cualquiera de las variables analizadas**. Incluso si se usa meramente para la reducción de datos, el análisis factorial es más eficiente cuando las dimensiones definidas conceptualmente pueden ser representadas por los factores obtenidos.

#### 12.4. Análisis factorial: Diseño

##### Paso 2: Diseño

El diseño de un análisis factorial implica tres decisiones básicas:

1. El cálculo de los datos de entrada (**matriz de correlación**) para alcanzar los objetivos específicos de la agrupación de variables o encuestados;
2. El diseño del estudio en términos de **número de variables**, las **propiedades de medición** de las variables y los tipos de las **variables permisibles**; y
3. **El tamaño de muestra necesario**, tanto en términos absolutos como para la función del número de variables en el análisis.



### 12.4.1. Las correlaciones entre las variables o los encuestados

La primera decisión en el diseño de un análisis factorial se concentra en la **aproximación** que se usa para calcular la **matriz de correlación** tanto para el análisis factorial de **tipo R** como para el de **tipo Q**. Usted puede:

1. Utilizar la matriz de datos de entrada a partir del cálculo de las correlaciones entre las variables, empleando, por tanto, un **análisis factorial de tipo R**.
2. Elegir la matriz de correlación de las correlaciones entre los encuestados individuales, o **análisis factorial tipo Q**, el resultado será una matriz factorial que identifica individuos similares. **Por ejemplo**, si los encuestados individuales se identifican por un número, la pauta de factores de resultado podría indicarnos que los individuos 1, 5, 6 y 7 son similares. Del mismo modo, los encuestados 2, 3, 4 y 8 quizá cargarían juntos sobre otro factor, y clasificaríamos estos individuos como similares. A partir de los resultados del **análisis factorial Q**, se pueden **identificar grupos o clusters** de individuos que muestran una pauta similar sobre las variables que se incluyen en el análisis. A estas alturas, una pregunta lógica sería, **¿cómo se diferencia el análisis factorial tipo Q del análisis cluster, dado que ambas aproximaciones comparan la pauta de respuestas a través de una serie de variables y clasifican a los encuestados en grupos?** La respuesta es que el **análisis factorial tipo Q** se basa en las **intercorrelaciones** entre los encuestados, mientras que el **análisis cluster** forma grupos que se basan en una medida de similitud basada en la **distancia entre las puntuaciones de los encuestados sobre las variables analizadas**. Para ilustrar esta diferencia, consideremos la **Figura 12.3**, que contiene las puntuaciones de cuatro encuestados sobre tres variables diferentes. Un **análisis factorial tipo Q** de estos cuatro encuestados daría dos grupos con estructuras de **covarianza** similares, agrupando a los encuestados **A y C frente a B y D**. Por contraste, la aproximación de **cluster** sería sensible a las distancias reales entre las puntuaciones de los encuestados y llevaría a la agrupación de las parejas más cercanas. Por tanto, con la aproximación del análisis cluster, los encuestados **A y B estarían situados en un grupo y C y D en el otro grupo**, por lo que se debe estar consciente de estas diferencias. Con la disponibilidad de otras técnicas de agrupación y el uso general del análisis factorial para la reducción de datos y el resumen, la exposición restante de este capítulo se centra en el **análisis factorial tipo R, la agrupación de variables** en vez de la agrupación de encuestados.

**Figura 12.3. Aproximación cluster**

	Variables		
Encuestados	1	2	3
<b>A</b>	2	2	3
<b>B</b>	3	1	1
<b>C</b>	7	7	8
<b>D</b>	8	6	6

8	X		X	Encuestado C
7	X	X		
6		X	X	Encuestado D
5				
4				
3	X		X	Encuestado A
2	X	X		
1		X	X	Encuestado B
0	V1	V2	V3	

Fuente: propia

### 12.4.2. La selección de variables y cuestiones de medición

A estas alturas, deberán resolverse preguntas:

1. ¿Cómo se miden las variables? Y
2. ¿Cuántas variables deberían ser incluidas?

Por regla general, las variables a incluir en el análisis factorial tienen que ser de **escala métrica**. En algunos casos, se pueden utilizar **variables ficticias (codificadas 0-1)**, aunque se consideran como **no métricas**. **Si todas las variables son ficticias**, entonces las formas especializadas del análisis factorial, tales como el **análisis factorial Booleano**, son más apropiadas [BMDP Statistical Software 1992]. Además, deberá intentar **minimizar el número de variables que se incluyen**; no obstante, **debe también mantener un número razonable de variables por factor**. Si se está diseñando un estudio para valorar una estructura propuesta, deberá incluir varias variables (**5 o más**) que puedan representar cada **factor propuesto**. El poder del análisis factorial se basa en **encontrar pautas entre grupos de variables y resulta de poca utilidad en la identificación de factores compuestos de una única variable**. Finalmente, cuando se diseña una investigación que utiliza análisis factorial, debería, si cabe, **identificar varias variables claves (denominadas indicadores claves o variables marcadoras)** que reflejan con detalle los factores subyacentes hipotéticos, de forma que se facilite la validación de los factores derivados y la valoración sobre la significación práctica de los resultados.

### 12.4.3. Tamaño muestra

Se recomienda **NO** usar el análisis factorial **para una muestra inferior a 50 y preferentemente de 100 o más observaciones**. Como regla general:

1. El mínimo es tener por lo menos un número de observaciones **5 veces mayor que el número de variables a ser analizadas**
2. El tamaño aceptable deberá estar en un ratio de **diez a uno**. Algunos investigadores proponen incluso un mínimo de **20 casos por cada variable**. Hay que recordar, sin embargo, que con **30 variables**, por ejemplo, hay **435 correlaciones en el análisis factorial**.

3. Con un nivel de **significación de 0.05**, es posible que incluso **20 de estas correlaciones sean consideradas significativas y aparecerían en el análisis factorial simplemente por casualidad.**
4. Deberá siempre procurar obtener el **ratio más alto de casos por variable para minimizar las posibilidades de “sobreajustar” los datos** (es decir, derivar los factores que son específicos a la muestra con poca capacidad de generalización).

De todas formas, se emplea **una serie de variables menor** al estar guiado por consideraciones **conceptuales y prácticas**. Aun así, se tienen tamaños muestrales más pequeños y/o ratios más bajos de casos y variables, debiéndose interpretar los resultados con cautela. La cuestión del tamaño muestral será abordada también en una sección posterior sobre la interpretación de las cargas de los factores.

## 12.5. Análisis factorial: Supuestos

### Paso 3: supuestos de aplicabilidad

Los supuestos básicos subyacentes del análisis factorial son **más de tipo conceptual que estadístico**, ya que estadísticamente hablando:

1. **Se pueden obviar los supuestos de normalidad, homocedasticidad y linealidad siendo conscientes de que su incumplimiento produce una disminución en las correlaciones observadas.** En realidad, **sólo es necesaria la normalidad** cuando se aplica una prueba estadística a la **significación de los factores**; sin embargo, **raramente se utilizan estas pruebas. De hecho, es deseable que haya cierto grado de multicolinealidad**, dado que el objetivo es identificar series de variables interrelacionadas.
2. Adicionalmente a las bases estadísticas para las **correlaciones de la matriz de los datos**, debe asegurarse también de que la matriz **tiene suficientes correlaciones para justificar la aplicación del análisis factorial.**
3. **Si la inspección visual revela que no hay un número sustancial de correlaciones > 0.30, entonces el análisis factorial es probablemente inapropiado.**
4. Las correlaciones entre las variables también pueden ser analizadas con el **cálculo de las correlaciones parciales entre las variables**; esto es, las correlaciones entre variables cuando se tienen en cuenta los efectos de las otras variables. Si los **factores “verdaderos”** existen en los datos, entonces:

**-Si la correlación parcial es pequeña**, entonces se puede explicar la variable mediante los factores (valores teóricos con cargas para cada variable).

**-Si la correlación parcial es alta**, entonces **no existen factores subyacentes “verdaderos”**, y el **análisis factorial es inapropiado. SPSS proporciona la matriz de correlación anti-imagen**, que es el **valor negativo de la correlación parcial.**

En cada caso, las **correlaciones parciales o anti-imagen** mayores son indicativas de una matriz de datos que **no es quizá adecuada para el análisis factorial.**

Otra manera de determinar la conveniencia del análisis factorial es:

1. **Examinar la matriz de correlación entera.** Utilice el **contraste de esfericidad de Bartlett**, como prueba estadística para la **presencia de correlaciones entre las variables**, como una de estas medidas. Proporciona la **probabilidad**

**estadística de que la matriz de correlación de las variables sea una matriz identidad.** Debe tener en cuenta, sin embargo, que **el incremento del tamaño muestra** da lugar a que la prueba de **contraste de Bartlett** sea **más sensible a la detección de correlaciones** entre las variables.

2. Otra medida para cuantificar el **grado de intercorrelaciones** entre las variables y la conveniencia del análisis factorial es la **medida de suficiencia de muestreo (MSA)**. Este índice se extiende de **0 a 1**, con las siguientes directrices:

**1** cuando cada variable es perfectamente predicha sin error por las otras variables.

**>=0.80** sobresaliente;

**>=0.70** regular;

**>=0.60** mediocre;

**>=0.50** despreciable y,

**< 0.50**, inaceptable [Kaiser, 1970 y 1974]

El **MSA** aumenta conforme:

-Aumenta el tamaño muestra

-Aumentan las correlaciones medias

-Aumenta el número de variables o

-Desciende el número de factores [Kaiser, 1974].

Las mismas directrices de **MSA** pueden extenderse también a las variables individuales. Usted deberá examinar primero los valores **MSA** para cada variable y excluir aquellos que caen en la gama de inaceptables. Una vez que las variables individuales logran un nivel aceptable, se puede valorar el **MSA** general y se puede tomar una decisión sobre la continuación del análisis factorial.

Los **supuestos conceptuales** que subyacen en el análisis factorial se relacionan con la serie de **variables seleccionadas y la muestra elegida**. Un supuesto básico del análisis factorial es que existe una **estructura subyacente** en la serie de variables seleccionadas. Es su responsabilidad asegurarse de que **las pautas observadas sean válidas y conceptualmente apropiadas** para utilizar el análisis factorial porque **la técnica no tiene medios de determinar la conveniencia, aparte de las correlaciones entre las variables. Por ejemplo**, la mezcla de **variables dependientes e independientes** en un solo análisis factorial y posteriormente el uso de los factores derivados para apoyar las relaciones de dependencia **es inapropiado**.

Usted debe asegurarse de que **la muestra es homogénea** con respecto a **la estructura de factor subyacente. Por ejemplo**, la aplicación del análisis factorial resultaría **inapropiada** para una muestra de hombres y mujeres que tienen distintas opiniones sobre una serie de aspectos según el sexo. Cuando se combinan las **dos submuestras (hombres y mujeres)**, las correlaciones resultantes y la estructura de factores serán una **representación pobre** de la estructura única de cada grupo. Por tanto, **cuando se esperan grupos diferentes en la muestra, se deben practicar análisis factoriales separados** y los resultados deben ser comparados para identificar las diferencias no reflejadas en los resultados de la muestra combinada.

## 12.6. Análisis factorial: Estimación y Ajuste

### Paso 4: estimación y ajuste

Una vez especificadas las variables y la matriz de correlación preparada, deberá estar listo para **aplicar el análisis factorial que identifique la estructura subyacente de las relaciones** (ver **Figura 12.2**). Para realizarlo, es necesario tomar decisiones con relación a:

1. El **método de extracción de los factores (análisis factorial común vs análisis de componentes principales)** La selección del método de extracción depende del objetivo del investigador. Se utiliza el análisis de componentes principales cuando el objetivo es resumir la mayoría de la información original (varianza) en una cantidad mínima de factores con propósitos de predicción. Por el contrario, se utiliza el análisis factorial común para identificar los factores subyacentes o las dimensiones que reflejan qué es lo que las variables comparten en común.
2. El número de **factores seleccionados para representar la estructura subyacente** en los datos. Para cualquiera de estos métodos, tiene que determinar también el número de factores que representan la serie de variables originales. Tanto las **cuestiones conceptuales como empíricas afectan a esta decisión**.

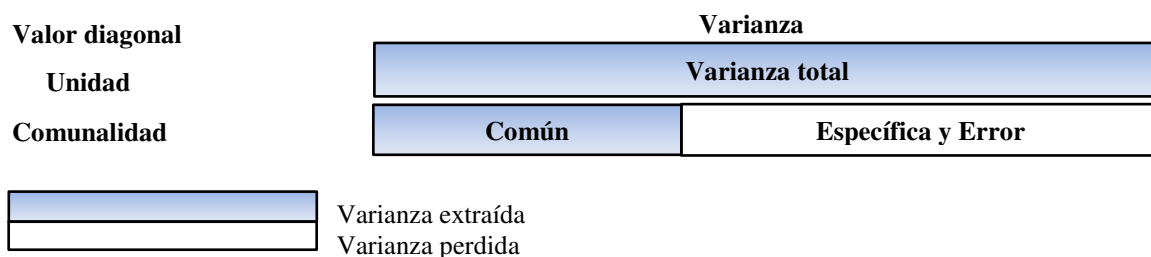
#### 12.6.1. Análisis factorial común vs. Análisis de componentes principales

Usted puede utilizar dos modelos básicos para obtener soluciones factoriales. Éstos se conocen como análisis factorial común y análisis de componentes principales. Con el fin de seleccionar el modelo apropiado, en primer lugar **debe comprender las diferencias entre los tipos de varianza**. Para los propósitos del análisis factorial, existen tres tipos de varianza total:

1. **Común**
2. **Específica (también conocida como única), y**
3. **Error.**

Estos tipos de varianza y su relación con el proceso de selección de modelo factorial se ilustran en **la Figura 12.4**.

**Figura 12.4. Tipos de varianza llevados a la matriz factorial**



Fuente: propia

Así, se tienen las siguientes definiciones:

1. **Varianza común** en una variable que se comparte con todas las otras variables en el análisis.
2. **Varianza específica** es aquella asociada solamente con una variable específica.

3. **La varianza de error** es aquella que se debe a la poca fiabilidad en el proceso de recolección de datos, al error de medición o un componente aleatorio en el fenómeno medido.

El **análisis de componentes principales** considera la **varianza total** y estima los factores que contienen **proporciones bajas de la varianza única** y, en algunos casos, la **varianza de error**. No obstante, **los primeros factores no contienen la suficiente varianza única o de error** como para distorsionar la estructura de factores en su conjunto. Específicamente, con el análisis de componentes principales, **se insertan las unidades en la diagonal de la matriz de correlación**, para que se traiga **la varianza completa en la matriz de factores**, tal y como se indica en la **Figura 12.4**.

En el **análisis factorial común**, **por el contrario**, se incorporan las varianzas compartidas en la diagonal. **Las comunales** son estimaciones de la varianza compartida o común entre las variables. Los factores que resultan del **análisis factorial común** se basan solamente en la **varianza común**.

La selección de un modelo u otro se basa en dos criterios:

1. Los **objetivos del análisis factorial**, y
2. En el **grado de conocimiento anterior acerca de la varianza** en las variables.

El **análisis de componentes principales** es apropiado cuando:

- El interés primordial se centra en **la predicción** o el **mínimo número de factores** necesarios para justificar la **porción máxima de la varianza** representada en la serie de variables original,
- Cuando el conocimiento previo sugiere que la **varianza específica y de error** representan una proporción relativamente pequeña de la varianza total.

El modelo de **análisis factorial común**, es apropiado cuando:

- Por el contrario, el **objetivo principal es identificar las dimensiones latentes** o las construcciones representadas en las variables originales y **tiene poco conocimiento acerca de la varianza específica y de error y por tanto quiere eliminar esta varianza**
- Con unos supuestos más restrictivos y la utilización exclusiva de las dimensiones latentes (**varianza compartida**), se basa más en la **teoría**.

Aunque teóricamente válido, no obstante, el **análisis factorial común** tiene varios problemas:

1. Adolece de **indeterminación de factores**. Esto significa que para cualquier encuestado individual, se pueden calcular varias puntuaciones de factores diferentes a partir de los resultados del modelo factorial [Mulaik, y McDonald 1978].
2. **No existe una sola solución única**, tal y como ocurre con el **análisis de componentes principales**; no obstante, y en la mayor parte de los casos, **las diferencias no son sustanciales**.
3. El cálculo de las **varianzas compartidas** no siempre se pueden estimar o pueden no ser válidas (es decir, **valores mayores que 1 o menores que 0**), lo que requiere la **supresión de la variable del análisis**.

Las complicaciones del análisis factorial común han contribuido al **uso generalizado del análisis de componentes principales**. Aunque todavía los expertos siguen

discutiendo acerca de cuál de los modelos factoriales es el más apropiado [Borgatta, et al. 1968, Gorsuch, 1990, Mulaik, 1990, Snook y Gorsuch, 1989], la investigación empírica ha mostrado resultados similares en muchos casos [Velicer y Jackson, 1990]. En la mayoría de las aplicaciones, ambos análisis llegan a **resultados esencialmente idénticos si el número de variables excede de 30** [Gorsuch, 1983], o **las varianzas compartidas exceden de 0.60 para la mayoría de las variables**. Si el investigador está preocupado por los supuestos del análisis de componentes principales, entonces debe **aplicar también el análisis factorial común** para valorar su estructura de representación. Cuando se llega a una decisión acerca del modelo factorial, debe estar preparado para **extraer los factores sin rotación iniciales**. Con el examen de la **matriz de factores sin rotación**, el investigador puede explorar las posibilidades de **reducción de datos** para una serie de variables y obtener una estimación preliminar de los factores a extraer. Así, se debe esperar para determinar el número de factores **final hasta que se haga una rotación** de los resultados y se interpreten los factores

### 12.6.2. Criterios para el cálculo del número de factores a ser extraídos

¿Cómo decidimos el número de factores que se deben extraer? Cuando una gran serie de variables se somete a la extracción de factores:

1. En primer lugar el método **extrae las combinaciones de las variables que explican la cantidad mayor de la varianza** y después continúa con combinaciones que justifican cantidades de varianza cada vez menores.
2. Para decidir cuántos factores se deben extraer, Usted empieza generalmente con algún criterio predeterminado, **tal como el porcentaje de varianza** o el **criterio de raíz latente**, para llegar a un número de factores específico
3. Después de estimar la solución inicial, se calculan **varias soluciones de prueba adicionales** -normalmente **un factor menos que el número inicial y dos o tres factores más que los que se estimaron inicialmente**.
4. Posteriormente, en función de la información que se obtiene de estos análisis previos, se **examinan las matrices de factores** y se **escoge el número de factores que represente mejor los datos**.
5. Por analogía, la elección del número de factores **es como enfocar un microscopio**. Un ajuste demasiado alto o demasiado bajo hará más oscura una estructura que es obvia cuando el ajuste es acertado.
6. Al examinar un número de estructuras factoriales diferentes que se derivan de varias soluciones de pruebas, Usted puede comparar y contrastar para llegar a la mejor representación de los datos.
7. Hasta aquí, se puede decir que todavía no se ha desarrollado una base cuantitativa exacta para decidir el número de factores a extraer.

### 12.6.3. Descripción de los criterios para el cálculo del número de factores a ser extraídos

**1. Criterio de raíz latente.** Es la más utilizada, por su sencillez de aplicación. La racionalidad que se usa para el criterio de raíz latente es **que cualquier factor individual debería justificar la varianza de por lo menos una única variable**. Cada variable contribuye con un valor de **1** para el autovalor total. Por tanto, **sólo se**



**consideran los factores que tienen raíces latentes o autovalores mayores que 1;** explican al menos una variable, se considera que todos los factores con **raíces latentes menores que 1 (explican menos de una variable) no son significativas** y por tanto, se desestiman a la hora de incorporarlos a la interpretación.

El uso del **autovalor** para establecer un corte es más fiable cuando el **número de variables está entre 20 y 50.**

Si el número de variables es **menor que 20**, hay una tendencia a que este método **extraiga un número de factores conservador** (demasiado poco); por el contrario, **si hay más de 50 variables, lo más común es que se extraigan demasiados factores.**

**2. Criterio a priori.** Es un criterio simple y a la vez razonable bajo ciertas circunstancias. Con su aplicación, **Usted ya sabe cuántos factores hay que extraer antes de iniciar el análisis factorial.** Usted simplemente instruye al computador para **parar el análisis cuando se haya extraído el número de factores deseado.** Esta aproximación resulta de utilidad cuando:

- Se prueba una teoría o una hipótesis acerca del número de factores para ser extraído.
- También puede ser justificado con el objetivo de replicar el trabajo de otro investigador y extraer el mismo número de factores que se encontraron previamente.

**3. Criterio de porcentaje de la varianza.** Es una aproximación que se basa en obtener un porcentaje acumulado especificado de la **varianza total extraída.** El propósito es asegurar una **significación** práctica de los factores derivados, asegurando que explican por lo menos una cantidad especificada de la varianza. No existe al momento, un umbral absoluto para todas las aplicaciones. Sin embargo:

- En las **ciencias naturales**, el procedimiento de factores **normalmente no debería ser detenido hasta que los factores extraídos cuenten por lo menos con un 95 % de la varianza** o hasta que el factor justifique solamente una porción pequeña (**menos del 5 por ciento**).

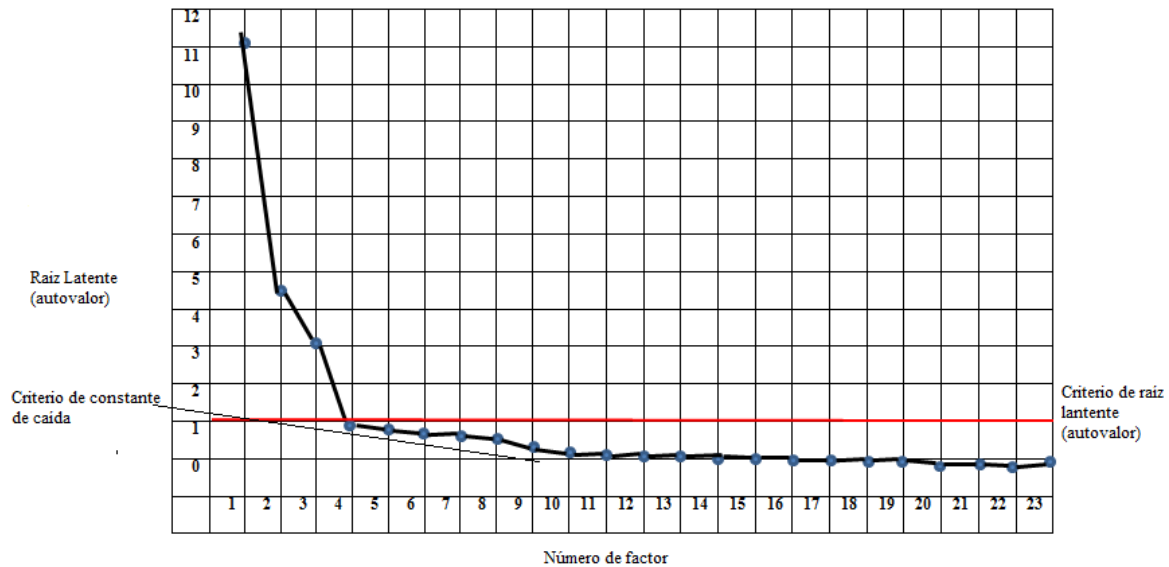
- En las **ciencias sociales**, donde la información muchas veces es menos precisa, **es normal considerar una solución que represente un 60 % de la varianza total (y en algunos casos incluso menos) como satisfactoria.**

Una variante de este criterio implica la selección de suficientes factores para **obtener una comunalidad** para cada una de las variables. Si las razones teóricas o prácticas requieren una cierta varianza compartida para cada variable, entonces la investigación **incluira tantos factores** como sean necesarios para representar de forma adecuada cada una de las variables originales

**4. Criterio de contraste de caída** Recuerde que con el modelo de **análisis de componentes principales**, los factores posteriores que han sido extraídos contienen tanto la **varianza común como la varianza única.** Aunque todos los factores contienen por lo menos alguna **varianza única**, la proporción de la varianza única es sustancialmente **más alta en los factores posteriores que en los primeros.** El contraste de caída se utiliza para **identificar el número óptimo de factores que pueden ser extraídos antes de que la cantidad de la varianza única empiece a dominar la estructura de la varianza común** [Cattell, 1966]. Se estima el contraste de caída con el trazo de **raíces latentes** en función del número de factores en su orden de extracción, y se utiliza la forma de la **curva consiguiente** para evaluar el punto de corte.

La **Figura 12.5** representa los primeros 18 factores extraídos de una investigación realizada.

**Figura 12.5. Gráfico de autovalor para el criterio de contraste de caída.**



Fuente: SPSS 20 IBM

Si empezamos con el primer factor, el trazo tiene inicialmente una inclinación descendente y a continuación se convierte paulatinamente en una línea más o menos horizontal. Se considera que el punto en que **la curva empieza a rectificar se indica el máximo número de factores a extraer**. En el caso que nos ocupa, se incluirán los primeros **10** factores. Por encima de **10**, se incluiría una **proporción de la varianza única demasiado grande, por lo que estos factores no son aceptables**. Es importante señalar que :

- Con el **uso del criterio de raíz latente**, solamente se tienen en cuenta **ocho factores**.
- Con el **uso del criterio de contraste de caída** nos proporciona **dos factores más**. Por regla general, el contraste de caída normalmente tiene como resultado que se incluyan uno y a veces dos o más factores adicionales que cuando se utiliza el **criterio de raíz latente [Cattell, 1966]**.

**5. Heterogeneidad de la muestra.** La existencia de varianza compartida entre las variables es el núcleo tanto de los modelos de **factores comunes como de los de componentes. Un supuesto subyacente es que la varianza compartida se extiende a lo largo de toda la muestra**. Si la muestra es heterogénea al menos con respecto a un subconjunto de variables, los primeros factores representarán aquellas variables que son más homogéneas a lo largo de toda la muestra. Las variables con mayor capacidad de discriminar entre subconjuntos muestrales cargarán sobre los últimos factores, en muchas ocasiones aquellos no seleccionados de acuerdo a los criterios comentados más arriba [Dillon et al.1989]. Cuando el objetivo sea identificar factores que discriminen entre subconjuntos muestrales, Usted deberá extraer factores adicionales entre aquellos señalados por los métodos anteriormente expuestos y examinar la capacidad de los factores adicionales para discriminar entre

grupos. Si resultan ser peores al discriminar, la solución puede estar en proceder de nuevo y eliminar estos últimos factores.

**6. Resumen de los criterios de selección de factores.** En la práctica, rara vez se utiliza un único criterio al determinar cuántos factores extraer. En su lugar, inicialmente se emplea un criterio como el de la **raíz latente como orientación** en un primer intento de interpretación. Después de haber interpretado los factores, como se expone en la siguiente sección, **se valora su carácter práctico**. También se interpretan los factores identificados mediante otros criterios. Elegir el número de factores **está interrelacionado con la valoración de la estructura**, lo que se revela en la etapa de **interpretación**. De esta forma, **se examinan varias soluciones factoriales con diferentes números de factores antes de que la estructura esté bien definida**.

**Nota a la selección del conjunto definitivo de factores:** puede resultar **inconveniente seleccionar tanto muchos como pocos factores** para representar los datos:

-**Pocos factores**, no se revela la estructura correcta y pueden omitirse importantes | dimensiones.

-**Demasiados factores**, las interpretaciones resultan más difíciles cuando se rotan los resultados.

Tal y como ocurre con otros aspectos de los modelos multivariantes, es importante la **parsimonia**. Una **excepción** a destacar es cuando el análisis de factores se emplea en exclusiva para la reducción de datos y se especifica la extracción de un nivel conjunto de varianza. Usted deberá siempre esforzarse en **conseguir el conjunto de factores lo más representativo y parsimonioso posible**.

## 12.7. Análisis factorial: Interpretación de factores

### Paso 5: Interpretación

Para interpretar los factores y seleccionar la solución factorial definitiva se debe:

1. **Calcular la matriz inicial de factores no rotados** para obtener una indicación preliminar acerca del número de factores a extraer. La matriz de factores contiene las cargas factoriales para cada variable sobre cada factor. Al calcularla Usted simplemente **deberá determinar la mejor combinación lineal de variables**, es decir, **encontrar aquella combinación particular de las variables originales que cuenta con el mayor porcentaje de varianza de los datos**. En consecuencia, **el primer factor** puede contemplarse como el mejor resumen de las relaciones lineales que los datos manifiestan. **El segundo factor** se define como la segunda mejor combinación lineal de las variables, sujeta a **la restricción de que sea ortogonal al primer factor**. Para ser ortogonal al primer factor, el segundo factor **debe derivarse de la varianza restante tras la extracción del primer factor**. Así, **el segundo factor** puede definirse como la combinación lineal de las variables que da cuenta del mayor porcentaje de varianza residual una vez se ha eliminado de los datos el efecto del primer factor. Los factores subsiguientes se definen de forma **análoga hasta haber agotado la varianza de los datos**. **Las soluciones factoriales no rotadas** alcanzan el objetivo de reducción de los datos, pero Usted debe preguntarse si la solución factorial no rotada (que se adecua a los requisitos

matemáticos deseables) facilita una información que ofrezca la interpretación más adecuada de las variables examinadas. **La mayor parte de las veces no resulta ser así.** Debe recordar que la **carga factorial**:

-Es el medio para interpretar la función que cada variable desempeña al definir cada factor.

-Es la correlación o grado de correspondencia entre cada variable y el factor, haciendo a una variable con mayor carga representativa del factor.

La **solución factorial no rotada** puede no suministrar un patrón significativo de cargas de las variables. Si se espera que los factores no rotados sean significativos, **puede especificar que la rotación no se lleve a cabo.**

**La rotación es deseable** porque simplifica la estructura de los factores, ya que es difícil determinar si los factores no rotados serán significativos.

2. **Hacer uso de un método de rotación** para lograr soluciones factoriales más simples y teóricamente más significativas. En muchos casos la rotación de los factores mejora la interpretación disminuyendo alguna de las ambigüedades que a menudo acompañan a las soluciones factoriales inicialmente no rotadas.
3. **Valorar** la necesidad de **especificar de nuevo el modelo de factores** debido a:
  - La eliminación de variables en el análisis,
  - El deseo de emplear un método de rotación diferente para la interpretación,
  - La necesidad de extraer un número diferente de factores, o
  - El deseo de cambiar de un método de extracción a otro. La especificación nueva del modelo factorial viene acompañada de la vuelta a la etapa de extracción, rotación de factores y de nuevo su interpretación.

### 12.7.1. Rotación de factores

Es una herramienta importante para la interpretación de los factores. Así, **se giran en el origen los ejes de referencia de los factores hasta alcanzar una determinada posición.** Como se indicó previamente, **las soluciones factoriales no rotadas extraen factores según su orden de importancia:**

-**El primer factor** tiende a ser un factor general por el que casi toda variable se ve afectada significativamente dando cuenta del mayor **porcentaje de varianza.**

-**El segundo y siguientes factores** se basan en la **varianza residual.** Cada uno explica porcentajes de varianza cada vez menores.

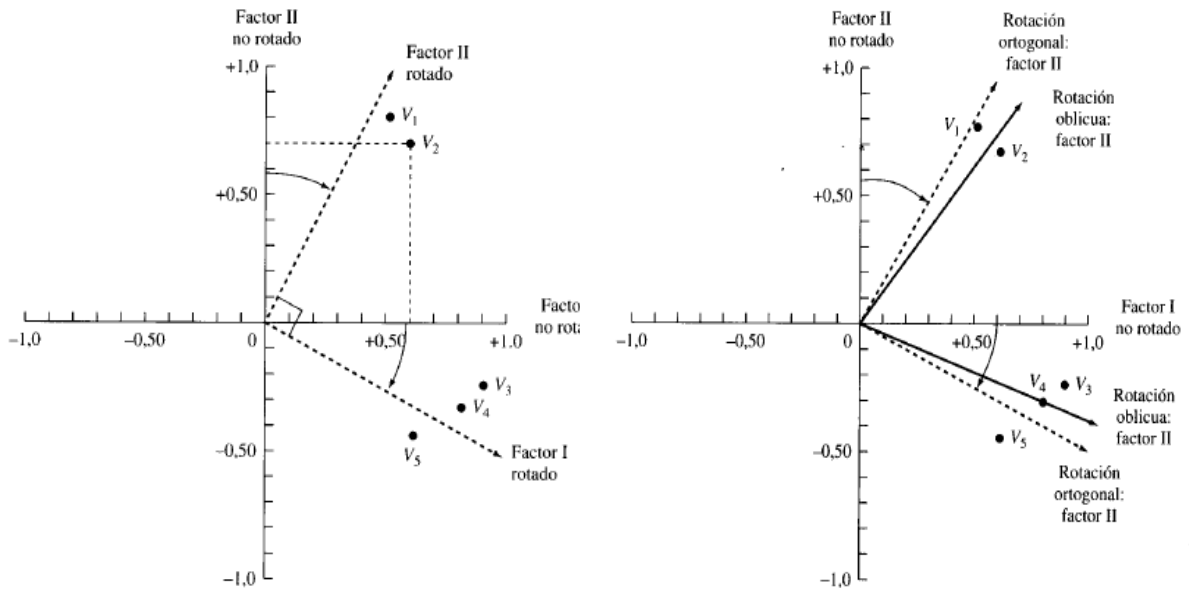
-**El efecto último de rotar la matriz de factores** es redistribuir la varianza de los primeros factores a los últimos para lograr un patrón de factores más simple y teóricamente más significativo.

-**El caso más simple de rotación es la rotación ortogonal,** en la que los ejes se mantienen formando un ángulo de **90 grados.**

-También es posible rotar los ejes y **no mantener el ángulo de 90 grados** entre los ejes de referencia.

-Cuando no se limita a ser ortogonal, **la rotación se denomina oblicua.** Las rotaciones de factores ortogonal y oblicua están ilustradas en las **Figuras 12.6,** respectivamente.

**Figura 12.6. Rotación factorial ortogonal y Rotación factorial oblicua**



Fuente: propia

De la **Figura 12.6** de rotación factorial ortogonal se han representado cinco variables en un **diagrama de factores bidimensional**, con la siguiente descripción:

1. **El eje vertical** representa el **factor no rotado II**, y el **horizontal el factor no rotado I**.
2. El **0** indica el origen de coordenadas yendo éstas de **-1.0 a +1.0**.
3. **El número sobre los ejes representa las cargas factoriales**.
4. Las **5 variables** están denominadas como **V1, V2, V3, V4 y V5**.
5. La carga factorial para la **V2** sobre el **factor no rotado II** está indicada horizontalmente mediante una línea discontinua del punto de los datos al eje vertical del **factor II**.
6. Análogamente se dibuja una línea vertical de la **V2** al eje horizontal del **factor no rotado I** para determinar la carga de la **V2** sobre el **factor I**.
7. Un procedimiento similar para las variables restantes determina las cargas factoriales para las soluciones **no rotadas y rotadas**, como se muestra en la **Tabla X**.
8. Sobre el **primer factor no rotado**, todas las variables cargan bastante alto.
9. Sobre el **segundo factor no rotado**, **V1** y **V2** cargan muy alto en el lado positivo. La **V5** tiene una carga moderadamente alta en el lado negativo, y las **V3** y **V4** tienen cargas considerablemente inferiores en el lado negativo.
10. Como resultado de la inspección de la **Figura 12.7. Rotación factorial ortogonal** se determina que hay dos grupos de variables. **V1** y **V2** van juntas, así como las **V3, V4 y V5**.

**Figura 12.7. Comparación entre cargas factoriales rotadas y no rotadas**

	Cargas factoriales no rotadas		Cargas factoriales rotadas	
	I	II	I	II
<b>V1</b>	0.5	0.8	0.03	0.94
<b>V2</b>	0.6	0.7	0.16	0.90
<b>V3</b>	0.9	-0.25	0.95	0.24
<b>V4</b>	0.8	-0.30	0.84	0.15
<b>V5</b>	0.6	-0.50	0.76	-0.13

Fuente: propia

Sin embargo, este patrón de variables no es tan obvio a partir de las **cargas de factores no rotados**. Rotando los ejes originales en el sentido de las agujas del reloj, como se indica en la **Figura 12.6. Rotación factorial ortogonal**, obtenemos un **patrón de carga factorial completamente distinto**. Nótese que al rotar los factores, los ejes mantienen el ángulo de **90 grados**. Este procedimiento implica que **los factores son matemáticamente independientes y que la rotación ha sido ortogonal**. Después de rotar el eje de factores, **X3 y X4** cargan muy poco sobre el **factor I**, y **X<sub>1</sub> y X<sub>2</sub>** cargan mucho sobre el **factor II**. Así, el patrón o agrupamiento de estas variables en **dos grupos** resulta más obvio que antes de la rotación, incluso la posición relativa o con figuración de las variables permanece inalterada.

Los mismos principios generales de las rotaciones ortogonales atañen a las **oblicuas**. El método de la **rotación oblicua es más flexible porque los ejes de factores no necesitan ser ortogonales**. También es más realista porque las dimensiones subyacentes teóricamente más importantes, se su ponen relacionadas entre sí.

En la **Figura 12.6 Rotación factorial oblicua** se comparan los dos métodos de rotación. Observe que la **rotación de factores oblicua** representa el agrupamiento de variables **con más precisión**, la cual se deriva del hecho de que cada eje de factores rotado está ahora más cerca del grupo respectivo de variables. Además, la **solución oblicua** provee de información sobre la medida en que los factores realmente están correlacionados uno con otro. La mayor parte de los investigadores están de acuerdo en que **soluciones factoriales no rotadas**, aunque más directas, **no resultan suficientes**; es decir, en muchos casos la rotación mejorará la interpretación paliando alguna de las ambigüedades que a menudo acompañan al análisis preliminar. **Las alternativas disponibles son la rotación ortogonal o la oblicua**.

El objetivo último de toda rotación es obtener algunos **factores teóricamente significativos** y, si es posible, **la estructura de factores más simple**.

La **rotación ortogonal** se emplea con más frecuencia dada su presencia en todos los programas informáticos de análisis factorial, mientras que **los métodos oblicuos no están tan extendidos**. Además, las **rotaciones ortogonales** se utilizan con más frecuencia porque los **procedimientos analíticos** para llevar a cabo **rotaciones oblicuas no están totalmente desarrollados** y están todavía **sujetos a controversia**. Existen varias aproximaciones distintas para llevar a cabo **rotaciones ortogonales u oblicuas**. Sin embargo, sólo un número escaso de procedimientos de

**rotación oblicua** está disponible en la mayoría de los programas estadísticos; por eso el investigador tendrá que aceptar probablemente alguno de los provistos.

### 12.7.2. Métodos de rotación ortogonal

En la práctica, **el objetivo de todos los métodos de rotación es simplificar las filas y columnas de la matriz de factores para facilitar la interpretación.** En una matriz de factores, las columnas representan los factores, con cada fila correspondiendo a las cargas de las variables para cada uno de los factores. Simplificando las filas, queremos **aproximar lo más posible a cero** tantos valores como sea posible (es decir, **maximizar la carga de una variable sobre un único factor**). Simplificando las columnas, queremos **aproximar lo más posible a cero** tantos valores como sea posible (es decir, haciendo que el número de **cargas “altas” sea el menor posible**). Se han desarrollado principalmente **3 aproximaciones**:

1. **QUARTIMAX.** Su objetivo es simplificar las filas de una matriz de factores; esto es, se centra en **rotar los factores iniciales de tal forma que una variable cargue alto sobre un factor y tan bajo como sea posible sobre los otros factores.** En estas rotaciones muchas variables pueden cargar alto o cerca sobre el mismo factor porque la técnica se centra en las filas. El método no ha demostrado gran capacidad para generar estructuras más simples. Su dificultad está en que tiende a producir un factor general, como el primer factor, sobre el que la mayor parte, si no todas las variables, tiene cargas mayores. Con independencia del concepto que cada cual tenga de estructuras **“más simples”**, inevitablemente se ha de tratar con agrupaciones de variables; un método que tiende a producir un factor general grande (por ejemplo, el **QUARTIMAX**) no responde a los objetivos de la rotación.
2. **VARIMAX.** En contraste al anterior, se centra en simplificar las columnas de la matriz de factores. Con **VARIMAX** se alcanza la máxima simplificación posible **si sólo hay ceros y unos en una columna.** Esto es, el método maximiza la suma de las varianzas de las cargas requeridas de la matriz de factores. Recuerde que en la aproximación **QUARTIMAX**, muchas variables pueden cargar alto o cerca de lo alto sobre el mismo factor porque la **técnica se centra en simplificar las filas.** Con **VARIMAX**, tiende a haber **altas cargas factoriales** (esto es, **cercanas a -1 o +1**) y algunas cargas cerca de **0** en cada columna de la matriz. Si la lógica está en que la interpretación es más fácil cuando las correlaciones **variable factor** están:  
**-Cercanas a -1 o +1**, indicando así una clara asociación positiva o negativa entre la variable y el factor; o  
**-Cercanas a 0** señalando una clara ausencia de asociación.

Esta estructura resulta esencialmente sencilla. Aunque la solución **QUARTIMAX** es analíticamente más simple que la solución **VARIMAX**, ésta parece suministrar una separación más clara de factores. En general, el experimento de Kaiser (1974) indica que el patrón factorial obtenido mediante la rotación **VARIMAX** tiende a resultar más robusto que el obtenido por el método **QUARTIMAX** cuando se analizan diferentes subconjuntos de variables. El método **VARIMAX** ha demostrado tener más éxito como aproximación analítica para lograr una rotación ortogonal de factores.

3. **EOUIMAX.** Se considera intermedio entre las aproximaciones **QUARTIMAX** y **VARIMAX**. En lugar de concentrarse bien en la simplificación de las filas, bien de



las columnas, **procura cumplir con las dos. EQUIMAX** no ha logrado una amplia aceptación y se emplea en muy raras ocasiones.

### 12.7.3. Métodos de rotación oblicua

Las rotaciones oblicuas **son similares a las rotaciones ortogonales**, excepto **permiten la existencia de factores correlacionadas en lugar de mantener la independencia entre los factores rotados**. Aunque en la mayor parte de los programas estadísticos hay varias alternativas de **aproximación ortogonal**, suele haber escasas de rotaciones oblicuas. Por ejemplo, SPSS cuenta con **OBLIMIN**; SAS con **PROMAX** y **ORTOBLIQUE**; y BMDP con **DQUART**, **DOBLIMIN** y **ORTOBUQUE**.

Los objetivos de simplificación son comparables a los de los métodos ortogonales, con el rasgo añadido de existencia de factores correlacionados. Con esta posibilidad, el investigador ha de tener un cuidado adicional al validar los factores rotados oblicuamente, puesto que cuenta con una forma adicional (**no ortogonalidad**) de proceder, específica a la muestra y no generalizable, especialmente en muestras pequeñas o de bajos ratios **casos/variable**.

### 12.7.4. Selección del método de rotación

**No se han desarrollado reglas concretas** que guíen al investigador en la selección de una técnica de rotación particular ortogonal u oblicua. En la mayoría de las ocasiones, el simplemente utiliza la técnica rotacional suministrada por el programa de computador. Muchos programas cuentan por defecto con la rotación **VARIMAX**, pero también resultan fácilmente accesibles los métodos rotacionales más comunes. Sin embargo, **no existe una razón analítica incuestionable a favor de un método de rotación u otro**. La elección de una rotación **ortogonal u oblicua** debería hacerse según las necesidades concretas de un problema de investigación determinado. Así, de acuerdo al objetivo, se tiene:

1. Si el objetivo del investigador es **reducir el número de variables originales**, con independencia de la significación resultante de los factores ó desea reducir un gran número de variables **a un conjunto pequeño de variables incorrelacionadas** para un uso posterior en el análisis de regresión u otras técnicas de predicción, **la solución ortogonal** resulta la más adecuada.
2. Sin embargo, si el objetivo último del análisis factorial es **obtener varios factores teóricamente significativos**, resulta apropiada una **solución oblicua**. Llegamos a esta conclusión dado que, realmente, muy pocos factores están incorrelacionados, como ocurre con la rotación ortogonal.

### 12.7.5. Criterios para la significación de la carga factorial

Al interpretar los factores, debe decidir el criterio en tomo a qué cargas factoriales merece la pena considerar, por lo que se expone a su consideración diversos aspectos relativos a la **significación práctica y estadística**, además de al número de variables, que afectan a la interpretación de las cargas factoriales, como sigue:

1. **Asegurar la significación práctica**. Esta consideración, no está basada en afirmaciones matemáticas, sino que tiene que ver más con la **significación práctica**. Consiste en un tipo de regla empírica empleado frecuentemente como

forma de realizar un **examen preliminar de la matriz de factores, donde las cargas factoriales, se consideran:**

**> ±:0.30 nivel mínimo;**

**±:0,40 más importantes;**

**>=±:0,50 significativas.** Así, cuanto mayor sea el tamaño absoluto de la carga factorial, más importante resulta la carga al interpretar la matriz factorial. Dado que la carga factorial es la correlación entre la variable y el factor, el cuadrado de la carga es la cuantía de la varianza total de la variable de la que da cuenta el factor. Así, una carga de **0.30** implica aproximadamente una explicación de un **10 %**; una carga de **0.50** quiere decir que un **25 %** de la varianza es debida al factor. Para que un factor explique un **50%** de la varianza ha de contar con una carga que supere el **70 %**. Usted deberá darse cuenta de que cargas extremadamente elevadas (**>0.80** o más) **no son normales** y que la significación práctica es un criterio importante. Estas orientaciones son de aplicación cuando el tamaño muestra supera las **100 observaciones**.

2. **Valoración de la significación estadística.** Como se indicó previamente, la carga factorial representa la correlación entre la variable original y su factor. Al determinar el **nivel de significación** para la interpretación de las cargas, se debería emplear una aproximación similar a la utilizada para la significación estadística de los **coeficientes de correlación**. Sin embargo, diversas investigaciones [Cliff y Hamburger,1967] han demostrado que **las cargas factoriales cuentan con errores estándar sustancialmente mayores que las correlaciones habituales**, por lo que las cargas factoriales deberían evaluarse con niveles considerablemente más estrictos. Usted debe utilizar el **concepto de potencia estadística** expuesto en el **Capítulo 2** para especificar cargas factoriales consideradas **significativas** según diferentes **tamaños muestrales**. Con el objetivo establecido en lograr un **nivel de potencia del 80%**, el uso de un nivel de significación de un **0.05** y la inflación probada de los **errores estándar de las cargas factoriales**, la **Figura 12.8** contiene los **ta-maños muestrales** necesarios para que cada valor de la **carga factorial** se considere significativa.

Por ejemplo, en una muestra de **100 observaciones**, las cargas factoriales de **0.55** o más son significativas. Sin embargo, en una muestra de **50**, la significación implica una carga factorial de **0.75**. En comparación con la anterior regla empírica que implicaba la significación para cargas del **0.30**, esta aproximación consideraría a una carga de **0.30** significativa si el tamaño muestra fuera de **350** observaciones. Existen varias orientaciones prudentes cuando se comparan con las de la sección previa o incluso con errores estándar asociados a los coeficientes de correlación convencionales. Por ello, estas orientaciones deberían emplearse como punto de partida en la interpretación de las cargas factoriales, considerando significativas cargas factoriales bajas y de forma añadida a la interpretación basada en otras consideraciones. La siguiente sección detalla el proceso de interpretación y la función que pueden desempeñar otras consideraciones.

**Figura 12.8. Directrices para la identificación de cargas factoriales significativas basadas en el tamaño muestra**

Carga factorial	Tamaño muestral necesario para la significación
0.30	350
0.35	250
0.40	200
0.45	150
0.50	120
0.55	100
0.60	85
0.65	70
0.70	60
0.75	50

Nota: La significación se basa en un nivel de significación de 0,05 (ex), un nivel de potencia del 80 por ciento y los errores estándar supuesta mente dos veces mayores que los coeficientes convencionales de correlación.

Fuente: Hair et al. 1999

- Ajustes basados en el número de variables** Una **desventaja** de las aproximaciones anteriores es que **no se considera el número de variables y los factores concretos que se analizan**. Se ha comprobado que, a medida que el investigador se **mueve del primer factor a los últimos factores**, debe **incrementar el grado aceptable para considerar a una carga como significativa**. El hecho de que la **varianza única** y la **varianza del error** empiecen a aparecer en los últimos factores significa que debería **incluirse algún ajuste al alza en el nivel de significación** [Kaiser, 1970]. Al decidir qué cargas son significativas también **es importante el número de variables que se analizan**. Según el número de variables analizadas, se incrementa el nivel aceptable para considerar significativa una carga que decrece. **El ajuste por número de variables** crece en importancia según uno se mueve del primer factor extraído a los últimos. **Resumiendo** los criterios para la significación de las cargas factoriales, se pueden establecer las siguientes orientaciones:

  - A **mayor tamaño muestra, menor puede ser la carga** para ser considerada como **significativa**;
  - A **mayor número de variables analizadas, menor ha de ser la carga** para ser considerada como **significativa**;
  - A **mayor número de factores, mayor ha de ser el tamaño de la carga** de los últimos factores para considerarse como **significativa** en la interpretación.

#### 12.7.6. Interpretación de la matriz de factores

La interpretación de las relaciones complejas representadas en la matriz de factores no es una tarea fácil. Sin embargo, siguiendo el procedimiento señalado a

continuación, se puede simplificar considerablemente el procedimiento de interpretación.

1. **El examen de la matriz de cargas factoriales.** Cada columna de números en la matriz de factores representa un factor aislado. Las **columnas** de números son las **cargas factoriales** de cada **variable sobre cada factor**. Con el fin de identificar, el computador normalmente **imprime identificando los factores de izquierda a derecha por los números 1, 2, 3, 4**, etc. También las **variables** por su número de **arriba a abajo**. Para facilitar aún más la interpretación, **se sugiere escribir el nombre de cada variable en el margen izquierdo** al lado del número de cada variable. Si se ha utilizado una **rotación oblicua**, se presentan **dos matrices de cargas factoriales**:

-La primera es la matriz de **patrones factoriales**, que contiene las cargas que representan la contribución única de cada variable al factor.

-La segunda es la matriz de **estructura factorial**, que contiene las **correlaciones simples** entre variables y factores, pero estas cargas contienen tanto la **varianza única** entre variables y factores como la **correlación entre factores**. Según crece la **correlación entre factores**, es más difícil distinguir qué variables cargan únicamente sobre cada factor en la matriz de estructura factorial. Muchos investigadores suministran los resultados de la **matriz de patrones factoriales**.

2. **Identificación de la mayor carga para cada variable.** La interpretación debería comenzar con la **primera variable sobre el primer factor y moverse horizontalmente de izquierda a derecha**, observando la mayor carga para cada variable sobre cada factor. **Cuando se identifica la mayor carga (en valor absoluto), debe subrayarse si es significativa.** Entonces la atención debe centrarse en la **segunda variable**, y de nuevo moviéndose de izquierda a derecha **horizontalmente**, comprobar la mayor carga de cada variable sobre cada factor y subrayarla. Este procedimiento debe continuar para toda variable **hasta que todas las variables se hayan subrayado** una vez en la mayor carga sobre un factor. Recuérdese que para **tamaños muestrales menores a 100**, la menor carga factorial que se considere significativa debería ser en la mayor parte de las ocasiones de  **$\pm 0,30$** . El proceso de subrayar sólo la mayor carga como significativa para cada variable es un ideal que debería perseguirse **pero rara vez se consigue**. Cuando cada variable tiene sólo una carga sobre un factor que es considerado significativo, la interpretación del significado de cada factor se simplifica considerablemente. En la práctica, sin embargo, muchas variables cuentan con **varias cargas de tamaño moderado**, todas las cuales son significativas, y el trabajo de interpretar los factores es mucho más complicado. La dificultad surge porque **una variable con varias cargas significativas debe tenerse en cuenta** al interpretar (etiquetar) todos los factores sobre los cuales tiene una carga significativa. Muchas soluciones factoriales no concluyen con una solución de estructura simple (una única alta carga para cada variable sólo sobre un factor). Por eso el investigador deberá continuar, tras encontrar la mayor carga para cada variable, evaluando la matriz de factores para encontrar todas las cargas significativas para una variable sobre todos los factores. Por último, **el objetivo es minimizar** el número de cargas significativas sobre cada fila y la matriz de

factores (esto es, **hacer que cada variable se asocie sólo con un factor**). Una variable con varias cargas altas es candidata a ser eliminada.

3. **Valoración de la comunalidad** Una vez que las variables se han agrupado en sus respectivos factores, Usted debe **examinar la matriz de factores** para **identificar variables que no hayan sido incluidas** en ningún factor. La **comunalidad** representa la **proporción de varianza con la que contribuye cada variable a la solución final**. Debe observar la comunalidad de cada variable para **evaluar si alcanza niveles aceptables de explicación**. Por ejemplo, puede especificar que al menos **sea explicada la mitad de la varianza de cada variable**. Esto significa **identificar todas las variables con comunalidades menores a 0.50** como carentes de explicación suficiente. **Si hay variables que no cargan** sobre ningún factor o cuyas comunalidades se juzgan **demasiado bajas**, caben **2 alternativas**:

-**Interpretar la solución tal cual es y simplemente prescindir de esas variables**; o

-**Evaluar cada una de esas variables para su supresión eventual. Prescindir de variables**.

Puede resultar apropiado si el objetivo es únicamente la **reducción de datos**, pero el investigador todavía debe percatarse de que las variables en cuestión están pobremente representadas en la solución factorial. La consideración sobre su eliminación debe depender de la contribución conjunta de las variables para el investigador además del **índice de comunalidad**. Si la variable en cuestión es de escasa importancia para el objetivo del estudio o cuenta con una comunalidad inaceptable, podría ser eliminada y se procedería posteriormente a especificar el modelo factorial excluyendo esa variable.

4. **Etiquetación de los factores**. Cuando se ha obtenido una solución factorial en que todas las variables tienen una carga significativa sobre un factor, el investigador procura atribuir un significado al patrón de cargas factoriales. Las variables con mayores cargas se consideran más importantes y tienen mayor influencia sobre el nombre o etiqueta seleccionada para representar al factor. Por eso, el investigador examinará todas las variables agrupadas en un factor particular y, poniendo mayor énfasis en las variables con mayor carga factorial, intentará asignar un nombre o etiqueta al factor que refleje con precisión las variables cargadas sobre el factor. Los signos se interpretan como otros coeficientes de correlación. Sobre cada factor, signos iguales significan que las variables están positivamente relacionadas, signos opuestos significan que las variables están negativamente relacionadas. En soluciones ortogonales los factores son independientes unos de otros. Por tanto, los signos de las cargas factoriales se relacionan sólo con el factor en el cual aparecen, no con otros factores en la solución. Esta etiqueta no viene asignada por el análisis factorial realizado por el programa de computador; en su lugar, la etiqueta se fabrica intuitivamente de acuerdo a la conveniencia para representar: las dimensiones subyacentes de un factor concreto. El resultado final será el nombre o etiqueta que representa cada uno de los factores obtenidos con tanta precisión como sea posible. En algunas ocasiones, no es posible asignar un nombre a cada uno de los

factores. Cuando surge tal situación, el investigador desearía etiquetar un factor o factores derivados de la solución como **"indefinidos"**.

En tales casos el investigador interpreta sólo aquellos factores que son significativos y elude aquellos indefinidos o menos significativos. Al describir la solución factorial, el investigador indica que esos factores se obtuvieron pero que eran indefinidos y que sólo se interpretan aquellos factores que representan relaciones significativas.

Como se expuso anteriormente, la selección de un número concreto de factores y el método de rotación están interrelacionados. Se pueden llevar a cabo varias rotaciones adicionales de prueba y comparando la interpretación factorial para las diferentes rotaciones ensayadas, el investigador puede seleccionar el número de factores a extraer. En resumen, la habilidad para asignar algún significado a los factores, o para interpretar la naturaleza de las variables, son consideraciones extremadamente importantes para determinar el número de factores a extraer.

## **12.8. Análisis factorial: Validación**

### **Paso 6: validación**

Comprende la evaluación del **grado de generabilidad** de los resultados para la población y la influencia potencial de causas o individuos sobre los resultados globales. Este aspecto es esencial en cada uno de los métodos multivariantes, pero es especialmente relevante en los **métodos de interdependencia** por describir una estructura de datos que debería ser representativa también de la población. El método más directo de validación de los resultados consiste en adoptar una **perspectiva de confirmación**, valorando la **replicabilidad** de los resultados, bien **dividiendo la muestra con los datos originales**, bien con una **muestra adicional**. **La comparación de los resultados de dos o más modelos factoriales siempre ha sido problemática**. Sin embargo, existen varias alternativas para realizar una comparación objetiva. El auge del **análisis factorial confirmatorio (AFC)** a través de la modelización de **ecuaciones estructurales** supone una alternativa, pero **generalmente es más complicado** y requiere software adicional como **LISREL o EQS** [Bentler, 1992, Joreskog y Sorbo 1993]. Además del **AFC**, se han propuesto otros métodos que van desde un simple índice de adecuación [Cattell et al. 1969] a programas (**FMATCH**) diseñados especialmente para valorar la correspondencia entre matrices de factores [Smith, 1989]. Estos métodos cuentan con un **uso ocasional**, debido en parte a:

**-La percepción de ausencia de sofisticación y**

**-La no disponibilidad de software** o programas analíticos que automaticen las comparaciones.

Por eso, cuando el **AFC** no es apropiado, estos métodos facilitan una base objetiva para la comparación.

Otro aspecto de la **generalización es la estabilidad** de los resultados del modelo factorial, la cual **depende primeramente del tamaño muestra y del número de casos por variable**. Es usual que los investigadores siempre estén obsesionados por contar con el mayor tamaño muestra posible y desarrollar modelos parsimoniosos que **incrementen la ratio-casos-por-variable**. Si el tamaño de la muestra lo permite,

puede **dividir aleatoriamente la muestra en dos subconjuntos y estimar los modelos factoriales de cada uno**. La comparación de las **dos matrices factoriales** resultantes suministrará una valoración de la robustez de la solución a lo largo de la muestra.

Además de la **generabilidad**, otro aspecto de importancia para la validación del análisis factorial es la **detección de observaciones influyentes**. Las discusiones sobre la **detección de atípicos** y de las **observaciones influyentes en la regresión** se deben aplicar también al análisis factorial. Debe procurar **estimar el modelo con y sin observaciones** identificadas como **atípicas** para **valorar su influencia** sobre los resultados. Son de aplicación al análisis factorial varias medidas de influencia que reflejan la posición relativa de una observación respecto a las otras (por ejemplo, el **ratio de la covarianza**). Finalmente, se han propuesto métodos específicos de análisis factorial para identificar observaciones influyentes [Chatterjee et al.1991], pero su complejidad ha restringido su aplicación.

## **12.9. Análisis factorial: Usos adicionales de los resultados**

### **Paso 7:**

Dependiendo de los objetivos de partida al aplicar el análisis factorial, puede **detenerse en la interpretación de los factores o proseguir con uno de los métodos de reducción de datos**. Si el objetivo simplemente consiste en identificar combinaciones lógicas de variables y una mejor comprensión de las relaciones entre variables, **entonces basta con la interpretación de los factores**. Ésta facilita una **base empírica** para enjuiciar la estructura de las variables y la influencia de esta estructura cuando se interpretan los resultados a partir de otras técnicas multivariantes. Si el objetivo, sin embargo, es identificar variables apropiadas para aplicaciones subsiguientes de otras técnicas estadísticas, se empleará alguna forma de reducción de datos, como:

1. Examinar la matriz de factores y seleccionar la variable con mayor carga factorial como un representante de una dimensión factorial particular, o
2. Reemplazar el conjunto original de variables por uno totalmente nuevo con menos variables creado a partir de escalas aditivas o de la puntuación de factores.

**Cada alternativa suministrará nuevas variables para ser usadas**, por ejemplo, como variables **independientes** en una **regresión** o en el **análisis discriminante**, o como variables **dependientes** en el **MANOVA**, o incluso como las variables agrupadas en el **análisis cluster**.

### **12.9.1. Selección de variables suplentes para el análisis subsiguiente**

Si el objetivo del investigador es sencillamente **identificar variables** apropiadas para la aplicación subsiguiente de otras técnicas estadísticas, cuenta con la alternativa de examinar **la matriz factorial y seleccionar la variable con la mayor carga factorial** sobre cada factor para que actúe como variable suplente del factor. Este enfoque es simple y directo sólo cuando una variable tiene una carga factorial que es sustancialmente mayor que las otras cargas factoriales. En muchas ocasiones, sin embargo, el proceso de selección es mucho más difícil **porque dos o más variables tienen cargas que son significativas y bastante cercanas entre sí**. Estos casos



**requieren un examen crítico** de las cargas factoriales de tamaño aproximado y sólo una como representativa de una dimensión concreta. Esta decisión debería basarse en el conocimiento previo de la teoría por parte del investigador que pueda sugerir que una variable con preferencia a otra pueda ser representativa de la dimensión. Además, puede contar con un conocimiento que le sugiera que una variable con una carga ligeramente inferior es de hecho más fiable que la variable con la mayor carga. En tales casos, puede elegir la variable con carga ligeramente inferior como la mejor variable suplente de un factor concreto. La aproximación de seleccionar una **única variable como representativa del factor**- aunque simple y manteniendo la variable original - **cuenta con varios inconvenientes potenciales**. En primer lugar, **no orienta sobre el error de medida** que aparece cuando se emplean medidas únicas (véase la siguiente sección para una discusión más detallada) y se corre, además, **el riesgo de confundir potencialmente los resultados** seleccionando sólo una única variable para representar un resultado posiblemente más complejo.

**Por ejemplo**, suponga que las variables que representan diseño de página web, analítica web, precio y valor, cargan en varios factores. La selección de cualquiera de estas variables aislada daría lugar a interpretaciones sustancialmente distintas en cualquier análisis subsiguiente, aunque las 4 pueden estar tan próximamente relacionadas como para excluir tal acción. En segundo lugar, en casos donde varias cargas elevadas complican la selección de una única variable, el investigador puede no tener otra elección que la de emplear el análisis factorial como la base para **calcular una escala aditiva** o la **puntuación de factores** para usar como variables suplentes. El objetivo, como en el caso de seleccionar una única variable, es representar de la mejor forma posible la naturaleza básica del factor o componente.

### 12.9.2. Creación de escalas aditivas

El concepto de escala aditiva se define como aquella escala que está formada por la combinación de varias variables individuales dentro de una única medida compuesta, esto es, **se combinan todas las variables que cargan alto sobre un factor**, y el total (o más normalmente la **puntuación media de las variables**) se emplea como **variable de sustitución**. Una **escala aditiva** cuenta con **3 ventajas concretas**:

1. Es una forma de **eludir en alguna forma el error de medida** inherente a todas las variables observadas. **El error de medida es el grado en el cual los valores observados no son representativos de los valores “verdaderos”** debido a cierto número de razones, desde **errores reales** (por ejemplo, errores en la entrada de los datos) a la **incapacidad de los individuos de suministrar información con precisión**. El error de medida **enmascara cualquier relación** (por ejemplo, correlaciones o comparación de medias de grupos) y dificulta la estimación en los modelos multivariantes. La **escala aditiva reduce el error de medida** utilizando **indicadores múltiples (variables)** para **reducir la desconfianza** sobre una única respuesta. Empleando la **“media”** o la respuesta **“típica”** a un conjunto de variables relacionadas, **el error de medida que podría tener lugar en una única respuesta se reducirá**.
2. Su **capacidad para representar los múltiples aspectos de un concepto** en una única medida. En muchas ocasiones empleamos más variables en nuestros

modelos multivariantes en un intento de representar las muchas “*facetas*” de un concepto que sabemos es bastante complejo. Pero al actuar así, complicamos la interpretación de los resultados debido a la redundancia de la información asociada con el concepto. Por eso, nos gustaría no sólo obtener una descripción mejor de los conceptos utilizando múltiples variables, sino también mantener la parsimonia en el número de variables de nuestros modelos multivariantes. La escala aditiva, cuando se construye apropiadamente, **combina los múltiples indicadores en una medida única representando lo que se mantiene en común a lo largo del conjunto de medidas.**

3. El proceso de **construcción de la escala** está fundamentado teórica y empíricamente en una serie de disciplinas que incluyen **la teoría psicométrica, la sociología y el marketing**. Aunque un tratamiento completo de las técnicas y aspectos involucrados están más allá del alcance de este libro, existen fuentes excelentes para un estudio más extenso de estas materias [American Psychological Association 1985, Churchill, 1979, Hattie, 1985, Peter, 1981, Robinson et al.1991]. Adicionalmente hay una serie de compilaciones de escalas existentes que pueden aplicarse en varias situaciones [Bearden et al. 1993, Bruner y Hensel 1993, Robinson y Shaver 1973]. Aquí expondremos, sin embargo, **4 aspectos básicos** en la construcción de cualquier escala aditiva: **la definición conceptual, la dimensionalidad, la fiabilidad y la validación.**  
**-Definición conceptual.** Es el punto de partida para construir una escala aditiva. La definición conceptual especifica las bases teóricas de la escala aditiva definiendo el concepto que es representado en términos aplicables al contexto de investigación. En la investigación académica, las definiciones teóricas están basadas en investigación previa que define el carácter y naturaleza de un concepto. En un ámbito de gestión empresarial, los conceptos concretos pueden definirse con relación a los objetivos propuestos, tales como la imagen, el valor o la satisfacción. En cualquier caso, la definición conceptual es la que orienta y concreta el carácter y tipo de **ítems** que son candidatos a ser incluidos en la escala. **-La validación del contenido** es la evaluación de la correspondencia de las variables incluidas en la escala aditiva con su definición conceptual. Esta forma de validación, también conocida como **validación aparente**, sirve para apreciar subjetivamente la correspondencia entre los **ítems** individuales y el concepto a través de evaluaciones de expertos, contrastes previos con múltiples subpoblaciones, u otros medios. **El objetivo es asegurar que los ítems de la escala abarquen algo más que aspectos empíricos, también consideraciones teóricas y prácticas** [Churchill, 1979, Robinson et al.1991].  
**-Dimensionalidad.** Un supuesto subyacente y requisito esencial para construir una escala aditiva es que **los ítems sean unidimensionales**, o sea, **fuertemente asociados unos con otros representando un único concepto** [Hattie, 1985, McDonald, 1981]. El análisis factorial sirve de apoyo realizando una valoración empírica de la dimensionalidad del conjunto de ítems determinando el número de factores y las cargas de cada variable sobre el factor o factores. El contraste de unidimensionalidad **consiste en que en una escala aditiva los ítems carguen de forma elevada en un único factor** [Anderson et al. 1987, Hattie,1985, McDonald,1981, Nunnally, 1979]. Si se propone que una escala aditiva cuente con

múltiples dimensiones, **cada dimensión reflejará un factor aislado**. El investigador puede evaluar la **unidimensionalidad** bien con un análisis **factorial exploratorio**, como se discutió en este capítulo, o bien un **análisis factorial confirmatorio**.

**-La fiabilidad** es el **grado de consistencia entre las múltiples medidas de una variable**. Una forma de fiabilidad es el **test-retest** por el cual la consistencia se mide entre las **respuestas de un individuo en 2 momentos del tiempo**. El objetivo es asegurar que las respuestas no varían demasiado a lo largo de períodos temporales por lo que una medida tomada en cualquier momento del tiempo es certera.

Una **segunda medida de la fiabilidad** más utilizada es la consistencia interna que se aplica a la **consistencia entre las variables en una escala aditiva**. La motivación para la consistencia interna es que los **ítems** individuales o indicadores de la escala deberían estar midiendo las mismas construcciones y, de esta forma, estar altamente intercorrelacionadas [Churchill, 1979, Nunnally, 1979]. Debido a que no hay un único ítem que sea una medida perfecta de un concepto, debemos de legar en un conjunto de medidas de diagnóstico para valorar la consistencia interna. En primer lugar, existen varias medidas que se relacionan con cada ítem aislado, incluyendo la correlación ítem-total (la correlación del ítem con la puntuación de la escala aditiva) o la correlación inter-ítem (la correlación entre **ítems**). La práctica empírica sugiere **que la correlación ítem-total exceda de 0.50** y que **las correlaciones inter-ítem excedan de 0,30** [Robinson et al.1991]. Otro tipo de medida de diagnóstico es el **coeficiente de fiabilidad** que valora la consistencia de la escala entera, el **alfa de Cronbach** [Nunnally, 1979, Peter, 1979], que es la medida más extensamente utilizada. El acuerdo general sobre el límite inferior para al alfa de **Cronbach** es de **0.70** [Robinson et al.1991, Robinson y Shaver 1973], aunque puede bajar a **0.60** en la investigación exploratoria [Robinson et al.1991]. El **alfa de Cronbach** tiene una relación positiva con **el número de ítems de la escala**. Debido a que al **aumentar** el número de **ítems**, incluso con el **mismo grado de intercorrelación**, **se incrementará el valor de la fiabilidad**, los investigadores deben **imponer requisitos más restrictivos para escalas con un mayor número de ítems**. También están disponibles medidas de fiabilidad derivadas del **análisis factorial confirmatorio**. Dentro de ellas están la **fiabilidad compuesta** y la **varianza media extraída**, ambas discutidas con mayor detalle la técnica de **ecuaciones estructurales**.

Cada uno de los programas estadísticos más utilizados contiene ahora módulos de **evaluación de la fiabilidad**, de tal forma que el investigador está provisto de un análisis completo de medidas tanto específicas de cada ítem **como medidas de fiabilidad globales**. **En toda escala aditiva debe analizarse la fiabilidad** para asegurar su adecuación antes de proceder a una evaluación de su validación.

**-Validación**. Habiendo asegurado que una escala: se adecúa a su definición conceptual; es unidimensional, y alcanza los niveles necesarios de fiabilidad. Usted debe realizar una evaluación final: **la validación de la escala**. La validación es la medida en que una escala o un conjunto de medidas representan con

precisión el concepto de interés. Ya hemos comentado una forma de validación (**de contenido o validación por confrontación**) en relación a las definiciones conceptuales. Otras formas de validación se miden empíricamente por la **correlación** entre los conjuntos de variables definidas teóricamente. Las **3 formas más extensa mente admitidas de validación son: la convergente, la discriminante y la nomológica** [Campbell y Fiske 1959, Peter, 1981].

-**La validación convergente** valora el grado en el cual dos medidas del mismo concepto están correlacionadas. Aquí debe buscar medidas alternativas de un concepto y correlacionarlas con la escala aditiva. Altas correlaciones indican que la escala está midiendo el concepto que se pretendía.

-**La validación discriminante** es el grado en el cual dos conceptos conceptualmente parecidos difieren. El contraste empírico de nuevo es la correlación entre las medidas, pero esta vez la escala aditiva está correlacionada con una medida parecida, pero conceptualmente distinta. Ahora la correlación debería ser baja, demostrando que la escala aditiva es conceptualmente distinta de otro concepto parecido.

-Finalmente, la **validación nomológica** refleja el **grado** en que la escala aditiva hace **predicciones precisas de otros conceptos en un modelo basado en la teoría**. Así, debe identificar relaciones apoyadas en investigación previa o en principios aceptados y **evaluar si la escala cuenta con las correspondientes relaciones**.

En resumen:

1. La **validación convergente** confirma que la escala está correlacionada con otras medidas conocidas del concepto,
2. La **validación discriminante** asegura que la escala es suficientemente distinta de otros conceptos parecidos que sean distintos, y
3. la **validación nomológica** determina si la escala demuestra las relaciones cuya existencia se deriva de la teoría y/o de investigación previa.

**Existen varios métodos** para evaluar la **validación**, que van desde el **multirasgo, las matrices multimétodo (MTMM)** a las aproximaciones basadas en **ecuaciones estructurales**. Aunque vayan más allá del alcance de este libro, existen una serie de fuentes que orientan sobre el conjunto de métodos disponibles y de los aspectos implicados por las técnicas concretas [Campbell y Fiske 1959, Joreskog, y Sorbo 1993, Peter, 1981].

Las **escalas aditivas**, cuentan con aplicación creciente en investigación aplicada y también en gestión empresarial. La capacidad de la escala aditiva para plasmar conceptos complejos en una única medida **reduciendo el error de medida supone un añadido valioso en cualquier análisis multivariante**. El **análisis factorial** ofrece al investigador una **evaluación empírica** de las relaciones entre las variables, esencial en la formación de los fundamentos conceptuales y empíricos de la escala aditiva por medio de la evaluación de la validación del contenido y la dimensionalidad de la escala.

### 12.9.3. Cálculo de la puntuación factorial

Otra alternativa para crear un conjunto más pequeño de variables es **reemplazar el conjunto original por medio del cálculo de la puntuación factorial, que constituye también medidas compuestas de cada factor calculadas para cada sujeto. La puntuación factorial** representa el grado en el cual **cada individuo puntúa en el grupo de ítems que cuentan con cargas elevadas sobre un factor. Por tanto, mayores valores en las variables con altas cargas sobre un factor resultarán en una mayor puntuación factorial.** Una diferencia clave que de la **puntuación factorial vs. la escala aditiva** es que la primera se calcula con base a las **cargas factoriales**, mientras que la segunda lo hace **combinando sólo las variables elegidas**. En consecuencia, aunque el investigador sea capaz de caracterizar un factor por medio de las variables con cargas mayores, **se debe también atender a las cargas de las otras variables**, aunque menores, y su influencia sobre la puntuación factorial. La mayoría de los programas estadísticos puede calcular **puntuaciones factoriales** fácilmente para cada elemento muestral. Seleccionando la alternativa de **puntuación factorial**, se graban estas puntuaciones para su uso en análisis subsiguientes. **Una desventaja de la puntuación factorial es que no se pueden replicar fácilmente en otros estudios debido a que están basados en la matriz factorial obtenida en cada estudio.** La replicación de la misma matriz factorial en distintos estudios requiere un importante trabajo de programación.

### 12.9.4. Selección entre los 3 métodos

Para elegir entre **las 3 alternativas (12.9.1/12.9.2/12.9.3)** para **reducir datos** deberá tomar una serie de decisiones:

1. Seleccione **una única variable suplente** para cada factor o **calcule una medida compuesta**. La **variable suplente única** tiene las **ventajas** de ser **der un sencillo tratamiento e interpretación**, pero tiene las **desventajas** de **no representar las otras "facetas" del factor y su propensión al error de medida**.
2. Si opta por **calcular una medida compuesta**, debe elegir entre la **puntuación factorial y las escalas aditivas**. **Puntuación factorial** tiene la **ventaja** de representar un **compuesto de las cargas de las variables sobre el factor**, aunque esto supone también una **desventaja potencial** al contar todas las variables con algún **grado de influencia en el cálculo de la puntuación factorial y hacer la interpretación más difícil**. La **escala aditiva** está a medio camino entre **variable suplente** y la **alternativa de la puntuación factorial**. Es una medida compuesta, como la puntuación factorial, **reduciendo por tanto el error de medida** y representando **múltiples facetas de un concepto**. Aunque similar a la aproximación de **variable suplente incluye sólo las variables que cargan alto sobre el factor y excluye aquellas con escasos efectos**. Además, su fácil replicación entre muestras es similar al enfoque de **variable suplente**.
3. Finalmente, como las **variables suplentes**, las **escalas aditivas** no son necesariamente **ortogonales**, mientras que **los factores pueden ser ortogonales o incorrelados**, si se necesita evitar complicaciones en su uso en otras técnicas multivariantes.

4. La regla de decisión, es que:
- Si los datos se emplean sólo en la muestra original o se debe mantener la ortogonalidad, la puntuación factorial es la adecuada.
  - Si se desea la **transferibilidad** o la **generalización**, las **escalas aditivas** o las **variables suplentes son más apropiadas**.
  - Si la **escala aditiva** está bien construida, validada y es fiable, es probable que sea la mejor alternativa.
  - Si la **escala aditiva** no está contrastada y revisada, con poca o ninguna prueba de fiabilidad o validación, deberían considerarse en su lugar las variables suplentes si no es posible un análisis añadido que mejore la **escala aditiva**.

## 12.10. Análisis factorial: Resumen para aplicar

### ¿Qué es el análisis factorial?

- El **análisis factorial** tiene como objetivo principal simplificar las múltiples y complejas relaciones que puedan existir en un conjunto de variables observadas  $X_1, X_2, X_3, \dots, X_N$ . Para lograrlo, trata de encontrar dimensiones comunes o factores que ligan a las variables aparentemente no relacionadas.
- Concretamente, se trata de encontrar un conjunto de  $k < p$  factores no directamente observables  $F_1, F_2, F_3, \dots, F_k$  que expliquen suficiente a las variables observadas perdiendo el mínimo de información, de modo que sean fácilmente interpretables (**Principio de Interpretabilidad**), y que sean los menos posibles, es decir,  $k$  pequeño (**Principio de Parsimonia**).
- Además, los factores han de extraerse de forma que **resulten independientes entre sí**, es decir, que sean **ortogonales**. En consecuencia, el **análisis factorial** es una técnica de reducción de datos que examina la interdependencia de variables, y proporciona conocimiento de la estructura subyacente de los datos.
- Es un conjunto de métodos estadísticos que aborda el problema de **cómo analizar la estructura de las interrelaciones (correlaciones) entre un gran número de variables**, con la definición de **dimensiones subyacentes comunes llamadas factores**.

### ¿Para qué se usa?

- Simplificar un conjunto de datos reduciendo el número de variables, bien por un exceso de variables que dificulta el análisis de la información, o bien por representar la misma información de manera redundante.
- Encontrar la estructura subyacente en los datos analizados
- En ocasiones se utiliza como un método intermedio para dejar los datos listos en el empleo de otro método siguiente.
- Se pierde información, pero se gana en la facilidad de interpretación al menor costo posible

### Tipos de Análisis Factorial

- **Exploratorio:** Se utilizara principalmente para identificar factores, sin restricciones o hipótesis previas, y se caracteriza porque no se conoce a priori el número de factores, y es en la aplicación empírica donde se determina este número.

- **Confirmatorio:** Se utilizan cuando se desea verificar la existencia de una estructura subyacente en los datos, anticipada hipotéticamente, y los factores están fijados a priori, utilizándose contrastaciones empíricas para su corroboración.
- Cuando por **razonamiento teórico, por experiencias u otras** investigaciones similares se **formulan hipótesis sobre la dimensionalidad o estructura subyacente de un fenómeno**, es entonces cuando se utiliza el análisis factorial para confirmar.

El análisis factorial parte de la distinción de la variabilidad de las variables observadas y, por lo tanto de la varianza de las mismas variables, en dos tipos esenciales:

- **Una parte común:** explicada por un conjunto de factores comunes que afectan a todas las variables. Esto es en el entendido que no captan toda la variabilidad, sino sólo la común.
- **Una parte específica:** exclusiva para cada variable y sin relación con las demás, explicadas por factores específicos o únicos que informan sobre la especificidad o unicidad de cada variable. Son **factores independientes y ortogonales**.

### **Etapas del Análisis Factorial**

**Primera Etapa:** el propósito general del análisis factorial es encontrar la forma de condensar la información contenida en un amplio número de variables en la menor cantidad posible de factores.

**Segunda Etapa:** esta etapa involucra **3 decisiones básicas:**

- Calcular la **matriz de correlación** entre las variables;
- Medición** de las variables y;
- El **tamaño de la muestra** necesaria.

**Tercera Etapa:** esta etapa se centra en las **características y composición de las variables**, como la **multicolinealidad, la matriz de correlación anti-imagen y la prueba de esfericidad de Bartlett**.

**Cuarta Etapa:** esta etapa involucra **2 decisiones básicas:**

- El **método de extracción** de los factores y;
- El **número de factores seleccionados** para mejorar el ajuste de los datos.

**Quinta Etapa:** la interpretación de los factores es una etapa crucial, y se realiza a través de:

- La estimación del **factor de la matriz**;
- La rotación de los factores y;
- La reespecificación.

**Sexta Etapa:** esta etapa involucra el grado de generalización de los resultados obtenidos a la totalidad de la población objeto de estudio.

**Séptima Etapa:** esta etapa proporciona las bases empíricas para el análisis de la estructura de las variables utilizadas, y el impacto en la aplicación de otras técnicas multivariadas.

### **Análisis Previo de los Datos:**

- Anteriormente ya se ha indicado que es necesario dar una serie de pasos previos antes de aplicar una técnica multivariable determinada. Algunos de ellos tienen que ver con la propia técnica y la comprobación del cumplimiento de sus hipótesis subyacentes: **normalidad, homoscedasticidad y linealidad**; otras comprobaciones son, incluso, previas al uso de la técnica y tienen que ver con la fiabilidad de los datos de partida: existencia de valores perdidos y de observaciones anómalas.
- Asimismo, debe señalarse que algunas de las técnicas de análisis que se expondrán en apartados posteriores, tienen sus propios procedimientos para la comprobación del cumplimiento de sus hipótesis o, por ejemplo, la detección de las observaciones anómalas, y así serán presentadas en su momento (por ejemplo, en la regresión lineal múltiple).

### 12.11. Análisis factorial. Ejemplos

#### Paso 1: Objetivos

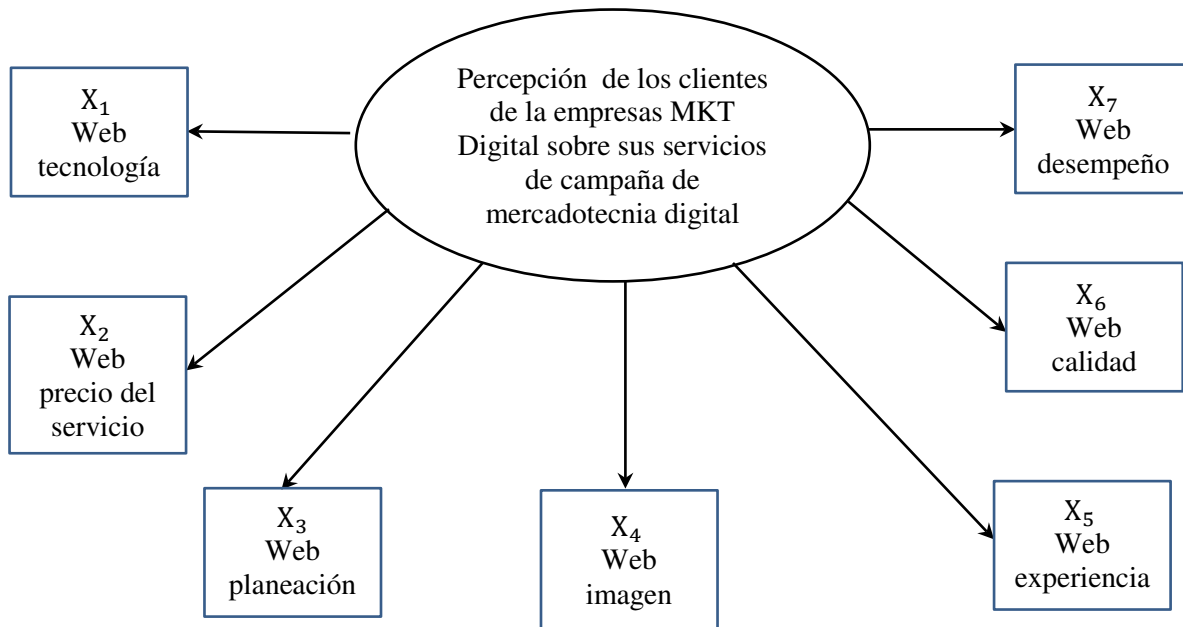
**Problema 1:** Para explicar de manera más sencilla y clara las diversas técnicas multivariables, se utilizará la base de datos de **WEB Diseño.sav**, de la compañía **MKT Digital**, la cual contiene datos de **100 clientes** a los que ha servido en sus campañas digitales., con la siguiente **estructura de base de datos**:

Descripción de variables	Tipo
<b>Clasificación de variables</b>	
V1.- Tipo de consumidor	No métrico
V2.- Tipo de Industria	No métrico
V3.- Tamaño de la firma	No métrico
V4.- Región	No métrico
V5.- Sistema de contratación	No métrico
<b>Variables de efectividad</b>	
X1.- Web tecnología	Métrico
X2.- Web precio del servicio	Métrico
X3.- Web planeación estratégica	Métrico
X4.- Web imagen	Métrico
X5.- Web experiencia del usuario	Métrico
X6.- Web calidad	Métrico
X7.- Web desempeño	Métrico
X8.- Web analítica	No métrico
X9.- Web contrataciones de clientes	Métrico
X10.- Web desempeño	Métrico
X11.- Web certificación	No métrico
X12.- Web plan de implementación	No métrico
X13.- Web seguridad	No métrico
X14.- Web situación de compra	No métrico
<b>Variables de resultado</b>	
V15.- Probabilidad de recomendación	Métrico
V16.- Probabilidad de compra futura	Métrico
V17.- Nivel actual de compra	Métrico
V18.- Futuras asociaciones/alianzas	No métrico



Así, se plantea la siguiente pregunta: **¿Cuál es la percepción que tienen los clientes de la empresa MKT Digital acerca de los servicios ofrecidos en sus campañas de mercadotecnia digital, en 7 atributos ( $X_1$  a  $X_7$ ) de efectividad? Ver Figura 12.9.**

**Figura 12.9. Modelo problema de análisis factorial exploratorio**



Fuente: propia

### Paso 2: Diseño

- La base de datos contiene 100 observaciones con un total de 23 variables, de las cuales se trabajará con 7/14 de las que corresponden al grupo de Efectividad. La base de datos WEB diseño.sav contiene las percepciones que tienen los 100 principales clientes
- Asimismo, se recolectó la información de percepciones a través de *focus groups* y entrevistas a profundidad aplicados a los principales clientes de la empresa MKT Digital
- Los gerentes de mercadotecnia digital, contestaron los 13 atributos utilizando una escala de 0 – 10, siendo 10 Excelente y 0 Pésimo, con 1 decimal.
- La información contempla aspectos como: tecnología, precio, planeación estratégica, imagen, experiencia del usuario, calidad, desempeño implicados en una campaña digital
- Con el resultado del análisis factorial, se espera detectar el mejor agrupamiento de éstas variables que le permitan al sector de mercadotecnia digital, ser más competitivo y prestar mejor servicio al cliente
- **Valores Perdidos:** La existencia de valores perdidos en un estudio de las Ciencias Sociales es algo prácticamente inevitable. Las consecuencias para la investigación

dependerá del patrón que sigan estos datos ausentes, cuántos son y por qué están perdidos.

- Como señalan **Tabachnick y Fidell (1996)**, el patrón de los valores perdidos es más importante que su cuantía; pues si su distribución es aleatoria en la matriz de datos no pueden causar mucho daño al análisis; sin embargo, si responden a un patrón determinado sí.
- **Valores perdidos menores o iguales al 10%** del total de los casos por lo general son ignorados y sustituidos por la media o la moda, según sea el caso, excepto cuando los valores perdidos se concentran en una pregunta determinada o un grupo de preguntas del cuestionario
- **Valores Atípicos u Outliers:** Son aquellos casos para los que una, dos o múltiples variables de una investigación determinada toman valores extremos que los hace diferir del comportamiento del resto de la muestra, y permiten al investigador sospechar que han sido alterados o generados por mecanismos distintos al resto de los datos (**Hawkins, 1980**).
- ¿Por qué es importante detectar los valores atípicos? Por las consecuencias que generan:
  - Distorsionan los resultados al **oscurecer el patrón de comportamiento** del resto de casos y obtenerse conclusiones que, sin ellos, serían completamente distintas y,
  - Pueden afectar gravemente a una de las condiciones de aplicabilidad más habituales de la mayor parte de las técnicas multivariadas: la normalidad.

### **Paso 3: Condiciones de Aplicabilidad:**

- **Normalidad:** La condición básica que debe asumirse en el análisis multivariable es la normalidad, y se refiere a que todos los datos de las variables métricas deben de seguir una distribución normal. Si la variación de la distribución normal es demasiado amplia, todos los resultados del análisis multivariable serán inválidos, porque la normalidad es un requisito esencial para los **estadísticos F y t**.
- La prueba de normalidad para una sola variable es fácil de realizar y existen diversos métodos para estimarse, pero la normalidad multivariable es más complicada de realizar y existen relativamente pocos métodos para estimarla.
- **Métodos: Análisis de la Asimetría y Curtosis, Gráficos q-q de residuos y Estadísticos de prueba de Kolmogorov-Smirnov-Lilliefors (KSL).**
- **Solución: Transformación potencial y logarítmica**

#### **Paso 1: Objetivos**

- **Problema 2:** De la base de datos **WEB Diseño.sav** donde **N>50** muestras, pruebe qué Hipótesis es aprobada:
  - 0.-Las variables **X<sub>1</sub>** a **X<sub>7</sub>** tienen una población con distribución normal
  - 1.-La variable **X<sub>1</sub>** a **X<sub>7</sub>** **NO** tienen una población con distribución normal

#### **Paso 2: Diseño; Paso 3: Condiciones de aplicabilidad.**

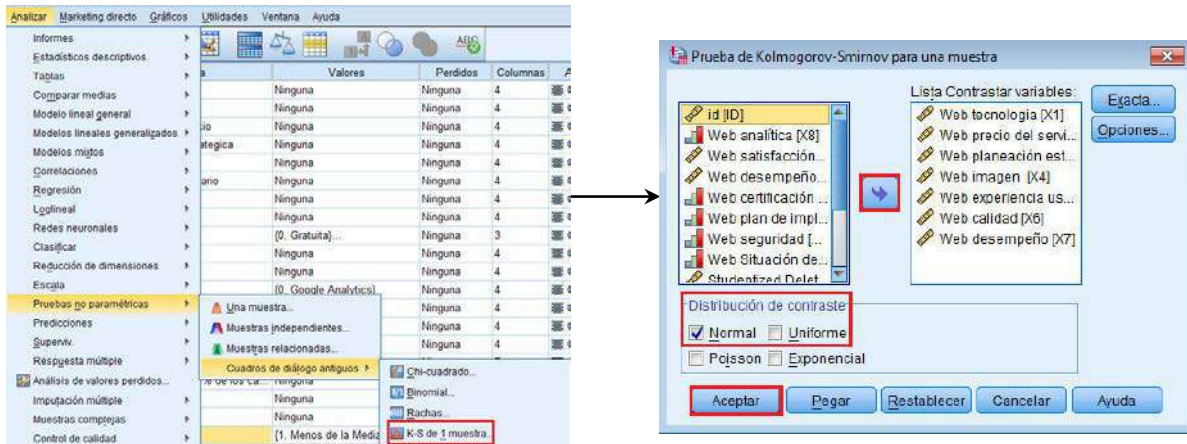
Dado que **N>50** muestras, la normalidad se analizará de acuerdo a **Kolmogorov-Smirnov**.

#### **Paso 4: Ejecución y ajuste**

**Teclear: Analizar->Pruebas no paramétricas->Cuadro de diálogo antiguos->K-S de 1 muestra; Seleccionar lista de variables de prueba (X<sub>1</sub> a X<sub>7</sub>);**

Seleccionar Distribución de prueba (Normal)->Aceptar->Continuar->Aceptar. Ver Figura 12.10

Figura 12.10 Prueba de normalidad



➔ Pruebas no paramétricas

[Conjunto\_de\_datos] C:\Users\Juan\Desktop\proy libro mc\WEB diseño copia.sav

	Web tecnología	Web precio del servicio	Web planeación estratégica	Web imagen	Web experiencia usuario	Web calidad	Web desempeño
N	100	100	100	100	100	100	100
Parámetros normales <sup>a,b</sup>	Media	3.515	2.364	7.894	5.248	2.916	6.971
	Desviación típica	1.3207	1.1957	1.3865	1.1314	.7513	1.5852
Diferencias más extremas	Absoluta	.063	.095	.095	.107	.085	.122
	Positiva	.055	.095	.085	.107	.058	.074
	Negativa	-.063	-.068	-.095	-.104	-.085	-.091
Z de Kolmogorov-Smirnov	.628	.945	.947	1.068	.854	1.221	.909
<b>Sig. asintót. (bilateral)</b>	<b>.825</b>	<b>.333</b>	<b>.331</b>	<b>.205</b>	<b>.460</b>	<b>.101</b>	<b>.380</b>

a. La distribución de contraste es la Normal.  
b. Se han calculado a partir de los datos.

Fuente: SPSS 20-IBM

**Paso 5: Interpretación**

Dado que las  $p > 0.05$  se acepta  $H_0$ .-Las variables  $X_1$  a  $X_7$  SI tienen una población con distribución normal

**Paso 1: Objetivos**

**Problema 3:** generar análisis de descriptivos y gráficos Q-Q de normalidad

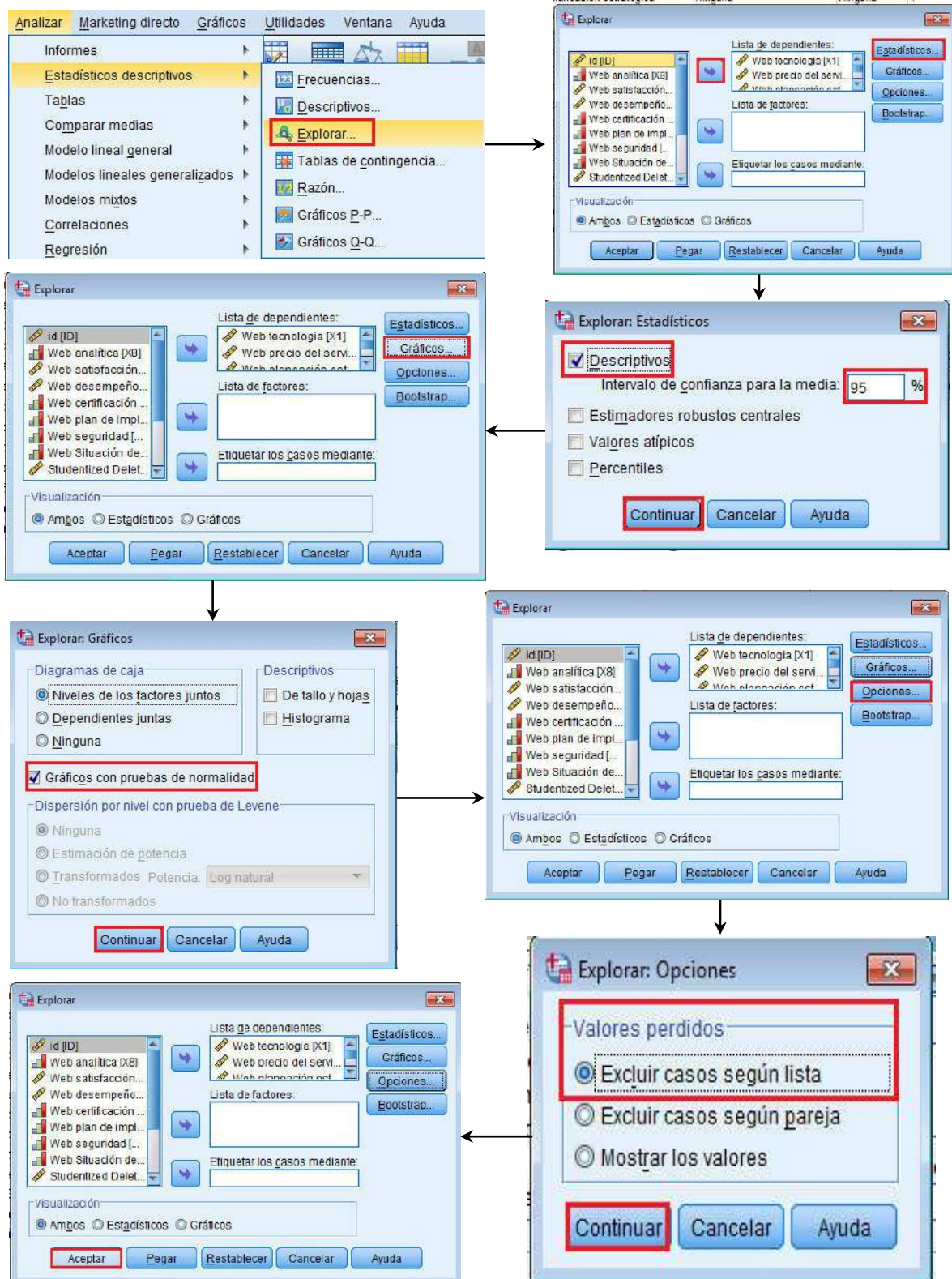
**Paso 2: Diseño; Paso 3: Condiciones de aplicabilidad. Se consideran sin cambio**

**Paso 4: Ejecución y ajuste**

Teclear: Analizar->Estadísticos descriptivos->Explorar->Seleccionar lista de dependientes ( $X_1$  a  $X_7$ )->Estadísticos->Seleccionar: Descriptivos->Confirmar: 95%-> Continuar->Gráficos->Seleccionar: Gráficos con pruebas de normalidad-

>Continuar->Opciones->Valores perdidos, seleccionar: Excluir casos según lista->Continuar->Aceptar. Ver Figura 12.11

Figura 12.11 Proceso para generar descriptivos y gráficos Q-Q de normalidad



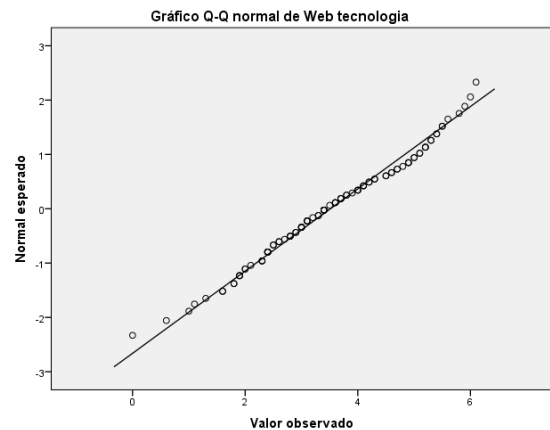
- SPSS genera la **Figura 12.12**, que contiene las tablas **Resumen de procesamiento de los casos**, **Descriptivos** y **Pruebas de normalidad**. Ver **Figura 12.12**

**Figura 12.12. Tabla Resumen del procesamiento de los casos; Descriptivos; Pruebas de normalidad y Gráficos Q-Q de normalidad**

➔ **Explorar**

[Conjunto\_de\_datos1] C:\Users\Juan\Desktop\proy libro mc\WEB diseño copia.sav

Resumen del procesamiento de los casos						
	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Web tecnologia	100	100.0%	0	0.0%	100	100.0%
Web precio del servicio	100	100.0%	0	0.0%	100	100.0%
Web planeación estratégica	100	100.0%	0	0.0%	100	100.0%
Web imagen	100	100.0%	0	0.0%	100	100.0%
Web experiencia usuario	100	100.0%	0	0.0%	100	100.0%
Web calidad	100	100.0%	0	0.0%	100	100.0%
Web desempeño	100	100.0%	0	0.0%	100	100.0%



Descriptivos			
		Estadístico	Error tip.
Web tecnologia	Media	3.515	.1321
	Intervalo de confianza para la media al 95%	Límite inferior	3.253
		Límite superior	3.777
	Media recortada al 5%	3.534	
	Mediana	3.400	

	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Web tecnologia	.063	100	.200	.985	100	.341
Web precio del servicio	.095	100	.028	.969	100	.017
Web planeación estratégica	.095	100	.027	.950	100	.001
Web imagen	.107	100	.007	.982	100	.183
Web experiencia usuario	.085	100	.069	.986	100	.366
Web calidad	.122	100	.001	.963	100	.007
Web desempeño	.091	100	.041	.971	100	.028

\*. Este es un límite inferior de la significación verdadera.

a. Corrección de la significación de Lilliefors

Fuente: SPSS 20 IBM



- **Homocedasticidad:** Debe definirse de manera distinta según se estén analizando datos no agrupados (caso de una regresión lineal múltiple), o datos agrupados (caso de un análisis de la varianza de un factor).
- En el primer caso la hipótesis de homocedasticidad puede definirse como el supuesto de que cada uno de los valores que puede tomar la distribución se mantiene constante para todos los valores de la otra variable continua.
- En el caso de datos agrupados la homocedasticidad implica que la varianza de la variable continua es más o menos la misma en todos los grupos que conforman la variable no métrica que es la que determina los grupos.
- En resumen, se puede decir que la homocedasticidad es la igualdad de varianza entre las variables independientes.
- **Métodos: Test de Levene**
- **Solución: Transformación logarítmica y potencial**

### Paso 1: Objetivos

**Problema 4:** Compruebe la homocedasticidad de la base de datos **WEB Diseño.sav** a partir del tamaño de las empresas vs. cada variable métrica de la misma.

$H_0$ : **No** existen diferencias significativas entre las varianzas de las variables con el tamaño de la empresa

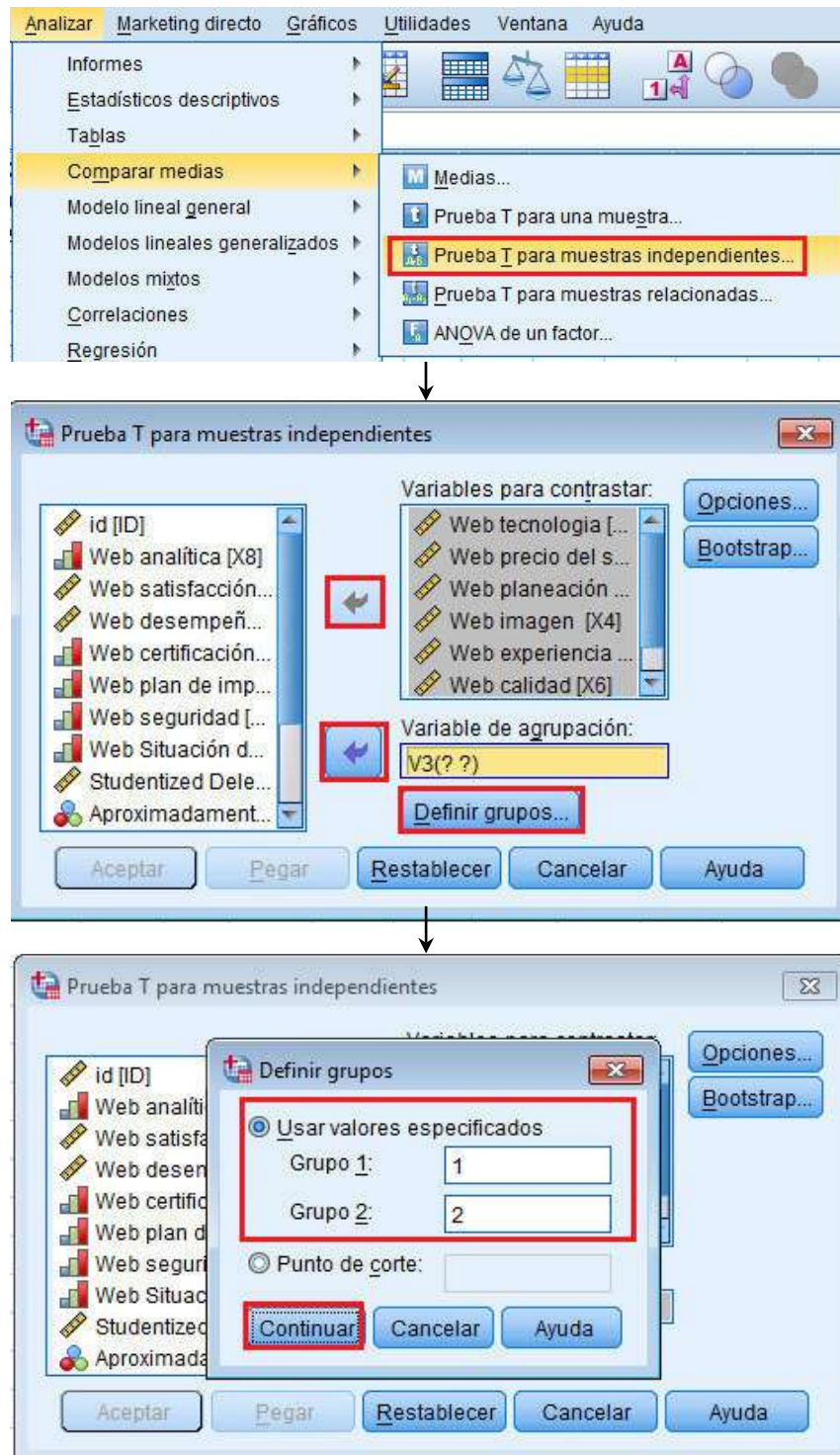
$H_1$ : **Sí** existen diferencias significativas entre las varianzas de las variables con el tamaño de la empresa

**Paso 2: Diseño. Paso 3: Condiciones de aplicabilidad.** Se toma en cuenta el ejemplo anterior y que cumple con las condiciones iniciales. Así, se aplicará test de **Levene** para comprobar la **homocedasticidad** (varianzas iguales de las variables dependientes). Se realiza a través de **t de Student** como sigue:

### Paso 4: Ejecución y ajuste

**Teclee: Analizar-> Comparar medias->Prueba T para muestras independientes->Selección Variables para contrastar métricas:  $X_1$ ; Selección Variable de agrupación nominal: Tamaño de la empresa (V3)->Definir grupos-> marcar las etiquetas a comparar (1,2,3,4,etc) por pares Continuar->Aceptar. Ver Figura 12.13.**

Figura 12.13.- Proceso para verificar homocedasticidad de la base de datos



Fuente:  
SPSS 20  
IBM

### Paso 5: Interpretación

- Dado que  $p > 0.05$  en todos los casos Se acepta la  $H_0$ : No existen diferencias significativas entre las varianzas de las variables con el tamaño de la empresa . Ver Figura 12.14

**Figura 12.14.- Resultados de Prueba de Levene**

Prueba de muestras independientes										
		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error tip. de la diferencia	95% Intervalo de confianza por la diferencia	
									Inferior	Superior
Web tecnología	Se han asumido varianzas iguales	.753	.389	1.244	63	.218	.4873	.3916	-2.952	1.26
	No se han asumido varianzas iguales			1.110	19.829	.280	.4873	.4391	-4.291	1.40
Web precio del servicio	Se han asumido varianzas iguales	.088	.768	-1.554	63	.125	-.5587	.3594	-1.2770	.15
	No se han asumido varianzas iguales			-1.653	25.483	.111	-.5587	.3380	-1.2540	.13
Web planeación estratégica	Se han asumido varianzas iguales	.918	.342	1.099	63	.276	.4087	.3717	-.3342	1.16
	No se han asumido varianzas iguales			1.029	21.047	.315	.4087	.3970	-.4167	1.23
Web imagen	Se han asumido varianzas iguales	.553	.460	-.061	63	.951	-.0200	.3269	-.6732	.63
	No se han asumido varianzas iguales			-.064	24.701	.950	-.0200	.3130	-.6651	.62
Web experiencia usuario	Se han asumido varianzas iguales	.170	.682	-.152	63	.880	-.0340	.2238	-.4812	.41
	No se han asumido varianzas iguales			-.150	22.653	.882	-.0340	.2266	-.5031	.43
Web calidad	Se han asumido varianzas iguales	.030	.862	.330	63	.742	.0740	.2241	-.3738	.52
	No se han asumido			.313	21.426	.757	.0740	.2360	-.4163	.56

Fuente: SPSS 20 IBM

- **Linealidad:** el supuesto de linealidad es fundamental para todas aquellas técnicas que se centren en el análisis de las **matrices de correlación o de varianzas – covarianzas, como el análisis factorial, regresión lineal o los modelos de ecuaciones estructurales**. La razón es sencilla: el coeficiente de **correlación de Pearson** sólo podrá captar una relación si ésta es lineal.
- Si la relación existe y es intensa pero, por ejemplo, **es curvilínea**, el **coeficiente de correlación de Pearson tomará un valor relativamente bajo** y el investigador puede interpretarlo como **ausencia de relación cuando, de hecho, ésta existe sólo que no es lineal**.
- Cuando la técnica empleada tiene **una variable dependiente**, como ocurre en el caso de la **regresión lineal múltiple**, existen diversos procedimientos para contrastar la linealidad de las relaciones basadas en el análisis de los residuos o residuales.
- **Métodos: Gráficos: de dispersión entre variables y**
- **Estadísticos: Coeficientes de correlación bivariados.**

### Paso 1: Objetivos

**Problema 5:** Verifique las correlaciones de las variables métricas de la base de datos y analícelas.

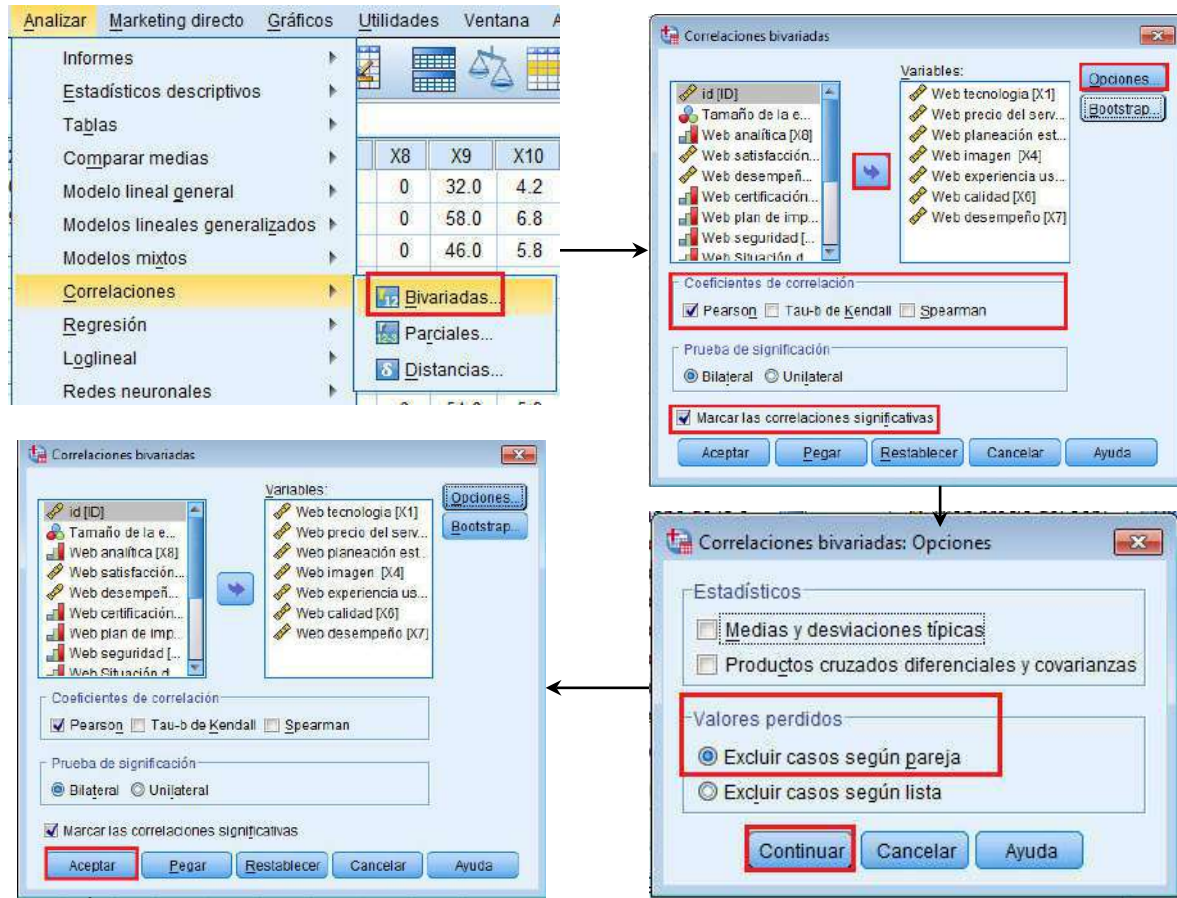
**Paso 2: Diseño. Paso 3: Condiciones de aplicabilidad.** Se toma en cuenta el ejemplo anterior y que cumple con las condiciones iniciales.



#### Paso 4: Ejecución y ajuste

Teclee: **Analizar->correlaciones->bivariadas->Coeficientes de correlación: Pearson->Marcar las correlaciones significativas->Opciones->Valores perdidos: Excluir casos según pareja->continuar. Ver Figura 12.15**

Figura 12.15.- Proceso para determinar la matriz de correlaciones



Correlaciones

		Web tecnología	Web precio del servicio	Web planeación estratégica	Web imagen	Web experiencia usuario	Web calidad	Web desempeño
Web tecnología	Correlación de Pearson	1	-.349**	.509**	.050	.612**	.077	-.483**
	Sig. (bilateral)		.000	.000	.618	.000	.444	.000
	N	100	100	100	100	100	100	100
Web precio del servicio	Correlación de Pearson	-.349**	1	-.487**	.272**	.513**	.185	.470**
	Sig. (bilateral)	.000		.000	.006	.000	.065	.000
	N	100	100	100	100	100	100	100
Web planeación estratégica	Correlación de Pearson	.509**	-.487**	1	-.116	.067	-.035	-.448**
	Sig. (bilateral)	.000	.000		.250	.510	.731	.000
	N	100	100	100	100	100	100	100
Web imagen	Correlación de Pearson	.050	.272**	-.116	1	.299**	.788**	.200
	Sig. (bilateral)	.618	.006	.250		.003	.000	.046
	N	100	100	100	100	100	100	100
Web experiencia usuario	Correlación de Pearson	.612**	.513**	.067	.299**	1	.240	-.055
	Sig. (bilateral)	.000	.000	.510	.003		.016	.586
	N	100	100	100	100	100	100	100
Web calidad	Correlación de Pearson	.077	.185	-.035	.788**	.240	1	.177
	Sig. (bilateral)	.444	.065	.731	.000	.016		.079
	N	100	100	100	100	100	100	100
Web desempeño	Correlación de Pearson	-.483**	.470**	-.448**	.200	-.055	.177	1
	Sig. (bilateral)	.000	.000	.000	.046	.586	.079	
	N	100	100	100	100	100	100	100

\*\* La correlación es significativa al nivel 0,01 (bilateral).

\* La correlación es significativa al nivel 0,05 (bilateral).

### Paso 1: establecimiento de objetivos

- **Objetivo:** Identificar la estructura de un grupo de variables a través de la reducción de datos. Así, **problema 5: determinar la percepción que tienen los gerentes de las empresas del sector de mercadotecnia digital acerca de los servicios ofrecidos a sus consumidores medidos en 7 atributos ( $X_1$  a  $X_7$ )** por las siguientes razones:
- Entender el agrupamiento de estas percepciones sobre los **7 atributos** generan una matriz de **correlaciones** que serán agrupadas en términos de lo que los gerentes de las empresas del sector de mercadotecnia digital tienen sobre los servicios ofrecidos a sus clientes.
- Reducir las **7 variables en el menor número de factores:** Si las **7 variables** se pueden representar en un pequeño número de factores o dimensiones, entonces se eligió a la técnica multivariable correcta y las otras técnicas ya no serían correctas.

### Paso 2: Diseño

- ¿Cómo deben medirse las variables a utilizar?  
**Deben ser métricas en general**, pero pueden introducirse no métricas (Variables Dummies), si bien no gozan de las mismas propiedades.
- ¿Cuál es el tamaño de la muestra?  
**No menos de 50 observaciones**, aunque lo mejor es **contar con más de 100 observaciones**.  
**No menos de 5 observaciones por cada variable**, si bien el **ratio óptimo es 10 a 1**  
En el caso de **WEB Diseño.sav** se tienen **100 observaciones** y **7 variables**, lo que da un ratio adecuado de **14 a 1**. (**100/7**)

### Paso 3: supuestos de Aplicabilidad

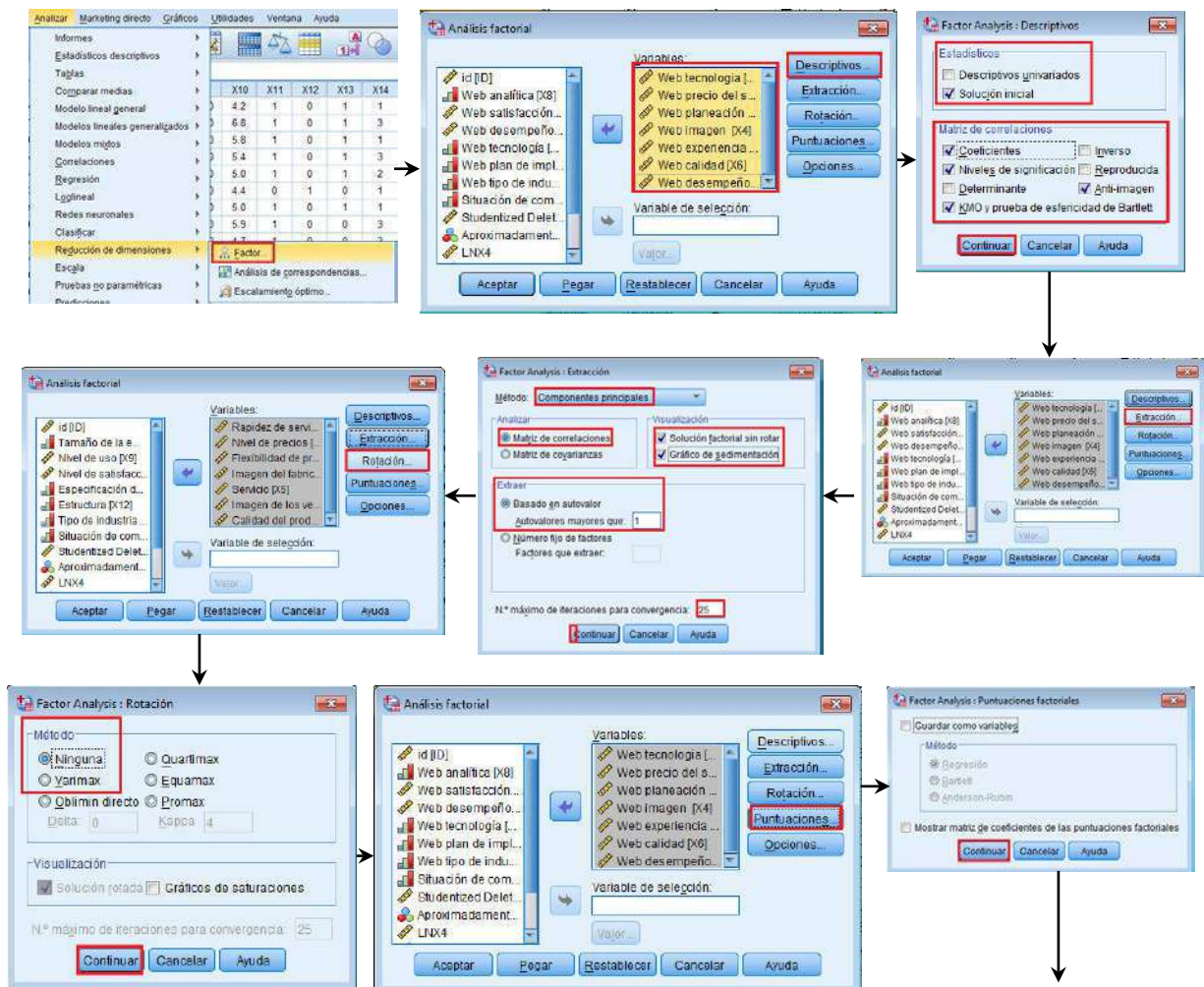
- **No se considera determinante la falta de normalidad, homocedasticidad pero sí la linealidad. Así que se debe cumplir:**
  1. Hay que asegurarse que existen suficientes correlaciones significativas entre las variables (**incorrelación implica un solo factor por cada variable**). Por linealidad **50%+1**
  2. Estas correlaciones **deben ser importantes (> a 0.3)** La **diagonal de la matriz de correlación anti-imagen** debe ser superior a **0.5**
  3. El test de **esfericidad de Bartlett** debe ser significativo (**0.01 ó 0.05 error / 99% o 95% confiabilidad**)
  4. El **KMO** debe ser **> 0.5 (lo ideal es > a 0.7)**
  5. Su **p** debe ser **inferior a los niveles críticos 0.05 o 0.01**. Debe saberse, sin embargo, que es un test muy sensible a incrementos en el tamaño de la muestra. Cuando esta se incrementa es más fácil que encuentra correlaciones significativas.
  6. Método de extracción con valores a **1; Varianza extraída superior al 60%** por gráfico de sedimentación

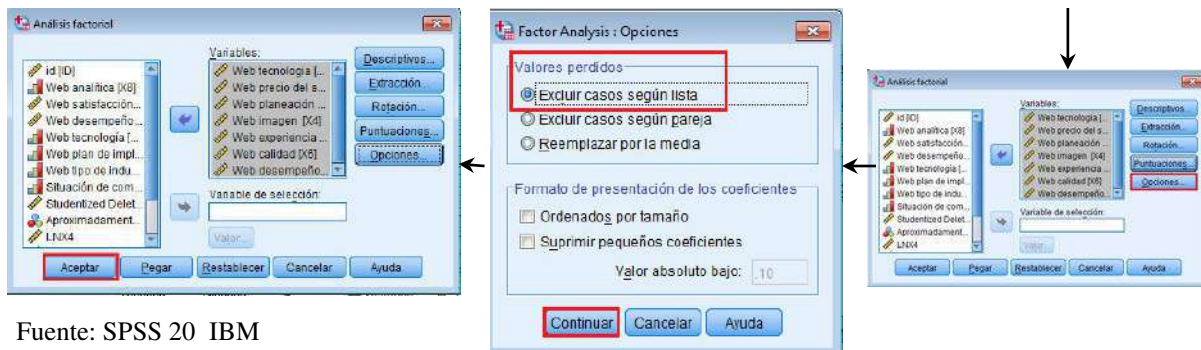
Deben cumplirse si no es así, **el análisis factorial no será posible.**

#### Paso 4: Ejecución y ajuste

Teclear: Analizar->Reducción de dimensiones->Selección de variables métricas->Descriptivos-> Seleccionar: Estadísticos: Solución inicial; Matriz de correlaciones: Coeficientes; Niveles de significación-> *KMO* y prueba de esfericidad de *Bartlett*; Anti-imagen->Continuar->Extracción->Método: componentes principales->Analizar: Matriz de correlaciones->Visualización: solución factorial sin rotar-> Seleccionar: Gráfico de sedimentación->Extraer: Autovalores mayores que: 1->Numero máximo de interacciones: 25->Continuar->Rotación->Método: Varimax->Visualización: solución rotada->No. máximo de iteraciones para convergencia: 25->Continuar->Puntuaciones->Continuar->Opciones->Valores perdidos (nota: no reemplazar por medias; no pérdida de datos >10%)->Excluir casos según lista (nota: evitar falta de datos)->Continuar->Aceptar. Ver Figura 12.16. De secuencia de instrucciones del análisis factorial

Figura 12.16. De secuencia de instrucciones del análisis factorial





Fuente: SPSS 20 IBM

### Paso 5: Interpretación

- a. SPSS genera la **Matriz de correlaciones**, de la que debe asegurarse que existen suficientes correlaciones significativas entre las variables (**incorrelación implica un solo factor por cada variable**). Por linealidad **50%+1**
  - Se observan **21 datos** por lado de la matriz bajo/sobre la diagonal
  - La mayoría deberá ser **50%+1=21/2+1=11 correlaciones deben ser significativas para cumplir**.
  - Existen 12 al 0.01 con este cumple también cumple, son significativas-
  - Existen 15 al 0.05 con este también cumple, son significativas. **Se recomienda usar. SI CUMPLE CONDICION. Ver Figura 12.17.**

Figura 12.17. Matriz de correlaciones

**matriz de correlaciones**

		Web tecnología	Web precio del servicio	Web planeación estratégica	Web imagen	Web experiencia usuario	Web calidad	Web desempeño
Correlación	Web tecnología	1.000	-.349	.509	.050	.612	.077	-.483
	Web precio del servicio	-.349	1.000	-.487	.272	.513	.185	.470
	Web planeación estratégica	.509	-.487	1.000	-.116	.067	-.035	-.448
	Web imagen	.050	.272	-.116	1.000	.299	.788	.200
	Web experiencia usuario	.612	.513	.067	.299	1.000	.240	-.055
	Web calidad	.077	.185	-.035	.788	.240	1.000	.177
	Web desempeño	-.483	.470	-.448	.200	-.055	.177	1.000
Sig. (Unilateral)	Web tecnología		.000	.000	.309	.000	.222	.000
	Web precio del servicio	X .000		.000	.003	.000	.032	.000
	Web planeación estratégica	X .000	X .000		.125	.255	.366	.000
	Web imagen	.309	X .003	.125		.001	.000	.023
	Web experiencia usuario	X .000	X .000	.255	X .001		.008	.293
	Web calidad	.222	X .032	.366	X .000	X .008		.039
	Web desempeño	X .000	X .000	X .000	X .023	.293	X .039	

Nota :   Datos <=0.01; X Datos <= 0.05

Fuente: SPSS 20 IBM

### b. Estas correlaciones **deben ser importantes (> a 0.3)**

- Se observan **21 datos** por lado de la matriz bajo/sobre la diagonal
- La mayoría deberá ser **50%+1=21/2+1=11 correlaciones deben ser significativas para cumplir**.



**-Existen sólo 9 /11 con valor  $\geq 0.3$  y 2 muy cercanos. CUMPLE MUY CERRADAMENTE LA CONDICION. Ver Figura 12.18.**

**Figura 12.18. Matriz de correlaciones**

**Matriz de correlaciones**

	Web tecnología	Web precio del servicio	Web planeación estratégica	Web imagen	Web experiencia usuario	Web calidad	Web desempeño	
Correlación	Web tecnología	1.000	-.349	.509	.050	.612	.077	-.483
	Web precio del servicio	X -.349	1.000	-.487	.272	.513	.185	.470
	Web planeación estratégica	X .509	X -.487	1.000	-.116	.067	-.035	-.448
	Web imagen	.050	.272	-.116	1.000	.299	.788	.200
	Web experiencia usuario	X .612	X .513	.067	.299	1.000	.240	-.055
	Web calidad	.077	.185	-.035	X .788	.240	1.000	.177
	Web desempeño	X -.483	X .470	X -.448	.200	-.055	.177	1.000
Sig. (Unilateral)	Web tecnología	.000	.000	.309	.000	.222	.000	
	Web precio del servicio	.000	.000	.003	.000	.032	.000	
	Web planeación estratégica	.000	.000	.125	.255	.366	.000	
	Web imagen	.309	.003	.125	.001	.000	.023	
	Web experiencia usuario	.000	.000	.255	.001	.008	.293	
	Web calidad	.222	.032	.366	.000	.008	.039	
	Web desempeño	.000	.000	.000	.023	.293	.039	

Nota: .272.-Muy cercano a 0.3; X.- 9 datos  $\geq 0.3$

Fuente: SPSS 20 IBM

**c. SPSS genera la tabla Matrices anti-imagen de la que se observa una diagonal con un valor que debe ser superior a 0.5**

- Sólo  $X_1, X_2$  y  $X_1$  a  $X_5$ . Con una sola que no cumpla ya no es posible continuar. **NO CUMPLE CONDICION. Ver Figura 12.19.**

**Figura 12.19. Tabla Matriz anti-imagen**

**Matrices anti-imagen**

	Web tecnología	Web precio del servicio	Web planeación estratégica	Web imagen	Web experiencia usuario	Web calidad	Web desempeño	
Covarianza anti-imagen	Web tecnología	.028	.028	.002	.015	-.025	-.006	-.002
	Web precio del servicio	.028	.032	.021	.014	-.026	-.005	-.020
	Web planeación estratégica	.002	.021	.608	.043	-.011	-.039	.086
	Web imagen	.015	.014	.043	.347	-.015	-.275	-.018
	Web experiencia usuario	-.025	-.026	-.011	-.015	.023	.005	.010
	Web calidad	-.006	-.005	-.039	-.275	.005	.371	-.044
	Web desempeño	-.002	-.020	.086	-.018	.010	-.044	.623
Correlación anti-imagen	Web tecnología	.345 <sup>a</sup>	.957	.018	.148	-.978	-.059	-.016
	Web precio del servicio	.957	.330 <sup>a</sup>	.155	.133	-.975	-.043	-.141
	Web planeación estratégica	.018	.155	.914 <sup>a</sup>	.094	-.091	-.083	.139
	Web imagen	.148	.133	.094	.558 <sup>a</sup>	-.172	-.766	-.040
	Web experiencia usuario	-.978	-.975	-.091	-.172	.288 <sup>a</sup>	.051	.088
	Web calidad	-.059	-.043	-.083	-.766	.051	.552 <sup>a</sup>	-.091
	Web desempeño	-.016	-.141	.139	-.040	.088	-.091	.927 <sup>a</sup>

a. Medida de adecuación muestral

Nota: .914<sup>a</sup> Valores de  $a > 0.5$

Fuente: SPSS 20 IBM

- d. El **test de esfericidad de Bartlett** debe ser significativo (0.01 ó 0.05 error / 99% o 95% confiabilidad)...**Sí es significativa al 0.01...CUMPLE**
- e. El **KMO** debe ser **> 0.5** (lo ideal es **> a 0.7**)....**NO CUMPLE. Ver Figura 12.20.**

**Figura 12.20. De resultados del análisis factorial**

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		.446
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	567.467
	gl	21
	Sig.	.000

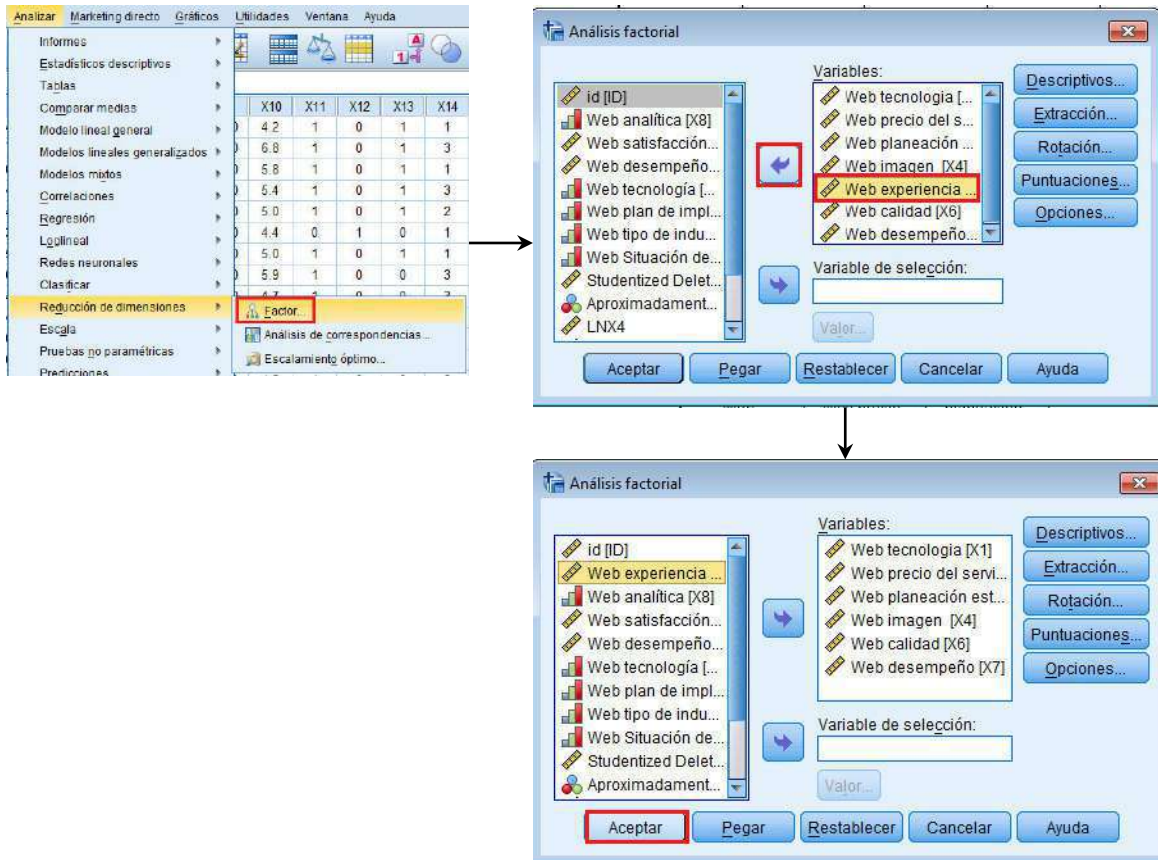
Fuente: SPSS 20 IBM

- f. Se concluye que hay problemas con las variables, por lo que se recomienda revisar las que se reportan en la matriz anti-imagen, que son:  **$X_1$ ,  $X_2$ ,  $X_5$**  La que se recomienda remover es la que tiene un valor bajo en la matriz anti-imagen:  **$X_5$  (.288). No significa que la variable no sirva.** De remover  **$X_5$**  se requiere realizar **TODO** el proceso anterior para volver a revisar. **Ver Figura 12.21.**

Figura 12.21. Separación de la variable X5 (Web experiencia) de WEB diseño.sav para **reinicio** de análisis factorial. Ver Paso 4.

**Paso 4: Ejecución y ajuste. Reinicio del análisis factorial.**

**Teclear: Analizar->Reducción de dimensiones->Factor->Seleccionar variable a remover->Flecha->Aceptar->Aceptar**



Fuente: SPSS 20 IBM

- g. Analizar la Matriz de correlaciones resultante. Debe asegurarse que existen suficientes correlaciones significativas entre las variables (**incorrelación implica un solo factor por cada variable**). Por linealidad **50%+1**:
  - Se observan **15 datos** por lado de la matriz bajo/sobre la diagonal
  - La mayoría deberá ser **50%+1=15/2+1=8 correlaciones deben ser significativas para cumplir**
  - Existen 8 al **0.01** con este cumple también cumple, son significativas
  - Existen 11 al **0.05** con este también cumple, son significativas. **Se recomienda usar. SI CUMPLE CONDICION. Ver Figura 12.22.**

**Figura 12.22. De resultados del análisis factorial**

**Matriz de correlaciones**

		Web tecnología	Web precio del servicio	Web planeación estratégica	Web imagen	Web calidad	Web desempeño
Correlación	Web tecnología	1.000	-.349	.509	.050	.077	-.483
	Web precio del servicio	-.349	1.000	-.487	.272	.185	.470
	Web planeación estratégica	.509	-.487	1.000	-.116	-.035	-.448
	Web imagen	.050	.272	-.116	1.000	.788	.200
	Web calidad	.077	.185	-.035	.788	1.000	.177
	Web desempeño	-.483	.470	-.448	.200	.177	1.000
Sig. (Unilateral)	Web tecnología		.000	.000	.309	.222	.000
	Web precio del servicio	X .000		.000	.003	.032	.000
	Web planeación estratégica	X .000	X .000		.125	.366	.000
	Web imagen	.309	X .003	.125		.000	.023
	Web calidad	.222	X .032	.366	X .000		.039
	Web desempeño	X .000	X .000	X .000	X .023	X .039	

Nota:   Datos con valor  $\leq 0.01$ ; X.- Datos con valor  $\leq 0.05$

Fuente: SPSS 20 IBM

**h.** Estas correlaciones **deben ser importantes ( $> a 0.3$ )**.

-Se observan 21 datos por lado de la matriz bajo/sobre la diagonal

-La mayoría deberá ser  $50\%+1=15/2+1=8$  correlaciones deben ser significativas para cumplir

**-Existen sólo 7/8 con valor  $\geq 0.3$  y 1 muy cercanos. CUMPLE MUY CERRADAMENTE LA CONDICION. Ver Figura 12.23.**

**Figura 12.23. De resultados del análisis factorial**

**Matriz de correlaciones**

		Web tecnología	Web precio del servicio	Web planeación estratégica	Web imagen	Web calidad	Web desempeño
Correlación	Web tecnología	1.000	-.349	.509	.050	.077	-.483
	Web precio del servicio	<span style="border: 1px solid red; padding: 0 2px;">-.349</span>	1.000	-.487	.272	.185	.470
	Web planeación estratégica	<span style="border: 1px solid red; padding: 0 2px;">.509</span>	<span style="border: 1px solid red; padding: 0 2px;">-.487</span>	1.000	-.116	-.035	-.448
	Web imagen	.050	X .272	-.116	1.000	.788	.200
	Web calidad	.077	.185	-.035	<span style="border: 1px solid red; padding: 0 2px;">.788</span>	1.000	.177
	Web desempeño	<span style="border: 1px solid red; padding: 0 2px;">-.483</span>	<span style="border: 1px solid red; padding: 0 2px;">.470</span>	<span style="border: 1px solid red; padding: 0 2px;">-.448</span>	.200	.177	1.000
Sig. (Unilateral)	Web tecnología		.000	.000	.309	.222	.000
	Web precio del servicio	.000		.000	.003	.032	.000
	Web planeación estratégica	.000	.000		.125	.366	.000
	Web imagen	.309	.003	.125		.000	.023
	Web calidad	.222	.032	.366	.000		.039
	Web desempeño	.000	.000	.000	.023	.039	

Nota:   Valores  $\geq 0.3$  ; X.- Valor cercano a 0.3

Fuente: SPSS 20 IBM



- i. **La diagonal de la matriz de correlación anti-imagen debe ser superior a 0.5 CUMPLE CONDICION. Ver Figura 12.24.**

**Figura 12.24. De resultados del análisis factorial**

**Matrices anti-imagen**

		Web tecnología	Web precio del servicio	Web planeación estratégica	Web imagen	Web calidad	Web desempeño
Covarianza anti-imagen	Web tecnología	.629	.047	-.210	-.046	-.022	.208
	Web precio del servicio	.047	.650	.190	-.078	.013	-.162
	Web planeación estratégica	-.210	.190	.613	.037	-.038	.092
	Web imagen	-.046	-.078	.037	.358	-.281	-.012
	Web calidad	-.022	.013	-.038	-.281	.372	-.046
	Web desempeño	.208	-.162	.092	-.012	-.046	.628
Correlación anti-imagen	Web tecnología	.721 <sup>a</sup>	.074	-.338	-.098	-.046	.331
	Web precio del servicio	.074	.787 <sup>a</sup>	.301	-.161	.027	-.253
	Web planeación estratégica	-.338	.301	.749 <sup>a</sup>	.079	-.079	.149
	Web imagen	-.098	-.161	.079	.542 <sup>a</sup>	-.769	-.025
	Web calidad	-.046	.027	-.079	-.769	.532 <sup>a</sup>	-.096
	Web desempeño	.331	-.253	.149	-.025	-.096	.779 <sup>a</sup>

a. Medida de adecuación muestral

Nota:   Valores >0.5

Fuente: SPSS 20 IBM

- j. **El test de esfericidad de Bartlett debe ser significativo (0.01 ó 0.05 error / 99% o 95% confiabilidad)...Si es significativa al 0.01...CUMPLE**
- k. **El KMO debe ser > 0.5 (lo ideal es > a 0.7)....SÍ CUMPLE . Ver Figura 12.25.**

**Figura 12.25. De resultados del análisis factorial**

**KMO y prueba de Bartlett**

Medida de adecuación muestral de Kaiser-Meyer-Olkin.	<span style="border: 1px solid red; padding: 0 5px;">.665</span>
Prueba de esfericidad de Bartlett	205.902
gl	15
Sig.	<span style="border: 1px solid red; padding: 0 5px;">.000</span>

- l. **Se concluye que es correcto y es posible continuar el análisis.**

Fuente: SPSS 20 IBM

**Paso 4: Estimación y ajuste. Continuación del análisis.**

- **Método de obtención de los Factores: Componentes Principales**, para resumir la mayor parte de la información en un menor número de factores.
- **¿Cuántos factores van a ser extraídos?:**
  - Autovalores mínimo de 1
  - En la **Figura 12.26**, se aprecian los valores a 1; la variable que **más información perdió** es la Web precio del servicio

**Figura 12.26. Comunalidades**

Comunalidades		
	Inicial	Extracción
Web tecnología	1.000	.658
Web precio del servicio	1.000	.580
Web planeación estratégica	1.000	.646
Web imagen	1.000	.882
Web calidad	1.000	.872
Web desempeño	1.000	.616

Método de extracción: Análisis de Componentes principales.

Fuente: SPSS 20 IBM

- En la **Figura 12.27**. Se observa la **creación de 2 grupos** , cuya varianza conjunta es del **70.879% >60% o 71% de explicación**, (Porcentaje de la varianza explicada **no menor del 60%**). sobre la percepción que tienen los gerentes de las empresas de mercadotecnia digital.

**Figura 12.27. Varianza total explicada**

Componente	Varianza total explicada								
	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción			Suma de las saturaciones al cuadrado de la rotación		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2.513	41.885	41.885	2.513	41.885	41.885	2.370	39.499	39.499
2	1.740	28.994	70.879	1.740	28.994	70.879	1.883	31.380	70.879
3	.598	9.959	80.838						
4	.530	8.830	89.667						
5	.416	6.928	96.595						
6	.204	3.405	100.000						

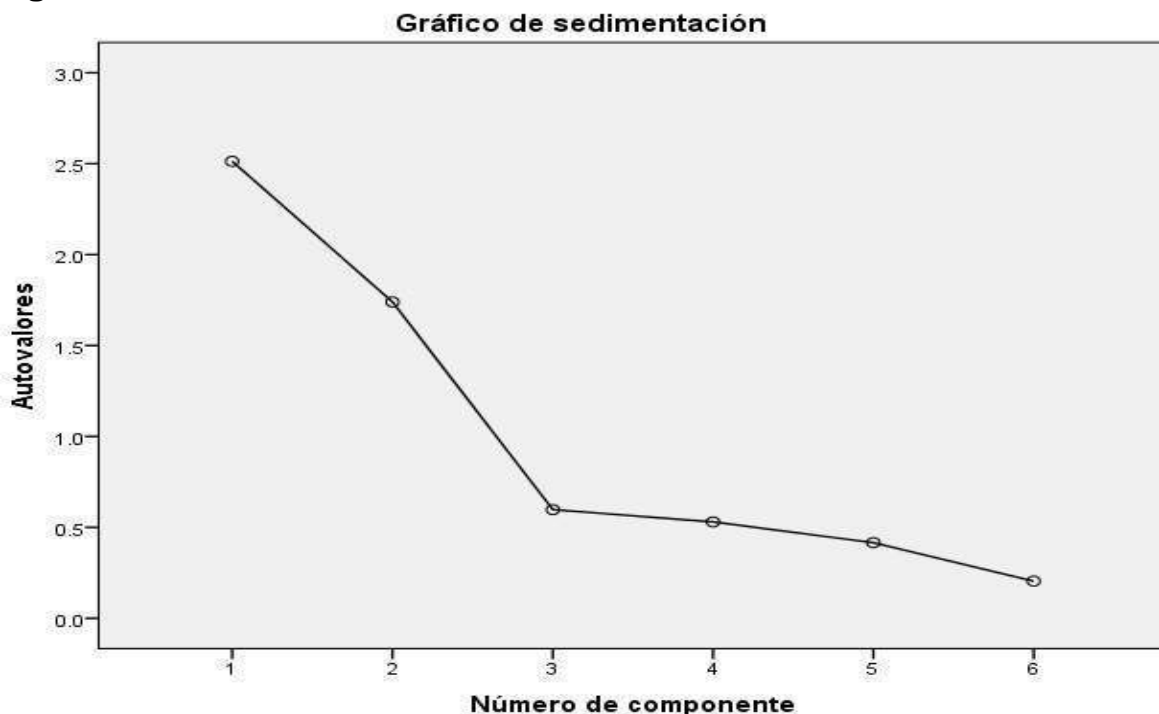
Método de extracción: Análisis de Componentes principales.

Fuente: SPSS 20 IBM

- **Análisis de gráfico de sedimentación**

Al analizar, se observa que la formación de grupos es determinado por las pendientes que existen entre los nodos siendo representativos los 2 primeros y la baja posibilidad de que se forme un tercer grupo, dado que la pendiente tiende a ser cero. Sin embargo, lo que se pretende es que la menor cantidad de grupos me explique más del 60% de la varianza extraída. Ver **Figura 12.28**.

**Figura 12.28. Gráfico de sedimentación**



Fuente: SPSS 20 IBM

### **Paso 5: Interpretación**

La **Matriz de componentes** reporta los valores que están listos a rotar. La **Matriz de componentes rotados** reporta el método de extracción, rotación y las iteraciones de convergencia (en nuestro caso **3/25**). Se deberá cuestionar ¿qué variables son de cada grupo? Como se recordará, la variable **X5** al no tomarse en cuenta, **NO** significa que no sirva, sino que **es tan parecida a los 2 grupos creados que no determina a cual asignarlo**. Se considera común a los grupos creados. Así, **el resto de las variables se clasificarán en el grupo que les corresponde de acuerdo a aquel que tenga el valor absoluto más alto**.

- ¿Qué variables explican qué factor?
- La rotación ayuda, pues fuerza a que las variables tengan una correlación cercana a **1** con un factor y cercana a **0** con los demás.
- **No hay una rotación menor que otra**. Lo ideal es la que mejor permita interpretar los factores.
- Analizar la **comunalidad** para ver qué varianza de cada variable recoge los factores.

- Analizar las cargas factoriales para interpretar los factores.
- Explicar sólo los factores con las variables que tengan una carga factorial importante. Ver **Figura 12.29 y 12.30**.

**Figura 12.29. Matriz de componentes**

**Matriz de componentes<sup>a</sup>**

	Componente	
	1	2
Web tecnología	-.627	.514
Web precio del servicio	.759	-.068
Web planeación estratégica	-.730	.336
Web imagen	.494	.799
Web calidad	.424	.832
Web desempeño	.767	-.167

Método de extracción: Análisis de componentes principales.

a. 2 componentes extraídos

**Matriz de componentes rotados<sup>a</sup>**

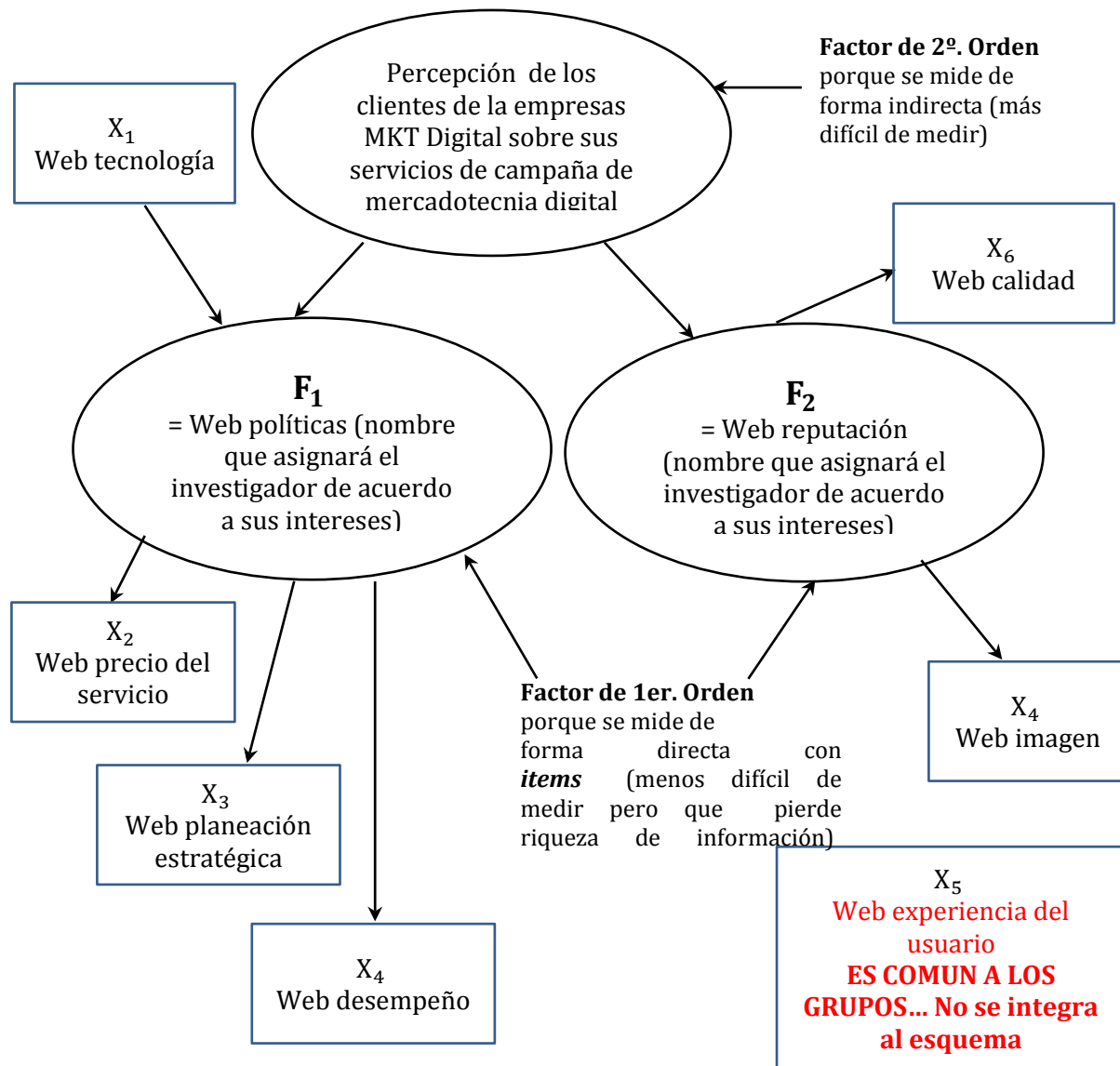
	Componente	
	1	2
Web tecnología	<b>-.787</b>	.194
Web precio del servicio	<b>.714</b>	.265
Web planeación estratégica	<b>-.803</b>	-.011
Web imagen	.102	<b>.933</b>
Web calidad	.025	<b>.934</b>
Web desempeño	<b>.764</b>	.179

Método de extracción: Análisis de componentes principales.

Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 3 iteraciones.

**Figura 12.30. Modelo final del análisis factorial exploratorio**



Fuente: propia

### Discusión

- Se genera un nuevo modelo con reagrupación de variables (**items**) y que en vez de ser explicado con 7 variables se explica con 2 nuevos factores y 6 variables. Se excluye 1 variable que no aporta valor al modelo
- Los 2 nuevos factores generados reciben el nombre que el investigador decida de acuerdo a sus intereses del proyecto
- Los factores finales se dividen en : **factores de 2º. Orden**, dado que se mide de forma **indirecta** (la literatura sugiere evitar trabajar con este tipo de factores ya que es más difícil de medir, sin embargo, es capaz de verificar relaciones de estructura, dado que permite ver qué variable es más importante de los factores de 1er orden y/o cuál de éstos aporta más al modelo) y **factores de 1er. Orden**

porque se mide de forma **directa** con **items** (la literatura sugiere su uso ya que es menos difícil de medir pero, pierde riqueza de información)

- **Nota:** Si se deseara un factor de 1er. Orden volverlo variable (**item**) haciendo una **media** de sus variables para incorporarlo al modelo de factor de 1er. Orden a variable. Esto haría que el factor de 2º. Orden se vuelva incluso de 1er Orden.
- La variable **X5** se vuelve **común (no se puede agregar)** a los **2 grupos y No se incluye en el modelo final, se considera que es básico, vital, primordial que brinde. Además, No puede haber factores con una sola variable (incorrelacionada)**, si se deseara tenerlo como constructo adicional. Lo que se elimina, es muy importante su discusión en el artículo a elaborar. Pudiera X5 incluirse como factor 2º. Orden midiéndose en términos de % NO con escala de **Likert**. Posible estudiar con ecuaciones estructurales en 2º. Orden (estudios aún escasos al momento)
- **Posible publicar en revistas CONACyT**
- Por análisis comparativo, se tienen resultados similares de cada una de las mitades respecto al modelo general.
- Lo anterior se refleja en : matriz de correlaciones, **KMO** y prueba de **Bartlett**, matriz anti-imagen, gráfico de sedimentación, comunalidades, varianza total explicada y matriz de componentes rotados Por lo tanto el modelo se presenta con mínimo error.
- Es posible aplicar ecuaciones estructurales para dar mayor explicación a las variables subyacentes,
- **Posible publicar en revistas CONACyT como estudio empírico.**
- **Posibles títulos:**
  - Construcción de un modelo empírico de percepción de los gerentes...**
  - Construcción de un modelo teórico es 100% académico**
- **Siguiente paso: buscar la teoría que justifique las nuevas relaciones los nuevos factores encontrados ya que estadísticamente está bien.**

#### **Paso 6: Validación**

- Para validar los resultados de un análisis factorial pueden emplearse dos métodos principales:
  - Uno de ellos es llevar a cabo un **análisis factorial confirmatorio** , mediante sistemas de ecuaciones estructurales (LISREL, AMOS, EQS) lo que está fuera del objetivo de este curso).
  - Otro procedimiento puede ser **separar la muestra en dos mitades aleatoriamente y llevar a cabo un análisis factorial con cada una de ellas.**
- Si el análisis de las cargas factoriales no difiere sustancialmente, podemos concluir que **los resultados son robustos y estables.**



## Paso 1: Objetivos

**Problema 6:** Validar el modelo resultante con el procedimiento de separar la muestra en dos mitades aleatoriamente y llevar a cabo un análisis factorial con cada una de ellas

**Paso 2: Diseño; Paso 3: Condiciones de aplicabilidad.** Se consideran que cumplen al caso.

## Paso 4: Ejecución y ajuste

**Teclar: Datos->Seleccionar casos->Seleccionar: muestra aleatoria de casos->Ejemplo->Tamaño de muestra: 50%->Continuar->Aceptar.** Ir a vista de datos para comprobar (se observaran marcas diagonales en el extremo izquierdo de la pantalla y un nuevo campo: **filter\$**, con valores: 1/0). Ver **Figura X. Proceso para validar resultados de análisis factorial por el proceso de las 2 mitades .Ver Figura 12.31.**

**Figura 12.31. Proceso para validar resultados de análisis factorial por el proceso de las 2 mitades**

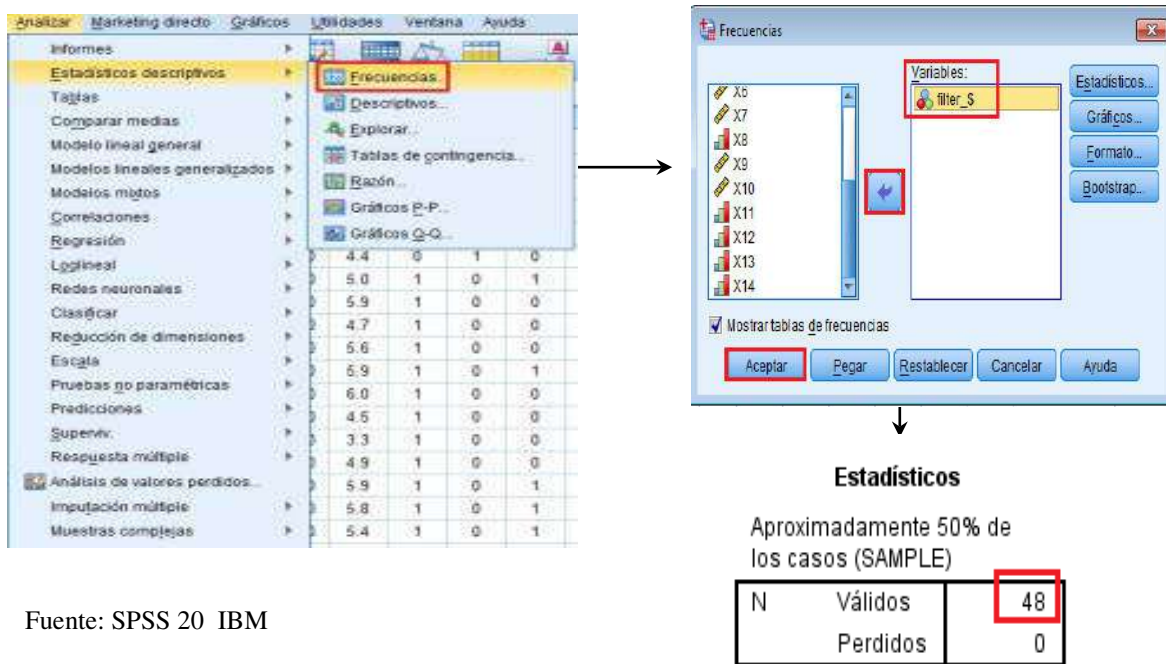
The figure illustrates the process of selecting a random sample in SPSS. It shows three main components:

- Top Left:** The 'Datos' menu with 'Seleccionar casos...' highlighted.
- Top Right:** The 'Seleccionar casos' dialog box with 'Muestra aleatoria de casos' selected.
- Middle Right:** The 'Seleccionar casos: Muestra aleatoria' dialog box with 'Aproximadamente 50 % de todos los casos' selected.
- Bottom Left:** A data view showing a table with columns ID, X1 through X14, and a new column 'filter\_\$' with values 1 or 0. The first 10 rows have 'filter\_\$' = 1, and the remaining 10 rows have 'filter\_\$' = 0.
- Bottom Right:** The 'Seleccionar casos' dialog box with 'Aceptar' highlighted.

ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	filter_\$
1	1	4.1	6	6.9	4.7	2.4	2.4	5.2	0	32.0	4.2	1	0	1	1
2	5	6.0	9	9.6	7.8	3.4	4.6	4.5	0	58.0	6.8	1	0	1	3
3	7	4.6	2.4	9.5	6.6	3.5	4.6	7.6	0	46.0	5.8	1	0	1	1
4	9	5.5	1.6	9.4	4.7	3.5	3.0	7.6	0	63.0	5.4	1	0	1	3
5	12	3.9	2.2	9.1	4.6	3.0	2.5	8.3	0	47.0	5.0	1	0	1	2
6	13	2.8	1.4	8.1	3.8	2.1	1.4	6.6	1	39.0	4.4	0	1	0	1
7	14	3.7	1.5	8.6	5.7	2.7	3.7	6.7	0	38.0	5.0	1	0	1	1
8	15	4.7	1.3	9.9	6.7	3.0	2.6	6.8	0	54.0	5.9	1	0	0	3
9	16	3.4	2.0	9.7	4.7	2.7	1.7	4.8	0	49.0	4.7	1	0	0	3
10	18	4.9	1.8	7.7	4.3	3.4	1.5	5.9	0	40.0	5.6	1	0	0	2
11	19	5.3	1.4	9.7	6.1	3.3	3.9	6.8	0	54.0	5.9	1	0	1	3
12	20	4.7	1.3	9.9	6.7	3.0	2.6	6.8	0	56.0	6.0	1	0	0	3
13	21	3.3	9	8.6	4.0	2.1	1.8	6.3	0	41.0	4.5	1	0	0	2
14	22	3.4	4	8.3	2.5	1.2	1.7	5.2	0	35.0	3.3	1	0	0	1
15	25	5.1	1.4	8.7	4.8	3.3	2.6	3.8	0	49.0	4.9	1	0	0	2
16	26	4.6	2.1	7.9	5.8	3.4	2.8	4.7	0	49.0	5.9	1	0	1	3
17	28	5.2	1.3	9.7	6.1	3.2	3.9	6.7	0	54.0	5.8	1	0	1	3
18	29	3.5	2.8	9.9	3.5	3.1	1.7	5.4	0	49.0	5.4	1	0	1	3
19	33	5.2	2.0	9.3	5.9	3.7	2.4	4.6	0	60.0	6.1	1	0	0	3
20	35	2.4	1.0	7.7	3.4	1.7	1.1	6.2	1	35.0	4.1	0	1	0	1

- Si se requiere saber cuántos datos son del filter\$=0 y/o filter\$=1, **Teclear: Analizar->Estadísticos descriptivos->Frecuencias->Seleccionar: filter\_\$ con flecha->Aceptar.** Se revisa la tabla de estadísticos en la que se reporta en nuestro caso, la división aproximada de la muestra como: **48%** (1s, es decir, sin rayar ) y **52%** (0s, es decir con rayar) por lo tanto como restante. Ver Figura 12.32. Proceso de conteo de la división aproximada de la muestra.

Figura 12.32. Proceso de conteo de la división aproximada de la muestra



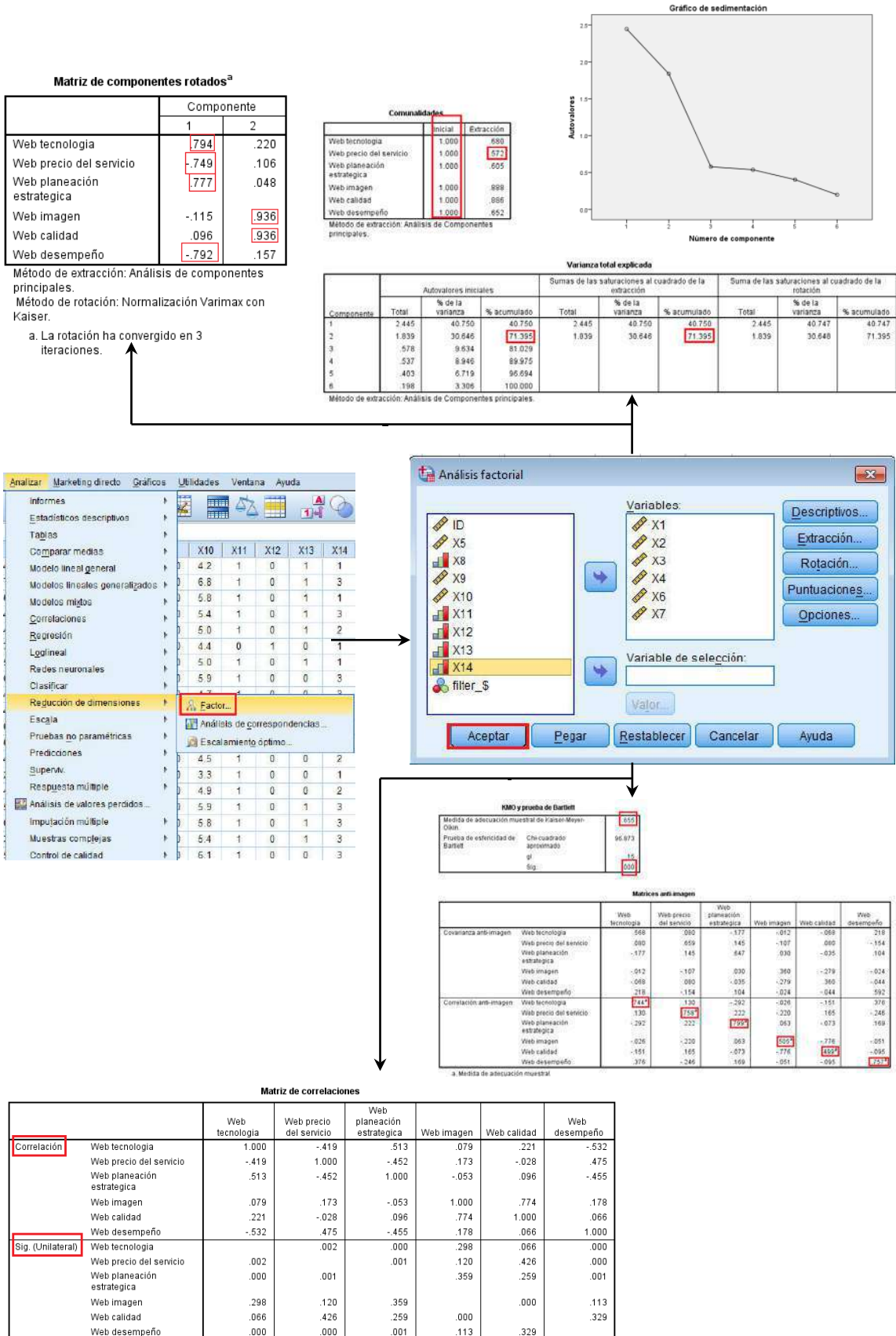
Fuente: SPSS 20 IBM

Realizar el proceso de análisis factorial nuevamente pero, con una de las muestras. **Teclear: Analizar->Reducción de dimensiones->Selección de variables métricas->Descriptivos-> Seleccionar: Estadísticos: Solución inicial; Matriz de correlaciones: Coeficientes; Niveles de significación-> KMO y prueba de esfericidad de Bartlett; Anti-imagen->Continuar->Extracción->Método: componentes principales->Analizar: Matriz de correlaciones->Visualización: solución factorial sin rotar-> Seleccionar: Gráfico de sedimentación->Extraer: Autovalores mayores que: 1->Número máximo de interacciones: 25->Continuar->Rotación->Método: Varimax->Visualización: solución rotada->No. máximo de iteraciones para convergencia: 25->Continuar->Puntuaciones->Continuar->Opciones->Valores perdidos (nota: no reemplazar por medias; no pérdida de datos >10%->Excluir casos según lista (nota: evitar falta de datos)->Continuar->Aceptar**

- **Nota:** Los valores con muestra filter\_\$=1/0 deberán **ser similares** a la muestra total. No considerar la variable X5 excluída previamente. Ver Figura 12.33



Figura 12.33. Proceso para validar resultados de análisis factorial con la primera mitad: filter\_\$=1.



- **Teclar: Datos->Seleccionar casos->Seleccionar: si se satisface la siguiente condición->Seleccionar filter\_\$->flecha->filter\_\$=0->Continuar->Aceptar.** Ir a visor de datos para comprobar rayado de extrema izquierda en los datos y filter\_\$= 0 rayados/1 sin rayar. Ver **Figura 12.34**.

**Figura 12.34. Proceso para validar resultados de análisis factorial con la otra mitad: filter\_\$=0**

The process is shown in four stages:

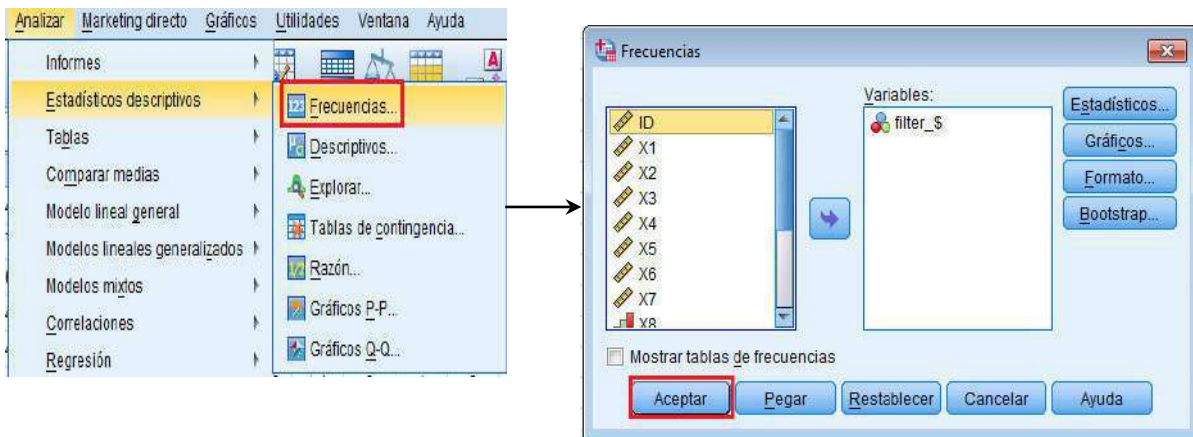
- Menu Selection:** The 'Datos' menu is open, and 'Seleccionar casos' is highlighted.
- Dialog Box:** The 'Seleccionar casos' dialog box is shown. Under 'Seleccionar', the option 'Si se satisface la condición' is selected. The condition 'filter\_\$=0' is entered in the text box. The 'Resultado' section has 'Descartar casos no seleccionados' selected.
- Confirmation:** The 'Si la opción' dialog box is shown, displaying a numeric keypad. The 'Continuar' button is highlighted.
- Data Grid:** The main SPSS window shows a data grid with columns ID, X1 through X14, and filter\_\$. Rows where filter\_\$ is 0 are highlighted in blue.

1. filter_\$	ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	filter_\$
0	1	4.1	6	6.9	4.7	2.4	2.4	5.2	0	32.0	4.2	1	0	1	1	0
0	2	5	6.0	9	9.6	7.8	3.4	4.6	4.5	0	58.0	6.8	1	0	1	3
0	3	7	4.6	2.4	9.5	6.6	3.5	4.5	7.6	0	46.0	5.8	1	0	1	1
0	4	9	5.5	1.6	3.4	4.7	3.5	3.0	7.6	0	63.0	5.4	1	0	1	3
0	5	12	3.9	2.2	9.1	4.6	3.0	2.5	8.3	0	47.0	5.0	1	0	1	2
1	6	13	2.8	1.4	8.1	3.8	2.1	1.4	6.6	1	39.0	4.4	0	1	0	1
1	7	14	3.7	1.5	8.6	5.7	2.7	3.7	6.7	0	38.0	5.0	1	0	1	1
1	8	15	4.7	1.3	9.9	6.7	3.0	2.6	6.8	0	54.0	5.9	1	0	0	3
1	9	16	3.4	2.0	9.7	4.7	2.7	1.7	4.8	0	49.0	4.7	1	0	0	3
1	10	18	4.9	1.8	7.7	4.3	3.4	1.5	5.9	0	40.0	5.6	1	0	0	2
1	11	19	5.3	1.4	9.7	6.1	3.3	3.9	6.8	0	54.0	5.9	1	0	1	3
1	12	20	4.7	1.3	9.9	6.7	3.0	2.6	6.8	0	55.0	6.0	1	0	0	3
1	13	21	3.3	9	8.6	4.0	2.1	1.8	6.3	0	41.0	4.5	1	0	0	2
1	14	22	3.4	4	8.3	2.5	1.2	1.7	5.2	0	35.0	3.3	1	0	0	1
0	15	25	5.1	1.4	8.7	4.8	3.3	2.6	3.8	0	49.0	4.9	1	0	0	2
0	16	26	4.6	2.1	7.9	5.8	3.4	2.8	4.7	0	49.0	5.9	1	0	1	3
0	17	28	5.2	1.3	9.7	6.1	3.2	3.9	6.7	0	54.0	5.8	1	0	1	3
0	18	29	3.5	2.8	9.9	3.5	3.1	1.7	5.4	0	49.0	5.4	1	0	1	3
1	19	33	5.2	2.0	9.3	5.9	3.7	2.4	4.6	0	60.0	6.1	1	0	0	3

Fuente: SPSS 20 IBM

- Para verificar descriptivos, Teclear: Datos->Seleccionar casos->Seleccionar: si se satisface la siguiente condición->Seleccionar filter\_\$->flecha->filter\_\$=0->Continuar->Aceptar. Ir a visor de datos para comprobar rayado de extrema izquierda en los datos y filter\_\$= 0 rayados/1 sin rayar. Resultados: 52 casos, el resto es de 1s. Ver Figura 12.35.

Figura 12.35. Proceso para verificar las frecuencias de la segunda mitad filter\_\$=0



Estadísticos

filter\_\$=0 (FILTER)

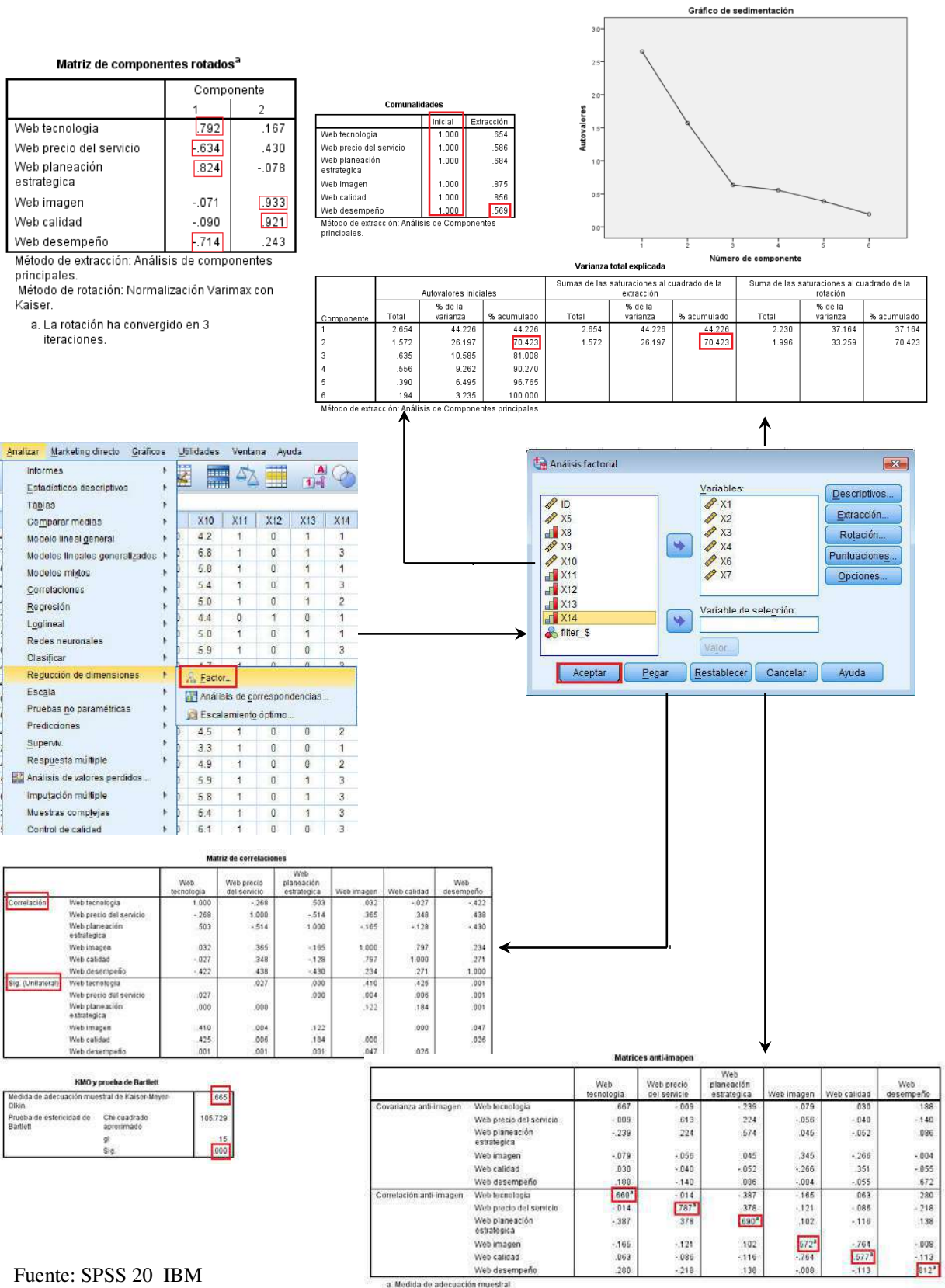
N	Válidos	52
	Perdidos	0

Fuente: SPSS 20 IBM

Realizar el proceso de análisis factorial nuevamente pero, con una de las muestras. Teclear: Analizar->Reducción de dimensiones->Selección de variables métricas->Descriptivos-> Seleccionar: Estadísticos: Solución inicial; Matriz de correlaciones: Coeficientes; Niveles de significación-> KMO y prueba de esfericidad de Bartlett; Anti-imagen->Continuar->Extracción->Método: componentes principales->Analizar: Matriz de correlaciones->Visualización: solución factorial sin rotar-> Seleccionar: Gráfico de sedimentación->Extraer: Autovalores mayores que: 1->Número máximo de interacciones: 25->Continuar->Rotación->Método: Varimax->Visualización: solución rotada->No. máximo de iteraciones para convergencia: 25->Continuar->Puntuaciones->Continuar->Opciones->Valores perdidos (nota: no reemplazar por medias; no pérdida de datos >10%->Excluir casos según lista (nota: evitar falta de datos)->Continuar->Aceptar. Ver Figura 12.36.



**Figura 12.36. Proceso para validar resultados de análisis factorial con la primera mitad: filter\_\$=0**



Fuente: SPSS 20 IBM

## Paso 5: Interpretación

Por análisis comparativo, se tienen resultados similares de cada una de las mitades respecto al modelo general.

Lo anterior se refleja en : matriz de correlaciones, **KMO** y prueba de **Bartlett**, matriz anti-imagen, gráfico de sedimentación, comunalidades, varianza total explicada y matriz de componentes rotados Por lo tanto el modelo se presenta con mínimo error. Es posible aplicar ecuaciones estructurales para dar mayor explicación a las variables subyacentes,

Posible publicar en revistas CONACyT como estudio empírico.

Posibles títulos:

-Construcción de un modelo empírico de percepción de los gerentes...

-Construcción de un modelo teórico es 100% académico

Siguiente paso: buscar la teoría que justifique las nuevas relaciones los nuevos factores encontrados ya que estadísticamente está bien.

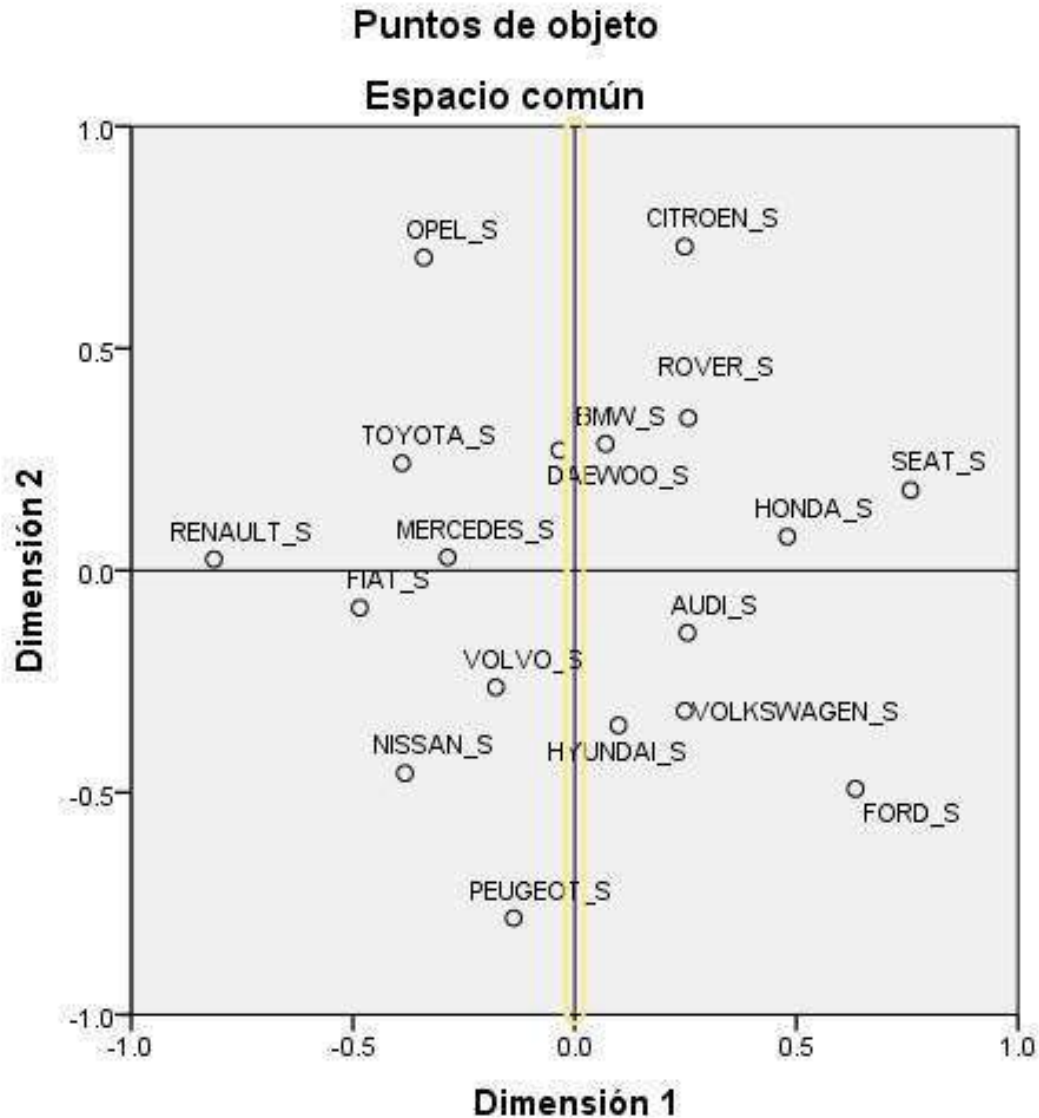
## Referencias

- Anderson, J. C., Gerbing, D. W. y Hunter J. E. (1987), On the Assessment of Unidimensional Measurement: Internal and External Consistency and Overall Consistency Criteria. *Journal of Marketing Research* 24 (November): 432-37.
- American Psychological Association (1985), *Standards for Educational and Psychological Tests*. Washington, D.C.: APA.
- Bearden, W. O., Netemeyer, R. G. y Mobic M. (1993), *Handbook of Marketing Scales: Multi-Item Measures for Marketing and Consumer Behavior*. Newbury Park, Calif.: Sage.
- Bentler, Peter M. (1992), *EQS Structural Equations Program Manual*. Los Angeles: BMDP Statistical Software.
- BMDP Statistical Software, Inc. (1992), *BMDP Statistical Software Manual*, Release 7, vols. 1 and 2, Los Angeles: BMDP Statistical Software.
- Borgatta, E. F., Kercher, K. y Stull D. E. (1986), A Cautionary Note on the Use of Principal Components Analysis. *Sociological Methods and Research* 15:160-68.
- Bruner, G. C., y Hensel P. J. (1993). *Marketing Scales Handbook, A Compilation of Multi-Item Measures*. Chicago: American Marketing Association.
- Campbell, D. T., y Fiske D. W. (1959). Convergent and Discriminant Validity by the Multitrait Multimethod Matrix. *Psychological Bulletin* 56 (March): 81-105.
- Cattell, R. B. (1966), The Scree Test for the Number of Factors. *Multivariate Behavioral Research* 1 (April): 245-76.
- Cattell, R. B., Baeur, K. R. Horn, J. L. y Nesselroade, J. R. (1969), Factor Matching Procedures: An Improvement of the s index; with tables. *Educational and Psychological Measurement* 29: 781-92.
- Chatterjee, S., Jamieson, L. y Wiseman F. (1991), Identifying Most Influential Observations in Factor Analysis. *Marketing Science* 10 (Spring): 145-60.
- Churchill, G. A. (1979), A Paradigm for Developing Better Measures of Marketing Constructs. *Journal of Marketing Research* 16 (February): 64-73.

- Cliff, N., y Hamburger, C. D. (1967), The Study of Sampling Errors in Factor Analysis by Means of Artificial Experiments. *Psychological Bulletin* 68: 430-45.
- Cronbach, L. J. (1951), Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 31: 93-96.
- Dillon, W. R. y Goldstein M. (1984), *Multivariate Analysis: Methods and Applications*. New York: Wiley.
- Dillon, W. R., Mulani, N. y Frederick, D. G. (1989), The Use of Component Scores in the Presence of Group Structure. *Journal of Consumer Research* 16:106-12.
- Gorsuch, R. L. (1983), Factor Analysis. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Gorsuch, R. L. (1990), Common Factor Analysis versus Component Analysis: Some Well and Little Known Facts. *Multivariate Behavioral Research* 25: 33-39.
- Hair, J.F.; Anderson, R.E.; Black, W.C. (1999). *Análisis Multivariante*. 5a. Ed. España: Prentice Hall.
- Hattie, J. (1985), Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement* 9: 139-64.
- IBM (2011a). *IBM SPSS Statistics Base 20*. EUA. Industrial Business Machines. Recuperado el 20161201 de: [ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM\\_SPSS\\_Statistics\\_Base.pdf](ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Base.pdf)
- IBM (2011b). *Guía breve de IBM SPSS Statistics 20*. EUA. Industrial Business Machines. Recuperado el 20161201 de: [ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM\\_SPSS\\_Statistics\\_Brief\\_Guide.pdf](ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf)
- IBM (2011c). *IBM SPSS Missing Values 20*. EUA. Industrial Business Machines. Recuperado el 20161201 de: [ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM\\_SPSS\\_Missing\\_Values.pdf](ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Missing_Values.pdf)
- Joreskog, K. G., y Sorbo, D. (1993), *LISREL 8*. Mooresville, Ind.: Scientific Software International.
- Kaiser, H. F. (1970), A Second-Generation Little Jifty. *Psychometrika* 35: 401-15.
- Kaiser, H. F. (1974), Little Jifty, Mark IV. *Educational and Psychology Measurement* 34: 111-17.
- McDonald, R. P. (1981), The Dimensionality of Tests and Items. *British Journal of Mathematical and Social Psychology* 34: 100-117.
- Mulaik, S. A. (1990), Blurring the Distinction Between Component Analysis and Common Factor Analysis. *Multivariate Behavioral Research* 25: 53-59.
- Mulaik, S. A., y McDonald R. P. (1978), The Effect of Additional Variables on Factor Indeterminacy in Models with a Single Common Factor. *Psychometrika* 43: 177-92.
- Nunnally, J. L. (1978), *Psychometric Theory*. 2d ed. New York: McGraw-Hill.
- Nunnally, J. (1979), *Psychometric Theory*. New York: McGraw-Hill.
- Peter, J. P. (1979), Reliability: A Review of Psychometric Basics and Recent Marketing Practices. *Journal of Marketing Research* 16 (February): 6-17.
- Peter, J. P. (1981), Construct Validity: A Review of Basic Issues and Marketing Practices. *Journal of Marketing Research* 18 (May): 133--45.

- Robinson, J. P., Shaver, P. R., y Wrightsrnan, L. S. (1991), *Criteria for Scale Selection and Evaluation, in Measures of Personality and Social Psychological Attitudes*, (eds.). San Diego, Calif.: Academic Press.
- Robinson, J. P., y Shaver P. R. (1973), *Measures of Psychological Attitudes*. Ann Arbor, MI: Survey Research Center Institute for Social Research, University of Michigan.
- Rurnrnel, R. J. (1970). *Applied Factor Analysis*. Evanston, Ill.: Northwestern University Press.
- SAS Institute, Inc. (1990). *SAS User 's Guide: Statistics, Version 6*. Cary, N.C.: SAS Institute.
- Srnith, Scott M. (1989), *PC-MDS: A Multidimensional Statistics Package*. Provo, Utah: Brigham Young University Press.
- Snook, S. C., y Gorsuch, R. L. (1989), Principal Cornponent Analysis versus Cornrnon Factor Analysis: A Monte Cario Study. *Psychological Bulletin* 106:148-54.
- Stewart, D. W. (1981), The Application and Misapplication of Factor Analysis in Marketing Research. *Journal of Marketing Research* 18 (February): 51-62.
- Velicer, W. F., y Jackson, D. N. (1990), Cornponent Analysis versus Cornrnon Factor Analysis: Some Issues in Selecting an Appropriate Procedure. *Multivariate Behavioral Research* 25: 1-28.

## Capítulo 13. Análisis Multidimensional y de Correspondencias



### 13.1.-El análisis multidimensional. ¿Qué es?

Tomando como base a la fuente de información, es decir, los encuestados, consiste en una serie de técnicas que le ayudarán a identificar los atributos (**dimensiones**) subyacentes claves en las evaluaciones de los objetos de estudio. Así, se depende que es una técnica muy utilizada en la administración de la mercadotecnia que identifica las **dimensiones subyacentes determinantes** en las evaluaciones de los productos/servicios, compañías, etc. por parte de los clientes. Otras aplicaciones habituales incluyen:

1. Comparativos de lo capturado de la encuesta, detectando diversos atributos desde gustos, calidades, desempeño, texturas, hasta la evaluación de diferencias

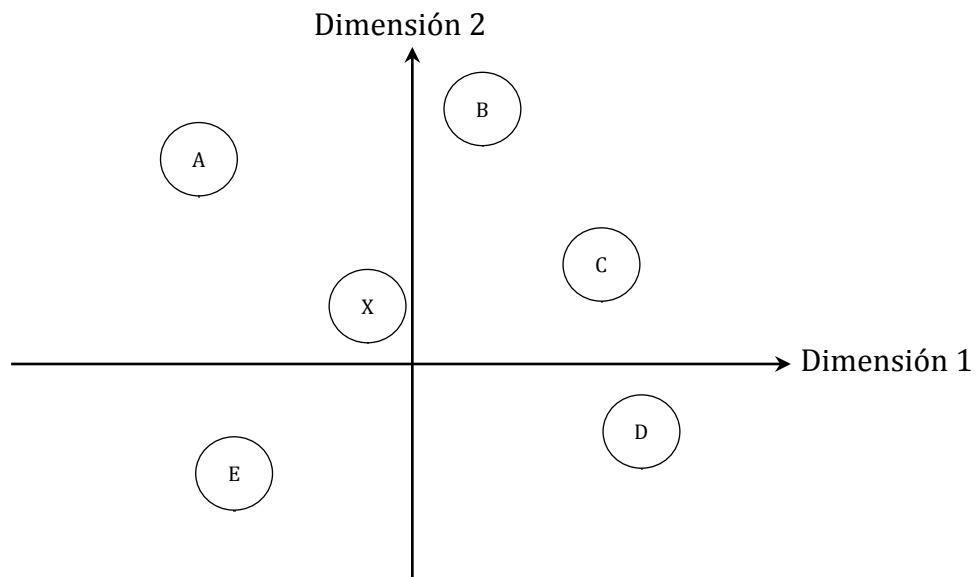


culturales entre diferentes grupos.

2. Obtenidos los datos, Usted podrá determinar: (a) cuáles dimensiones los encuestados usan al evaluar los objetos, (b) número de dimensiones pueden utilizarse en una situación particular, (c) establecer de cada dimensión, la importancia relativa y (d) relaciones perceptualmente de los objetos.

La técnica también es conocida en la elaboración de **mapas perceptuales**, que es un procedimiento que le permite definir la **imagen percibida relativa** de un conjunto de **objetos** (productos, servicios, empresas, etc.). La técnica tiene como objetivo, hacer la transformación de las percepciones emitidas por el consumidor por similitud o preferencia (por ejemplo, percepciones por diseño, colores, marcas, etc.) en un espacio multidimensional, referido a distancias. Por ejemplo, suponga que sus encuestados evalúan a los sujetos A y B, que los consideran lo más similares comparados con todos los posibles pares de objetos. Las técnicas de análisis multidimensional situarán los objetos A y B de tal forma que la distancia entre ellos en el espacio multidimensional es menor que la distancia entre cualquier otro par de objetos. El **mapa perceptual** resultante (o **mapa espacial**) muestra la situación relativa de todos los **objetos**, tal y como se muestra en la **Figura 13.1**.

**Figura 13.1. Ejemplo de mapa multidimensional**



Nota: percepción de 5 servicios de mercadotecnia digital y X, como percepción ideal.

Fuente: propia

Se afirma que, la técnica se basa en las comparaciones entre objetos. Se puede suponer que cualquier objeto (por ejemplo, producto/servicio, textura, color, etc.) **tiene dimensiones objetivas y percibidas**. Por ejemplo, la dirección de **MKT Digital** puede considerar su videojuego con dos opciones: sistema operativo IOS y Windows, en ambiente terrestre, aéreo, marino. Aunque son dimensiones objetivas, es posible que los clientes aprecien (o no) estos atributos. Los consumidores pueden percibir el videojuego de **MKT Digital** como **caro y difícil de manipular (dimensiones subjetivas)**. Se destaca que **2 productos /servicios** pueden tener las mismas características físicas (**dimensiones objetivas**), **pero ser apreciados de forma**

**diferente**, debido a que las distintas formas, colores, marcas, etc. se perciben de forma diferente, en calidad (**una dimensión percibida**) por los clientes. Dado lo anterior, se puede ver que es muy importante distinguir, como:

1. **Diferencias individuales:** dimensiones percibidas por los encuestados **vs.** las dimensiones objetivas propuestas por Usted, pueden NO coincidir con (o pueden no incluir).
2. **Interdependencia: las evaluaciones de las dimensiones** (incluso si las dimensiones objetivas son las mismas que las dimensiones percibidas) **pueden no concordar y pueden no ser independientes.** La interacción de ambas tienden a crear evaluaciones inesperadas. Por ejemplo, una forma de teléfono inteligente puede juzgarse más potente que otro, debido a que el primero tiene un diseño más que corresponde a los estereotipos del momento que otro, aunque contenga las mismas prestaciones de servicio.

Como se observa, entender las dimensiones percibidas y referirlas a dimensiones objetivas, si es posible y es el reto permanente de Usted como investigador. Así, necesitará de un análisis adicional para realizar evaluaciones de **qué atributos predicen la posición de cada objeto** en los 2 espacios: **objetivo y perceptual.** Dado que este proceso depende de hacer calificaciones muy agudas de las dimensiones (más arte que ciencia), por parte de los investigadores se recomienda precaución, en la interpretación de las dimensiones. Debido a que este proceso es más un arte que una ciencia, **Usted deberá resistirse a permitir que su percepción personal afecte a la dimensionalidad cualitativa de las dimensiones percibidas y ser lo más objetivo posible en esta crítica área.** Para saber más, ver IBM, 2011a; IBM, 2011b, IBM, 2011c.; IBM, 2012.

### 13.2. Análisis multidimensional. Cómo actúa

Para que tenga una idea de los alcances de la técnica, suponga que la empresa **MKT Digital**, le interesa comprender las percepciones de sus clientes de **6 videojuegos** que tiene en el mercado. Como estrategias para obtener información al respecto, se plantea:

1. Aplicar evaluaciones de los consumidores sobre cada uno de los videojuegos, por medio de un cierto número de atributos
2. Reunir sólo las percepciones de similitudes y diferencias conjuntas. Los datos se reúnen normalmente pidiendo a los encuestados respuestas globales a cuestiones como éstas:
  - Calificación de similitud de los productos A y B en una escala de 1 a 10.
  - El producto A es más similar a B que a C.
  - Prefiero el producto A al B.

Por la originalidad de la **segunda propuesta**, ésta es la que se acuerda seguir ya que se puede deducir el mapa perceptual, a partir de éste conjunto de respuestas que representan en una mejor medida, la similitudes entre los **6 videojuegos**. Sólo por fines de aprendizaje, describirá el **proceso de creación de un mapa perceptual** datos provenientes de **un sólo encuestado**, aunque este proceso debe aplicarse a varios encuestados o respuestas de un grupo de consumidores enfocado, mediante el siguiente proceso:

1. Sabiendo que  $N=6$  obtenga los datos a partir de definir un conjunto de **15 pares de videojuegos** a presentar a evaluación ( $N*(N-1)/2=6*5/2=15$  pares).
2. Solicitar a los encuestados, que califiquen los siguientes **15 pares de videojuegos**, donde una calificación de:
  - a. **1 se otorga al par de videojuegos que sea más parecido y**
  - b. **15 indica que el par es el menos parecido.** Ver Figura 13.2

**Figura 13.2. Tabla de datos por detección de nivel de similitud**

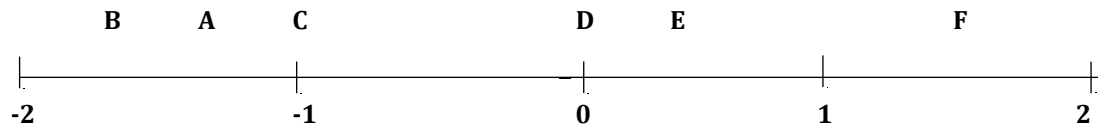
Videojuego	A	B	C	D	E	F
A	-					
B	2	-				
C	13	12	-			
D	4	6	9	-		
E	3	5	10	1	-	
F	8	7	11	14	15	-

Fuente: propia

En la **Figura 10.2** se muestran las calificaciones de todos los pares de videojuegos a un encuestado, quien considera, que los **videojuegos D y E son las más parecidas**, los videojuegos **A y B serían las siguientes más parecidas** y así sucesivamente hasta las videojuegos **E y F, que son las menos parecidas.**

3. De ilustrar la similitud entre los videojuegos gráficamente, en un primer intento se podría deducir **una sola escala de similitud y ajustar todos los videojuegos a esta escala.** Esto implica una **representación unidimensional de similitud**, donde **la distancia represente la similitud.** Por tanto, los objetos **cercanos** en la escala son más parecidos y **aquellos más alejados son menos parecidos**, por lo que el objetivo es determinar una escala de tal forma que las calificaciones estén mejor representadas para cada par de videojuegos en evaluación de manera tal que la cercanía de las calificaciones sea proporcional de manera gradual (por ejemplo, la calificación 1 es la más cercana, la calificación 2 es la siguiente más próxima, y así sucesivamente).
4. Haga prueba de colocación, por ejemplo, ubicar la primera prueba real llega con cuatro objetos, en este caso, videojuegos a elegir: A, B, C y D. La **Figura 10.2** muestra que el rango de orden de los pares es el siguiente: **BA(2) < DA(4) < DB(6) < DC(9) < CB(12) < CA(13)** (la línea que hay sobre cada par de letras indica que la expresión se refiere a la distancia "**similitud**" entre los pares). A partir de estos valores, debemos situar los cuatro videojuegos en una escala única, de tal forma que las más similares **BA(2)** sean las más cercanas y las menos similares **CA(13)** sean las más alejadas. La **Figura 13.3** contiene un **mapa perceptual unidimensional** que muestra cuando la persona que juzga la similitud entre los videojuegos, ha estado pensando en términos de una regla simple de similitud que contiene solo un atributo (**una dimensión**), tal como **el sistema operativo**, entonces todos los pares podrán ser situados **en una escala simple que reproduce los valores similares.**

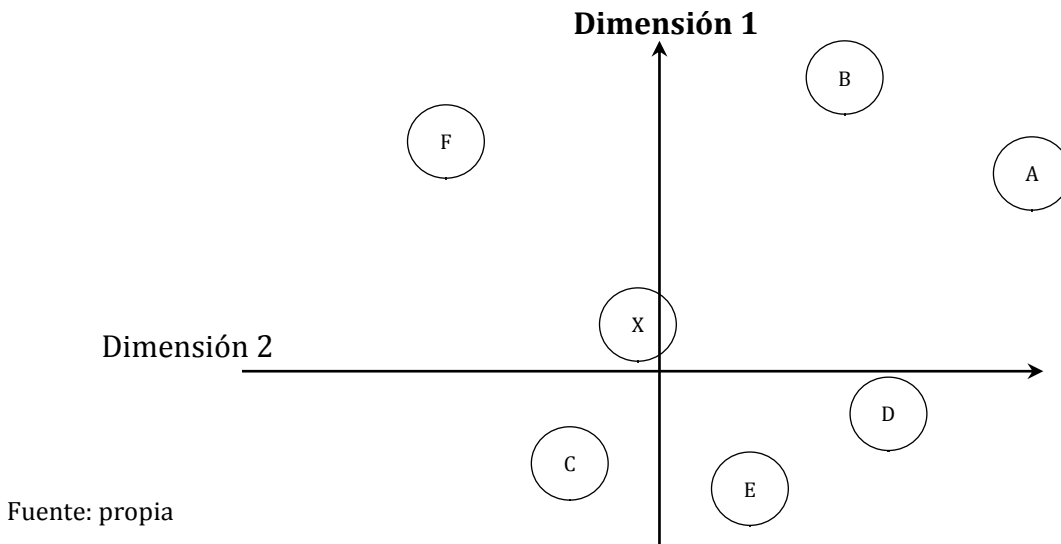
**Figura 13.3. Mapa perceptual unidimensional de 6 observaciones**



Fuente: propia

1. Dado que un análisis **unidimensional** no se ajusta bien a los datos, se debe intentar una solución **bidimensional**. Esto permite utilizar otra dimensión en la configuración de los **6 videojuegos**. Hacer a mano este procedimiento, suele ser de una gran inversión de tiempo y muy tedioso lograrlo; en la **Figura 13.4** se muestra la solución **bidimensional** producida típicamente por un software de análisis multidimensional. El resultado corresponde a las calificaciones de la **Figura 13.2** donde el encuestado usará con toda seguridad **2 dimensiones** para evaluar los videojuegos, ya que se considera que está basada en la **incapacidad de representar las percepciones de un encuestado en una dimensión**. Sin embargo, **todavía no somos conscientes de qué atributos utiliza el encuestado en la evaluación**. Aunque no tenemos información sobre cuáles son estas dimensiones, podemos ser capaces de **observar las posiciones relativas de los videojuegos e inferir qué atributos representan las dimensiones**. Por ejemplo, suponga que los videojuegos **A, B y F** fueran una forma de **combinación** de sistema operativo Windows-ambiente aéreo-control joystick botonera y los videojuegos **C, D y E** fueran estrictamente de sistema operativo IOS. Podríamos inferir a continuación que la dimensión **X** representa el tipo de videojuego (sistema operativo IOS vs **combinación**). Cuando observamos la posición de los videojuegos en la dimensión vertical, es posible que emerjan otros atributos como descriptores de la dimensión.

**Figura 13.4. Mapa perceptual bidimensional de seis observaciones**



Fuente: propia

El análisis multidimensional permite a los investigadores entender las similitudes entre los **6 videojuegos** tan sólo con **cuestionar sobre las percepciones de similitud**. A fin de **determinar qué atributos forman parte en realidad de las percepciones de similitud**. Aunque no se incorporan directamente las evaluaciones de los **atributos** en ésta técnica, podemos utilizarlos en análisis subsecuentes para ayudar a interpretar las dimensiones y los impactos que cada atributo tiene en las posiciones relativas de las cho- colatinas.

### **13.3. Análisis multidimensional vs. otras técnicas de enfoque interdependiente**

Puede compararse ésta técnica vs. otras técnicas de interdependencia (**análisis factorial y el análisis cluster**) en función de su aproximación a la estructura definida, del que se tiene:

1. **El análisis factorial** agrupa variables en **valores teóricos** que definen las **dimensiones subyacentes del conjunto original de variables**. Las variables que están muy correlacionadas se agrupan conjuntamente.
2. **El análisis cluster** agrupa observaciones de acuerdo con su perfil sobre un conjunto de variables (**valor teórico cluster**) en el cual las observaciones muy cercanas se agrupan juntas. Sin embargo, el análisis multidimensional difiere del análisis cluster en dos aspectos clave:

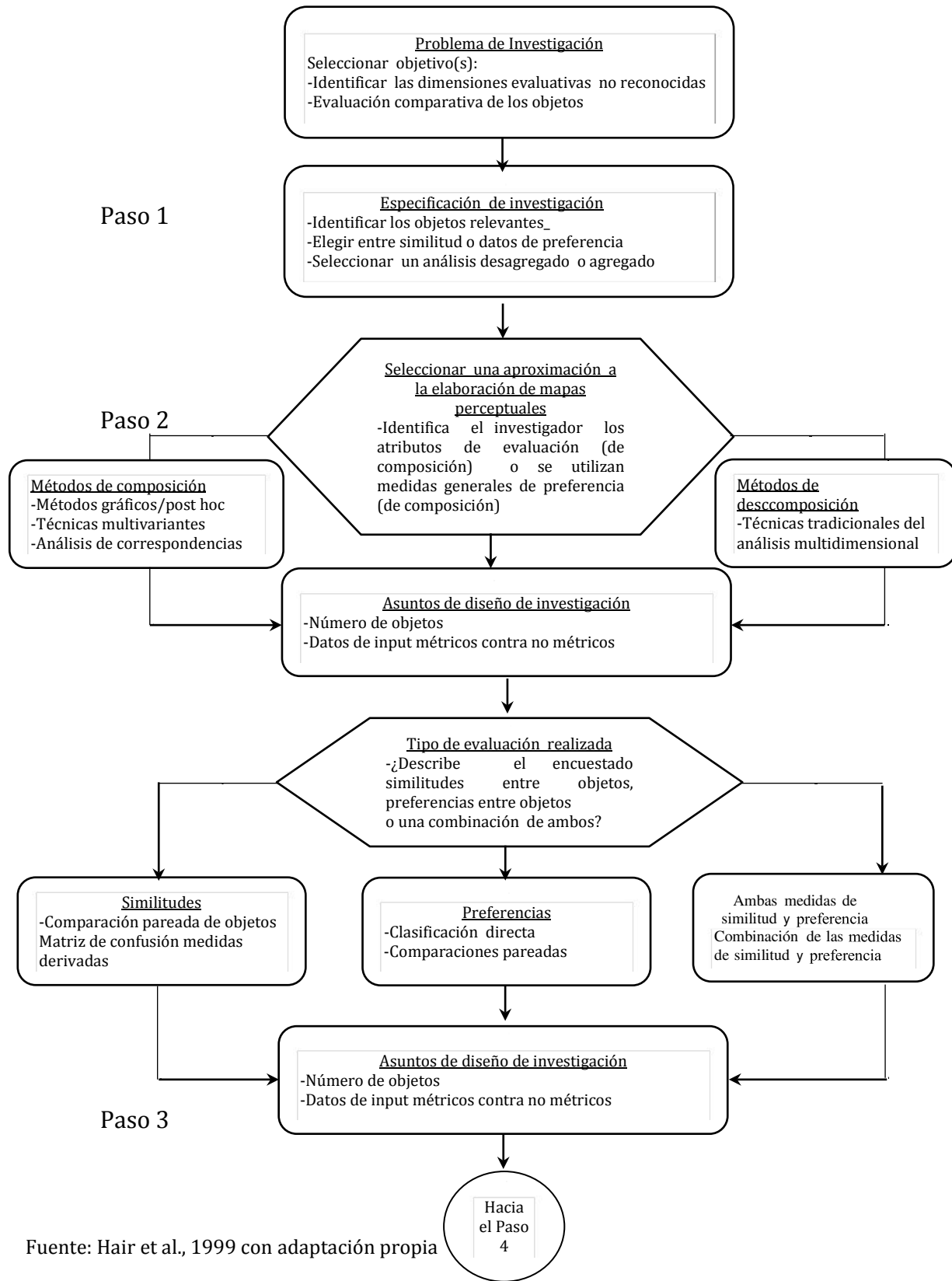
**-Se puede obtener una solución para cada individuo.** En la técnica, cada encuestado proporciona evaluaciones de todos los objetos que se están considerando, por lo que se puede obtener una solución para cada individuo, lo que no es posible para el análisis cluster o factorial. Como tal, el énfasis no se pone en los objetos mismos, sino en cómo el individuo percibe los objetos. La estructura que está siendo definida se basa en las dimensiones perceptuales de comparación de los individuos. Una vez que las dimensiones perceptuales están definidas, se pueden realizar las comparaciones relativas entre los objetos.

**-No utiliza el valor teórico.** El análisis multidimensional, a diferencia de otras

técnicas multivariantes, **NO** utiliza el valor teórico, sino por el contrario, las *“variables”* que compensarían al valor teórico (**las dimensiones perceptuales de comparación**) provienen **de inferir medidas globales de similitud entre objetos**. Haciendo una analogía simple, esto es como proporcionar la variable dependiente (**similitud entre objetos**) e imaginarse cuál debe ser la variable independiente (**dimensiones perceptuales**). La técnica tiene la ventaja de **reducir la influencia del investigador al no exigir la especificación de las variables que se utilizan en la comparación de objetos**,(situación que sí se presenta en el análisis cluster). Pero también tiene la desventaja de que **Usted no estará realmente seguro de qué variables está utilizando el encuestado para realizar las comparaciones**.

3. **La elaboración de mapas perceptuales** abarca varios métodos posibles, como ésta técnica la cual se puede visualizar en los **6 pasos** que hemos visto desde el **Capítulo 2**. Ver **Figuras 13. 5 y Figura 13.6**

**Figura 13.5.** Diagrama de flujo análisis multidimensional pasos 1-2-3



Fuente: Hair et al., 1999 con adaptación propia



#### 13.4. Análisis multidimensional. Paso 1: Objetivos

Se consiguen dos objetivos al aplicar la técnica con la determinación de los mapas perceptuales:

1. Lograr **aplicar una técnica exploratoria** que identifique las dimensiones no reconocidas que afectan al comportamiento.
2. Un medio de **obtener evaluaciones comparativas de objetos** cuando las bases específicas de comparación no se conocen o no están definidas.

En la técnica, **especificar los atributos de comparación de los encuestados NO es necesario**. Sólo lo que se requiere es la especificación de los objetos asegurándose de que **comparten una base común de comparación**. Esta flexibilidad hace a la técnica particularmente apropiada para realizar **estudios gráficos y de situación en los que las dimensiones de evaluación sean muy globales o muy emocionales y afectivas como para ser medidas por análisis convencionales**. Combinar la situación de objetos y sujetos en un único mapa conjunto, es la misión de algunos métodos de análisis multidimensional, lo que genera **posiciones relativas de los objetos y consumidores que producen, por tanto, un análisis de segmentación** mucho más directo, con las siguientes particularidades:

-**Decisiones clave en la fijación de objetivos. La falta de especificación en la definición de los niveles de evaluación de los objetos** es una característica común a cada objetivo. Hacer mapas perceptuales, tiene la gran ventaja de generar la habilidad para **“inferir” dimensiones sin la necesidad de definir atributos**. La flexibilidad y la naturaleza inferencial de la técnica, exigirá a Usted como investigador una mayor responsabilidad al momento de definir **“correctamente”** el análisis. Las **consideraciones prácticas y conceptuales son esenciales** para que la técnica consiga sus mejores resultados y, para asegurar este éxito, Usted debe definir un análisis multidimensional a través de **tres decisiones clave**:

- Seleccionar los **objetos que se van a evaluar**. Este punto, es el asunto más básico, y a la vez importante, en la elaboración de mapas perceptuales que se debe definir. Dado que el mapa perceptual es una técnica de situación relativa, Usted debe asegurarse que todas las empresas, productos, servicios u otros **objetos “relevantes”** estén incluidos. La relevancia está determinada por las preguntas que hace el investigador. Considere que la omisión o la inclusión de objetos inapropiados [Green 1975, Maholtra 1987]. Los mapas perceptuales resultantes de cualquier método pueden verse enormemente influenciados tanto por **la omisión como por la inclusión de objetos inapropiados** [Green 1975, Maholtra 1987]. ...Si se incluyen los **“objetos” irrelevantes o no comparables, el investigador está forzando la técnica no sólo para inferir las dimensiones perceptuales** que distinguen entre objetos comparables, sino también para **inferir aquellas dimensiones que distinguen entre objetos no comparables**. Se considera que **ésta tarea está más allá del alcance del análisis multidimensional y resulta en una solución que no concuerda con la realidad**.

-**Decidir si son las similitudes o las preferencias lo que se va analizar.**

Seleccionados los objetos de estudio, Usted debe elegir las bases de evaluación: **similitud frente a preferencia**. A manera de resumen, se ha discutido la elaboración de mapas perceptuales y análisis dimensional en términos de **juicios de similitud**. Al proporcionar datos de similitud, el **encuestado no aplica ningún aspecto de evaluación en términos de «bueno- malo» en la comparación**, la cual evalúa dentro de los datos de preferencia, que asumen **que diferentes combinaciones de atributos percibidos se valoran más que otras combinaciones**. Estas bases de comparación se pueden utilizar para desarrollar mapas perceptuales, pero **con diferentes interpretaciones**:

**-Los mapas perceptuales basados en la similitud representan similitudes de atributos y dimensiones perceptuales de comparación pero no reflejan ninguna visión directa en los determinantes de la elección.**

**-Los mapas perceptuales basados en la preferencia reflejan elecciones preferidas pero pueden no corresponder de ningún modo a posiciones basadas en la similitud**, dado que los encuestados pueden basar sus elecciones en dimensiones o criterios enteramente diferentes de aquellos en los que basan sus comparaciones. **No existe una base óptima de evaluación**, pero la decisión entre cuál aplicar debe realizarse teniendo en cuenta la investigación a realizar, ya que son fundamentalmente diferentes en lo que representan.

**-Análisis agregado vs desagregado**. Ya sea decidir por datos de **preferencia o similitud**, se está tratando **las percepciones de estímulos de los encuestados y creando representaciones de proximidad** entre estímulos en un **espacio t-dimensional** (número de dimensiones  $t$ , es menor que el número de estímulos). El método de análisis **desagregado** genera este resultado, **sujeto a sujeto (produciendo tantos mapas como sujetos)**. Se destaca, que una de las características distintivas de la técnica, es su **capacidad para estimar soluciones para cada encuestado**, que generalmente se representan por separado:

**-La ventaja es la representación de los elementos genuinos de cada percepción del encuestado.**

**-La desventaja es que deberá identificar los elementos comunes a todos los encuestados.**

-Las técnicas de análisis multidimensional también se **pueden combinar encuestados y crear menos mapas perceptuales mediante un proceso denominado análisis agregado**. La agregación puede realizarse tanto **antes** como **después** del análisis de los datos del encuestado. Antes del análisis, la aproximación más simple para el investigador es encontrar las **evaluaciones "medias"** para todos los encuestados y obtener una **solución simple para el grupo de encuestados en conjunto**.

3. **Para identificar grupos de encuestados similares**, el investigador puede analizar mediante el **método cluster** las respuestas de los encuestados para encontrar un número reducido de encuestados o **representativos o medios** y a

continuación **desarrollar mapas para cada individuo y agrupar los mapas de acuerdo con las coordenadas de los estímulos en los mapas.** Es más recomendable utilizar las **evaluaciones “medias”** en lugar de agregar los mapas de percepción individual, debido a que **el menor número de rotaciones del mismo mapa, puede crear problemas al obtener aglomerados razonables mediante el segundo método.** Se puede encontrar una forma especializada de análisis desagregado en **INDSCAL (análisis de diferencias individuales)** [Chang y Carroll 1969] y sus variantes, que tienen las características de ambos tipos de análisis. **INDSCAL supone que todos los individuos comparten un espacio común o de grupo (una solución agregada) pero que todos los encuestados individualmente ponderan las dimensiones, incluyendo las ponderaciones cero cuando ignoran totalmente una dimensión, mediante los pasos:**

**-Primer paso, INDSCAL obtiene el mapa perceptual compartido por todos los individuos,** de la misma forma que hacen otras soluciones agregadas. Sin embargo, **los individuos también están retratados en un mapa perceptual de un grupo especial.** Aquí, **la posición de los encuestados se determina por sus ponderaciones de cada dimensión.**

**-Los encuestados situados cerca emplean similares combinaciones de dimensiones para formar el espacio común de grupo.** Más aún, **la distancia de los individuos desde el origen es una medida aproximada de la proporción de la variación de ese sujeto tomada en cuenta en la solución.** Por tanto, **una posición alejada del origen indica un buen ajuste.**

**-Estar en el origen significa “no ajuste”,** debido a que todas las ponderaciones son cero. Si dos o más sujetos o grupos de sujetos están en el **origen,** se necesitaría configurar otros espacios de grupo para cada uno de ellos. En este análisis, **el investigador tiene presente no sólo una representación del mapa perceptual, sino también el grado en que está representado cada individuo por el mapa perceptual.**

**-Los resultados de cada encuestado pueden utilizarse a continuación para agrupar a los encuestados e incluso identificar diferentes mapas perceptuales en análisis subsecuentes.-**

La elección de análisis **agregados o desagregados** se basa en los objetivos de estudio, así, se recomienda:

**- Un análisis agregado** se recomienda si requiere conocer evaluaciones conjuntas de objetos y las dimensiones empleadas en dichas evaluaciones.

**-Un análisis desagregado,** si requiere **entender la variación entre individuos.**

### **13.5. Análisis multidimensional. Paso 2: Diseño**

La técnica le enfrentará a resolver **8 problemas de diseño básicos:**

- 1. Selección de un enfoque de descomposición (libre de atributos) o de composición (basada en atributos).** Dada la naturaleza de las respuestas obtenidas del individuo en relación con el objeto, es como se clasifican las técnicas de elaboración de mapas perceptuales, con los siguientes métodos: **- El método de descomposición,** como una aproximación que mide sólo la evaluación o impresión conjunta de un objeto y después **intenta obtener**

**posiciones espaciales en un espacio multidimensional que refleje estas percepciones**, asociando ésta técnica típicamente con el análisis multidimensional.

**-El método de composición** es una aproximación alternativa, que emplea varias técnicas multivariantes ya discutidas que se usan en la formación de una impresión o evaluación **basada en una combinación de atributos específicos**.

Cada aproximación tiene ventajas y desventajas aunque son más reconocidos **por su uso los métodos de descomposición**.

2. **Enfoque de descomposición o libre de atributos.** Asociado con las técnicas de análisis multidimensional, se basa **en medidas conjuntas o globales de similitud**, de las que se forman los **mapas perceptuales y el posicionamiento relativo de los objetos. Tiene 2 ventajas:**

**-Sólo requiere que los encuestados expresen percepciones conjuntas de los objetos; no detallan los atributos utilizados en esta evaluación.**

**-Dado que cada encuestado da una evaluación de similitudes completa entre todos los objetos, se pueden desarrollar mapas perceptuales para encuestados individuales o para conjuntos de éstos para formar un mapa de composición.**

Los métodos de descomposición **tienen también desventajas:**

**-El investigador NO tiene una base objetiva provista por el encuestado sobre la cual identificar las “dimensiones” básicas de evaluación de los objetos (la correspondencia de dimensiones objetivas y perceptuales).** En muchos ejemplos, la utilidad de los estudios libres de atributos está restringida porque los estudios arrojan poca luz para una acción específica. Por ejemplo, **la incapacidad de desarrollar un vínculo directo entre acciones de la empresa (la dimensión objetiva) y las posiciones de mercado de sus productos (la dimensión perceptual) muchas veces disminuye el valor de la elaboración de mapas perceptuales.**

3. Además, el investigador **tiene poca ayuda**, aparte de líneas generales y creencias a priori, en la determinación tanto de la **dimensionalidad del mapa perceptual como de la representatividad de la solución**. Algunas medidas globales de ajuste, **NO** son estadísticas, y por tanto las decisiones sobre la solución final implican una **interpretación por parte del investigador**. La **selección de un método específico** requiere decisiones en relación con **la naturaleza de las respuestas de los encuestados (calificación o clasificación)**, si se obtienen similitudes o diferencias y si se utilizan mapas perceptuales individuales o compuestos. Entre los programas más habituales de análisis multidimensional están PREFMAP, KYST, MDSCAL, MDPREF, INDSCAL, ALSCAL, MINISSA, POLYCON y MULTISCALE [Schiffman et al. 1981, Smith 1989]

4. **Enfoque de composición o basado en atributos.** Algunas técnicas multivariantes como el **análisis discriminante o el análisis factorial**, lo incluyen, así como métodos diseñados específicamente para la elaboración de **mapas perceptuales**, tales como el **análisis de correspondencias**. Un rasgo común de estos métodos es **la evaluación de similitud** en la cual se considera un **conjunto de atributos** que son similares entre objetos. La **descripción explícita de las dimensiones del espacio perceptual**, es una ventaja de esta aproximación. Como los encuestados aportan evaluaciones detalladas de los atributos de cada objeto, generan :

-Criterios de evaluación representados por las dimensiones de la solución son más fáciles de averiguar.

-Estos métodos ofrecen un método directo de re-presentación de ambos (**atributos y objetos**) en un **mapa único**, donde diversos métodos ofrecen un posicionamiento adicional de los grupos de encuestados.

-Para los responsables empresarial, esta información supone una visión única en un mercado competitivo.

**-Hay 4 desventajas fundamentales de las técnicas de composición.**

- a. **La similitud entre objetos está limitada sólo a atributos calificados por los encuestados.** Si se omiten los atributos más destacados, no hay oportunidad por parte del encuestado para introducirlos, como ocurre si se ofrece una única medida global.
- b. usted deberá **asumir algún método de combinación de estos atributos para representar la similitud conjunta**, y el método elegido **puede no representar el pensamiento de los encuestados.**
- c. El esfuerzo en la **recolección de los datos** es importante, especialmente a medida que el **número de objetos a elegir aumenta.**
- d. Por último, **los resultados habitualmente no están a disposición del encuestado individual.**

-Los métodos de composición pueden agruparse en **3 grupos básicos:**

a. **Enfoques básicos o post hoc**-se incorporan **gráficos de diferencia semántica o parrillas de importancia-realización**, que descansan en el juicio del investigador y **representaciones univariantes o bivariantes de los objetos.**

b. **Técnicas estadísticas multivariantes convencionales.** Estas técnicas, especialmente el **análisis factorial y análisis discriminante**, son particularmente útiles en el desarrollo de una estructura dimensional entre los numerosos atributos, para representar a continuación objetos sobre estas dimensiones.

c. **Métodos especializados de elaboración de mapas perceptuales.** En esta clase de técnicas destaca el **análisis de correspondencias**, desarrollado específicamente para proporcionar una elaboración de **mapas perceptuales con datos analizados sólo cualitativa o nominalmente como entrada de datos.**

## 5. Selección entre técnicas de composición y descomposición

Puede desarrollarse con **técnicas de composición o descomposición la elaboración de mapas perceptuales**, aunque cada técnica tiene **ventajas y desventajas específicas** que deben considerarse a la vista de los objetivos de investigación. Si la elaboración de mapas perceptuales forma parte del "*espíritu*" de uno de **los dos objetivos básicos** discutidos anteriormente:

-los **enfoques de descomposición o libres de atributos son los más apropiados.**

-Si los objetivos de la investigación se encaminan a la **representación entre objetos sobre un conjunto definido de atributos**, entonces las técnicas de **composición son la alternativa preferida.** Los métodos de composición en los capítulos pasados han ilustrado sus usos y aplicaciones junto con sus puntos

fuertes y sus debilidades. **El investigador siempre debe recordar las alternativas que están disponibles en el caso de que los objetivos del investigador cambien.** Por tanto, aquí nos centramos en los **enfoques de descomposición**, seguido por una discusión del **análisis de correspondencias**, una técnica de composición ampliamente utilizada que es particularmente apropiada para la elaboración de **mapas perceptuales**. Por último, los especialistas equivalentes los términos **elaboración de mapas perceptuales y análisis multidimensional** a menos que se indique lo contrario.

6. **Los Objetos.** El investigador deberá hacer :

-Una serie de cuestionamientos del objeto de estudio a evaluar, antes de comenzar cualquier estudio **de elaboración de mapas perceptuales**, tales como: **¿son los objetos realmente comparables?** Esto representa un supuesto implícito dado que existen **características comunes**, que los encuestados tienen, tanto objetiva como percibida, para evaluar. **Es totalmente indebido que Usted force o induzca al encuestado a comparar creando pares de objetos no comparables** ya que su utilidad sería muy cuestionable. Una segunda cuestión hace referencia al número de objetos a evaluar.

-Al decidir **cuántos objetos hay que incluir**, el investigador debe tener en cuenta **un reducido número de objetos** que faciliten el esfuerzo por parte del encuestado, frente al número de objetos requerido para obtener una solución multidimensional estable. Se sugiere para **soluciones estables, tener más de 4 veces tantos objetos como dimensiones deseadas** [Green et al.1989]. Por tanto, se requieren **al menos 5 objetos para un mapa perceptual unidimensional, 9 objetos para una solución bidimensional**, y así sucesivamente. Cuando se utiliza el método de **evaluación de pares de objetos por similitud**, el encuestado debe hacer **36 comparaciones de 9 objetos por lo que es una tarea importante**. Una solución de **3 dimensiones** supone que habría que analizar al menos **13 objetos**, y por tanto evaluar **78 pares de objetos**. Por lo tanto, hay que **compensar entre la dimensionalidad adecuada a los objetos (y el número implicado de dimensiones subyacentes que pueden ser identificados) y el esfuerzo exigido por parte del encuestado**. La determinación de un nivel aceptable de ajuste, suele afectar un nivel aceptable de ajuste. **Tener menos del número de objetos sugerido para una dimensionalidad dada provoca un sobreajuste**, como el de la **regresión**, y estar fuera de las líneas recomendadas de **4 objetos por dimensión aumenta enormemente las posibilidades de obtener una solución equivocada**. Por ejemplo, un **estudio empírico** demuestra que cuando 7 objetos se ajustan a **3 dimensiones** con valores aleatorios de similitud, el **50%** de las veces se generan mapas perceptuales válidos y **niveles de stress aceptables**. Si los **7 objetos** con similitudes aleatorias se ajustaran a las **4 dimensiones**, los valores de stress caerían a cero, indicando **un ajuste perfecto**, en la mitad de los casos [Kruskal y Wish 1978], aunque en ambos casos **NO** exista un patrón real de similitud entre los objetos. Por tanto, tenga precaución a la hora de incumplir las pautas del número de objetos por dimensión y el impacto que tiene tanto sobre **las medidas de ajuste como sobre la validez de los mapas perceptuales resultantes**

7. **Métodos métricos vs. No métricos.** Originalmente, la técnica se basaba en datos

enteramente **no métricos**, lo que significa que exigían **sólo datos no métricos pero que a su vez proporcionaban sólo resultados no métricos (clasificación-orden)**. Los resultados **no métricos**, sin embargo, **limitaban la interpretabilidad del mapa perceptual**. Actualmente, todos los mapas utilizados hoy en día producen **resultados métricos** que pueden girarse desde el origen, el origen puede cambiarse añadiendo una **constante**, los ejes pueden ser volteados (**reflejo**), **o incluso la solución entera se puede reducir de manera uniforme, sin cambiar la posición relativa de los objetos**. Dado que todo el software actual genera resultados métricos, la distinción se basa en la medida de los datos de similitud. Los métodos **no métricos**, que se distinguen por el dato típicamente **generado por pares de objetos clasificados según rango**, son más flexibles en la medida en que no suponen ningún tipo de relación específica entre la distancia calculada y la medida de similitud. Sin embargo, los métodos **no métricos** contienen menos información para crear el mapa perceptual, es más probable que arrojen **soluciones subóptimas o degeneradas**. **Éste es un problema particular cuando existen amplias variaciones en los mapas perceptuales entre los encuestados o las percepciones entre los objetos no se distinguen o no están bien definidas**. Los **métodos métricos** suponen que los **datos y los resultados sean métricos**. Este supuesto nos permite **estrechar la relación entre la dimensionalidad del resultado final y los datos de entrada**. En lugar de suponer que sólo las relaciones ordenadas están preservadas en los datos de entrada, podemos asumir que **el resultado preserva el intervalo y las calidades del ratio de los inputs de entrada**. Incluso aunque los supuestos subyacentes del software con **programas métricos** son más difíciles de apoyar conceptualmente en muchos casos, los resultados de procedimientos métricos y no métricos aplicados a los mismos datos son a **menudo muy similares**. Por tanto, al seleccionar el tipo de datos de entrada debe considerarse tanto la situación de investigación (**variaciones de percepciones entre encuestados y distinciones entre los objetos**) como el **modo preferido de recogida de datos**.

8. **Recolección de datos. Sobre similitudes o preferencias.**

La distinción principal entre los programas de análisis multidimensional se basa en el tipo de datos (**métricos o no métricos**) utilizados para representar **similitudes o preferencias**. Así, preguntamos a los encuestados asuntos asociados con hacer **juicios de preferencia o basados en similitudes**. Para muchos de los métodos de obtención de datos, **no pueden recogerse ni los datos métricos (calificaciones) ni los no métricos (clasificaciones)**. En muchos casos, sin embargo, las respuestas están limitadas a un solo tipo de datos, del tipo:

-**Datos de similitudes**. Aquí, el investigador está intentando determinar **qué puntos son más parecidos a otros y cuáles son los más diferentes**. Los términos de **similitudes y diferencias a menudo se utilizan indistintamente** para representar medidas de las diferencias entre objetos. Implícita en la medida de similitud está la capacidad **de comparar** todos los pares de objetos. Por ejemplo, todos los pares de objetos del conjunto **A, B, C (es decir, AB, AC, BC)** se clasifican ordenan, entonces todos pares objetos pueden también compararse. Suponga que todos los pares se clasifican como **AB = 1, AC = 2, y BC = 3 (1 es el más parecido)**. Claramente, el par **AB** es más parecido que el par **AC**, el par **AB** es



más parecido que el par **BC** y el par **AC** es más parecido que el par **BC**. Diversos procedimientos para obtener percepciones de los encuestados de las similitudes entre estímulos se pueden utilizar. Cada procedimiento se basa en la noción de que las diferencias relativas entre cualquier par de estímulos deben medirse de tal forma que el investigador pueda determinar si el par es más o menos parecido que cualquier otro par. A continuación discutimos tres procedimientos habitualmente utilizados para obtener las percepciones de similitudes de los encuestados: comparación pareada de objetos, matriz de confusión y medidas derivadas:

**-Comparación pareada de objetos.** El método más ampliamente utilizado para obtener juicios de similitud es el método de pareado de objetos, en el cual se pregunta al encuestado simplemente que ordene o califique todos los pares de objetos. Si tenemos el estímulo **A, B, C, D y E**, podremos clasificar los pares desde el más similar al menos similar. Si, por ejemplo, al par **AB** se le da la calificación de **1**, asumimos que el encuestado ve ese par como contención de los dos estímulos que son más parecidos, en contraste con el resto de los otros pares. Este procedimiento ofrece a una medida no métrica de similitud. Las **medidas métricas** de similitud implican una calificación de similitud (es decir, de 1 "*muy similar*", 0 "*no tan similar*"). Ambas formas (métrica o no métrica) pueden ser utilizadas en la mayoría de los programas de análisis multidimensional.

**-Matriz de confusión.** El emparejamiento (o "*confusión*") de estímulos **I** con estímulos **J** se considera indicativo de similitud. También se conocen como conglomerados subjetivos; el procedimiento típico de obtener estos datos es situar los objetos cuya similitud se va a medir (por ejemplo, 10 videojuegos) en tarjetas, sea descriptivamente o mediante fotografías. Al encuestado se le pide que clasifique en montones todas las tarjetas de tal forma que todas las tarjetas de un montón representen chocolatinas similares. Algunos investigadores dicen a los encuestados que las coloquen en un número fijo de montones; otros dicen que las coloquen en tantos montones como el encuestado quiera. En cada situación, los datos resultan en una matriz de similitudes agregadas similar a una tabla de tabulación cruzada. **Estos** datos indican que productos aparecen juntos más a menudo y por tanto se les considera como más similares. La recogida de datos de esta forma permite solo el cálculo agregado de similitud, dado que las respuestas de todas las personas se combinan para obtener matrices de similitud.

**-Medidas derivadas** Las medidas derivadas de similitud se basan habitualmente en puntuaciones dadas por los encuestados a estímulos. Por ejemplo, se pregunta a los sujetos que evalúen tres estímulos (cereza, fresa y soda lima-limón) sobre un número de atributos (dieta-no dieta, dulce-agrio, sabor fuerte-ligero) utilizando análisis diferenciales semánticos. Las respuestas se evaluarán para cada encuestado (es decir, correlación, índice de acuerdo) para crear medidas de similitud entre los objetos. Aquí hay que hacer **3** importantes supuestos:

- a. El investigador ha seleccionado las dimensiones apropiadas a medir.
- b. Las escalas pueden ser ponderadas (sea igualmente o desigualmente) para conseguir datos de similitud para un sujeto o grupo de sujetos
- c. Incluso si se pudiera determinar la ponderación de las escalas, todos los

individuos tendrían las mismas ponderaciones.

De los tres procedimientos el de **medida derivada es el método menos deseable** para uso de la técnica ya que pretende que la evaluación de los objetos sea hecha de tal forma que el **investigador tenga una influencia mínima**.

**-Recogida de datos de preferencia.** Esta modalidad, implica que los **estímulos se juzgarían en términos de relaciones dominantes**; los estímulos se ordenan en términos de la preferencia **por alguna propiedad**. Por ejemplo, el **servicio A es preferida al servicio C**. Los dos procedimientos más comunes para obtener datos de preferencia son la **clasificación directa y las comparaciones pareadas**.

**-Clasificación directa.** Es un método muy popular de recolección de **datos de similitud no métricos** dado que es fácil de administrar **para un número moderado de objetos**. El encuestado clasifica los objetos de más preferido a menos preferido. Muy similar al procedimiento del **conglomerado subjetivo**, distinguiéndose en este caso a cada objeto se le debe dar **una clasificación única**.

**-Comparaciones pareadas.** A un encuestado se le presentan todos los pares posibles y se le requiere que indique qué miembro de cada par es el preferido.

De esta forma, el investigador **recoge datos explícitos para cada comparación**, lo que permite más detalle que las simples calificaciones directas. **La principal desventaja de este método es el gran número de tareas que hay que realizar** incluso con un número relativamente pequeño de objetos. Por ejemplo, **10 objetos resultan en 90 comparaciones pareadas**, que son demasiadas para la mayoría de las situaciones de investigación. Nótese que las **comparaciones pareadas se utilizan también en la recogida de datos de similitud**, como se citó en el ejemplo del principio del tema.

**-Datos de preferencia frente a datos de similitud.** Esta modalidad es muy útil debido a que las percepciones de los individuos respecto a un objeto en un **contexto de preferencia** tiene la posibilidad de ser diferente de la que se realiza en un **contexto de similitud**; es decir, **una dimensión particular puede ser muy útil** en la descripción de las diferencias entre dos objetos **pero puede no ser consecuente en la determinación de una preferencia**. Los **datos de preferencia** le permitirán ver la ubicación de los objetos en un **mapa perceptual** en el cual las preferencias se reflejan a través de las distancias. Por tanto, **2 objetos podrían aparecer como similares en un mapa preferencial y sin embargo percibidos como diferentes en un mapa de similitud**. Así, **resultarían dos mapas bastante diferentes**, de tal forma que dos marcas de **videojuegos** distintas estarían **apartadas en un mapa de similitud** pero, **con preferencia equivalente**, estarían situadas muy cerca la una de la otra en un **mapa de preferencia**.

En resumen, los **procedimientos de recogida de datos para datos tanto de similitud como de preferencia tienen el propósito común de obtener series de respuestas unidimensionales que representan los juicios de los encuestados**. Estos juicios servirán después como datos de muchos procedimientos de análisis multidimensional que definen la pauta de multidimensionalidad subyacente en relación con estos juicios.

### 13.6. Análisis dimensional. Paso 3: Condiciones de aplicabilidad

Aunque la técnica no tiene supuestos restrictivos en la metodología, referente al tipo de datos o la forma de las relaciones entre las variables, requiere que Usted acepte varios **principios acerca de la percepción**, tales como:

1. **Dimensionalidad y su variación.** Los encuestados entre sí **NO** perciben que un estímulo tenga la misma dimensionalidad (aunque suponga que la mayor parte de las personas emiten juicios en términos de un limitado número de características o dimensiones). Por ejemplo, alguien puede evaluar una casa en términos de sus acabados y diseño, mientras que otros no consideran estos factores sino que lo evalúan en términos del costo y la ubicación
2. **Variación en importancia.** Los encuestados no necesitan asignar el mismo nivel de importancia a una dimensión, incluso si todos ellos perciben esta dimensión. Por ejemplo, dos encuestados perciben un auto en términos de su nivel de ahorro de combustible y marca, pero algunos pueden considerar éstas dimensiones sin importancia mientras que otros pueden considerarla muy importante.
3. **Variación en el tiempo**-la emisión de juicios sobre un estímulo en términos tanto de dimensiones o niveles de importancia **NO tienen por qué permanecer estables en el tiempo.** En decir, **NO puede esperarse que los encuestados mantengan las mismas percepciones durante largos períodos de tiempo.**

A pesar de estos supuestos y de las diferencias que podemos esperar entre los individuos, **la técnica intenta representar percepciones espacialmente**, de tal forma que pueda analizarse cualquier relación subyacente. Como, se observa, el propósito de emplearla reside **no sólo en entender a cada individuo por separado, sino también en identificar las percepciones compartidas y las dimensiones evaluativas dentro de una muestra de encuestados.**

### 13.7. Análisis multidimensional. Paso 4: Estimación y ajuste

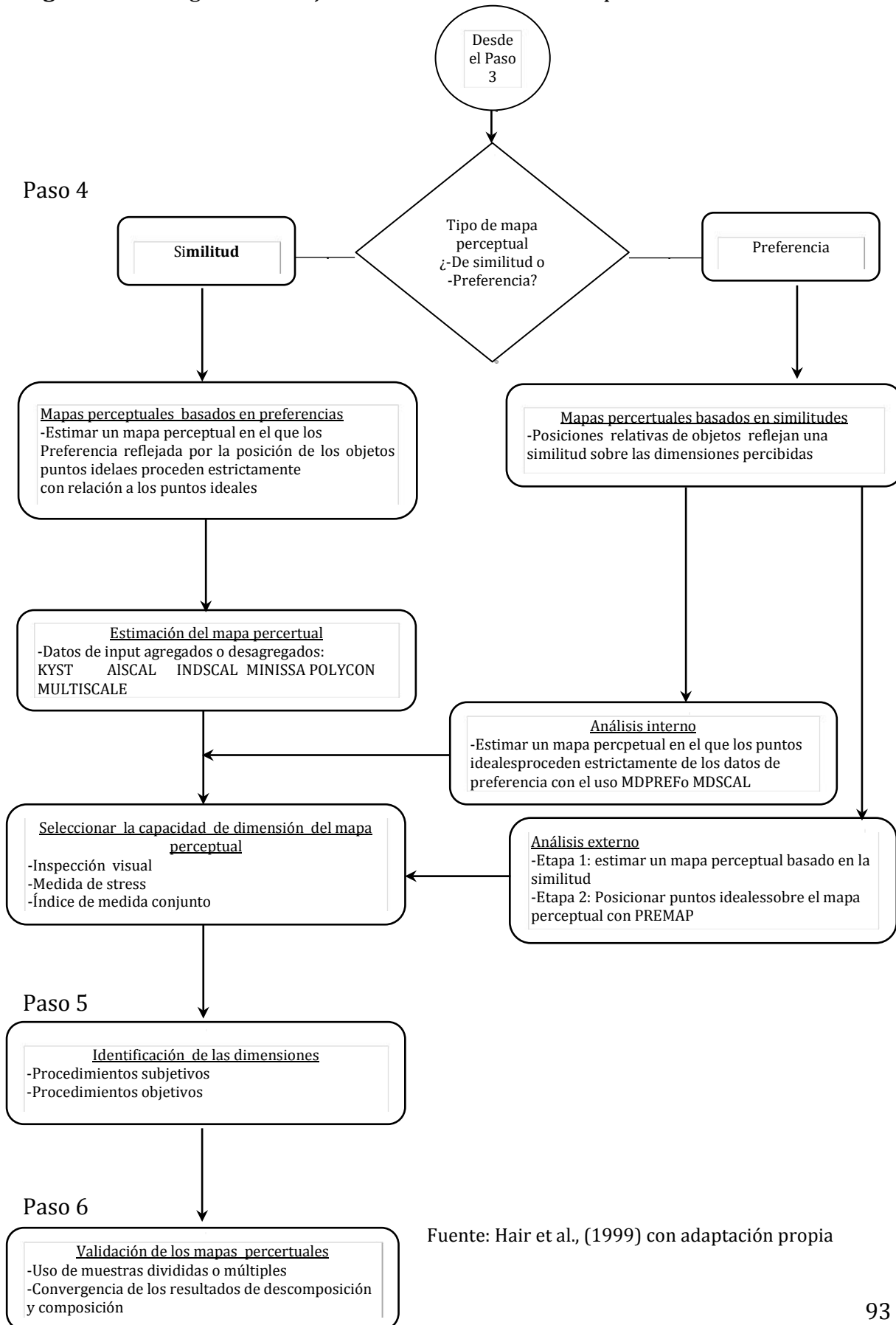
Existe una amplia variedad de software incorporado en programas estadísticos que realizan análisis multidimensional. El objetivo aquí es proporcionarle una visión de la técnica que permita una rápida comprensión de su uso en el software alternativo **SPSS**. Sin embargo, como con otras técnicas multivariantes, existe un amplio desarrollo continuo en conocimientos como en aplicación. Se anexan otras obras al respecto en: [Green et al. 1989, Green y Rao 1972, Lingoes 1972, Kruskal y Wish 1978, Schiffman et al. 1981].

1. **Objeto en un mapa perceptual. Determinación de su posición**  
La primera tarea del cuarto paso implica la situación de objetos que reflejen mejor las evaluaciones de similitud ofrecidas por los encuestados (ver **Figura 13.6**). Los programas del análisis multidimensional siguen un proceso común de determinación de posiciones óptimas. Este proceso puede describirse en **4 pasos**: **Paso 1: Configuración inicial de estímulos ( $S_k$ ) respecto a una dimensionalidad inicial ( $t$ ) deseada.** Existen varias opciones para obtener las configuraciones iniciales. Las 2 más utilizadas son:  
-Configuraciones, bien aplicadas por el investigador **basándose en datos previos**, o -Generadas por puntos pseudoaleatorios desde una **distribución**

**multivariante aproximadamente normal.**

**Paso 2: cálculo de las distancias entre los puntos de estímulo y comparar las relaciones (observadas vs. estimadas) con una medida de ajuste.** Una vez que se ha encontrado la configuración, las distancias de estímulos entre puntos  $d_{ij}$  en las configuraciones iniciales se comparan con **las medidas de distancia ( $d_{ij}$ )** de los juicios de similitud (**s**). Las dos distancias medidas se componen entonces mediante una **medida de ajuste, normalmente una medida de stress.**

**Figura 13.6.** Diagrama de flujo análisis multidimensional pasos 4-5-6



Fuente: Hair et al., (1999) con adaptación propia

**Paso 3:** si la **medida de ajuste no llega a un valor límite seleccionado**, hay que **encontrar una nueva configuración para la cual la medida de ajuste se minimice aún más**. Normalmente el software determina las direcciones en las cuales puede obtenerse la **mejora en el ajuste** y **mueve los puntos de la configuración en esas direcciones mediante pequeños incrementos**.

**Paso 4:** Con una **medida del stress satisfactoria, la dimensionalidad se reduce a uno y el proceso se repite hasta que se alcance la menor dimensionalidad con una medida aceptable del ajuste**.

Lo anterior provoca una gran necesidad de contar con un programa de software adecuado para el cálculo de los datos ya que se produce alta complejidad en su tratamiento. Por ejemplo, si debe evaluar **10 productos/servicios**, cada encuestado deberá clasificar los 45 pares de productos, desde los más parecidos (**1**) a los menos parecidos (**45**). Para hacer un análisis manual, se propone:  
-Coloque los **10 puntos** (representando los 10 productos) **aleatoriamente en una hoja de papel milimétrico** y a continuación mida las distancias entre dos pares de puntos (45 distancias).  
-Calcule el **stress de la solución**, a medida que se muestra la consonancia del orden-clasificación entre las distancias euclídeas (**línea directa**) de los **45 objetos** dibujados y las **45 clasificaciones originales**. (La **medida de stress** es simplemente una medida de lo bien (o mal) que las distancias representadas en un mapa concuerdan con las clasificaciones dadas por los encuestados).

-Si las distancias en línea recta no concuerdan con los rangos originales, **deberá modificar los 10 puntos e intentarlo de nuevo**.  
-El **proceso se hace imposible de calcular a medida que el número de objetos aumenta y las diferencias en percepción y las dimensiones utilizadas en la evaluación crecen**.

Es aquí donde se deduce la importancia de uso de un computador, que se utiliza **sólo para reemplazar los cálculos manuales y permitir una solución más detallada y precisa**. El criterio principal para encontrar la mejor representación de los datos en todos los casos es la **conservación de las relaciones ordenadas entre la clasificación original de los datos y las distancias obtenidas entre los puntos**.

-Al **evaluar un mapa perceptual**, el investigador debe ser consciente de las **soluciones degeneradas**, resultado de los **mapas perceptuales que no son representaciones precisas de las respuestas de similitud**.

-**Las inconsistencias**. Muy a menudo estas son provocadas por inconsistencias en los datos o una incapacidad del programa de análisis informático para alcanzar una solución estable. Están caracterizadas generalmente tanto por un **ciclo circular**, en donde **todos los objetos se muestran similares**, o una **solución de conglomerado**, en la cual **los objetos están agrupados en dos extremos de una dimensión única**. En ambos casos, el programa informático es incapaz de diferenciar entre los objetos por alguna razón. El

investigador deberá entonces a reiniciar el examen del diseño de la investigación para ver dónde se ha producido la inconsistencia.

-**Selección de la dimensionalidad del mapa perceptual.** El objetivo es seleccionar una configuración espacial en un número especificado de dimensiones. El determinar cuántas dimensiones se encuentran efectivamente representadas en los datos se realiza por una de las 3 técnicas:

- a. La evaluación subjetiva,
- b. Los gráficos de caída de las medidas de stress o
- c. Un índice de ajuste conjunto.

-**Primera aproximación.** El mapa espacial es un buen punto de partida para la evaluación. El número de mapas necesarios para la interpretación depende del número de dimensiones. Se produce un mapa para cada combinación de dimensiones. La obtención del mejor ajuste con el menor número posible de dimensiones, es la intención del investigador. Interpretar soluciones más de tres dimensiones es extremadamente difícil y generalmente no merece la pena la mejora en el ajuste. Normalmente, el investigador hace una evaluación subjetiva de los mapas perceptuales y determina si la configuración parece razonable. Es importante esta valoración porque en una etapa final será necesario explicar e interpretar las dimensiones.

-**Una segunda aproximación** es utilizar una medida de stress, que indica la proporción de la variación de las disparidades no tenidas en cuenta por el análisis multidimensional. Varía de acuerdo con el tipo de programa y los datos que se van a analizar y aplica el stress de Kruskal como la medida más utilizada en la determinación de un modelo de buen ajuste. Se define por:

$$\text{Stress} = \sqrt{(d_{ij} - \sim d_{ij})^2 / (d_{ij} - x d_{ij})^2}$$

Donde:

$x d_{ij}$  = la distancia media sobre el mapa

$\sim d_{ij}$  = distancia obtenida a partir de los datos de similitud

$d_{ij}$  = distancias originales facilitadas por los encuestados

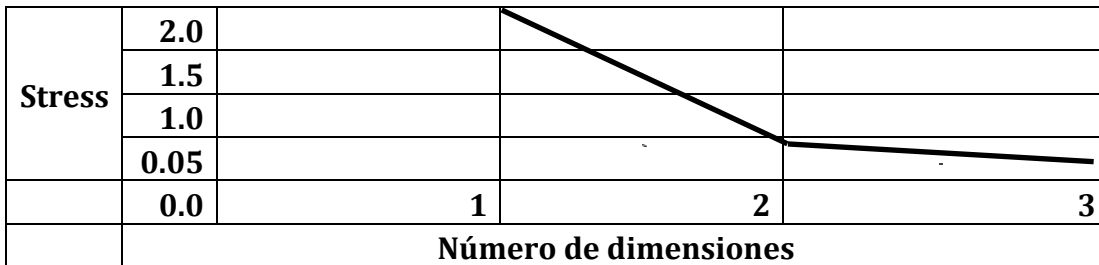
-El valor del stress se hace más pequeño a medida que la  $\sim d_{ij}$  se aproxima a la  $d_{ij}$  original. El stress se minimiza cuando los objetos están situados en una configuración de tal forma que las distancias entre los objetos se ajusten mejor a las distancias originales.

Cabe mencionar, que uno de los problemas que se encuentran al utilizar el stress es análogo al del  $R^2$  en la regresión múltiple; esto es, que el stress siempre mejora con mayores dimensiones. (Recuérdese que el  $R^2$  siempre aumenta con variables adicionales.) Se debe considerar entonces la relación entre el ajuste de la solución y el número de dimensiones. Como se hizo para la extracción de factores en el análisis factorial, podemos realizar un gráfico entre el valor del stress frente al número de dimensiones para determinar el mejor número de dimensiones para utilizarse en el análisis [Kruskal y Wish 1978]. Por ejemplo, en el gráfico de caída en la Figura 13.7, el



**codo** indica que existe una mejora sustancial en la bondad del ajuste cuando el número de dimensiones se aumenta de 1 a 2.

**Figura 13.7. Gráfico de caída para determinar la dimensionalidad apropiada.**



Fuente: Hair et al., 1999 con adaptación propia

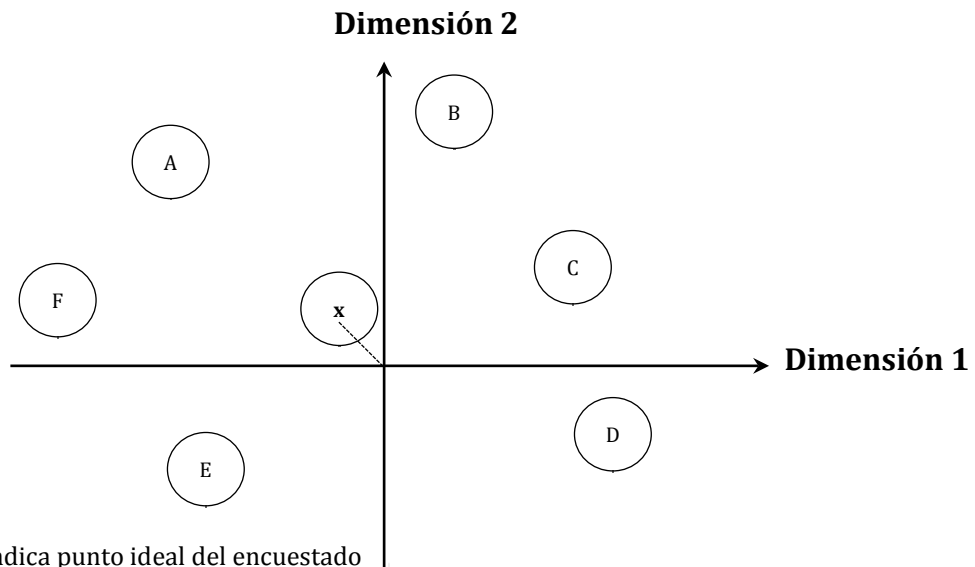
Como se puede apreciar, el mejor ajuste se obtiene con un número relativamente bajo de dimensiones. Ocasionalmente, se usa un índice de correlación al cuadrado como índice de ajuste. Una posible interpretación es como la proporción de variación de las disparidades (datos analizados de forma óptima) por la técnica. Es decir, es una medida de lo bien que se ajustan los datos al modelo de análisis multidimensional. La medida del  $R^2$  del análisis multidimensional representa esencialmente la misma medida de variación que en otras técnicas multivariantes. Por tanto, es posible utilizar criterios de medida similares; esto es, se consideran aceptables medidas de 0.60 o mayores. Así también, cuanto mayor sea el  $R^2$ , mejor será el ajuste.

- El análisis multidimensional y la incorporación de las preferencias.** Por el momento, las exposiciones anteriores se han concentrado en desarrollar mapas perceptuales basados en juicios de similitud y sin embargo, también es posible obtenerlos también a partir de las preferencias. En este sentido, se establece el objetivo en determinar la combinación preferida de características para un conjunto de estímulos que predice preferencias, dado un conjunto de configuración de objetos [Green y Carrnone 1969, Green et al. 1989]. Así, se desarrolla un espacio conjunto que representa tanto los objetos (estímulos) como los sujetos (puntos ideales). En este caso la homogeneidad la percepción de los individuos para el conjunto de objetos, es un supuesto crítico donde, ya que permite que se atribuyan todas las diferencias a las preferencias, no a las diferencias perceptuales, por lo que se debe tomar en cuenta:

  - Puntos ideales.** En la mayoría de las ocasiones, ha sido mal entendido o confundido. Podemos suponer que si localizamos (en el mapa perceptual estimado) el punto que representa la combinación más preferida de atributos percibidos, hemos identificado la posición de un objeto ideal. Igualmente, podemos suponer que la posición de este punto ideal (en relación con los otros

producto/servicios en el mapa perceptual derivado) **define preferencias relativas**, de tal forma que los productos/servicios que estén lejos de este ideal serían menos preferidos. Un **punto ideal se sitúa de tal forma que la distancia de cualquier objeto a este punto expresa cambios en las preferencias**. Considere, la **Figura 13.8**.

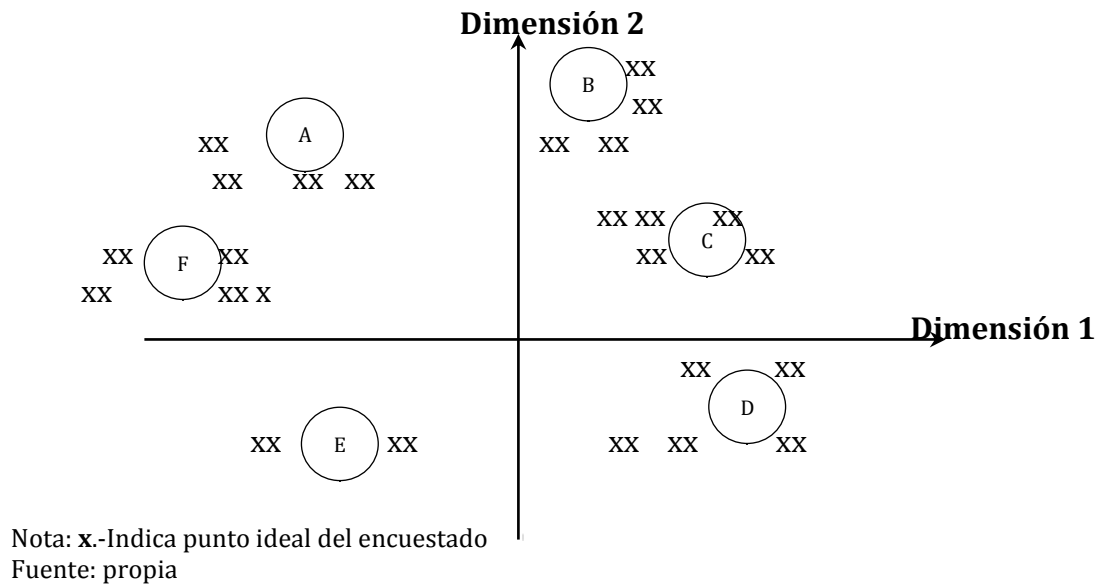
**Figura 13.8. Un punto ideal de un encuestado dentro del mapa perceptual**



Nota: x.-Indica punto ideal del encuestado  
Fuente: propia

Cuando los datos de preferencia sobre los **6 videojuegos** se obtuvieron de la persona indicada por el **punto (x)**, situado de manera que el aumento de la distancia con este punto indica una disminución de la preferencia. Uno puede suponer que este orden de preferencias de la persona es **A, E, F, B, C, D**. Suponga que el **videojuego ideal** está exactamente en el punto **(X)** (en la dirección mostrada por la línea punteada desde el origen). Esto es, el **punto ideal simplemente define la relación de preferencia ordenada por el encuestado entre el conjunto de los 6 videojuegos**. Cabe destacar, que aunque los puntos ideales pueden no ofrecer mucha ayuda, **los conglomerados de ellos pueden ser muy útiles en la definición de segmentos**. Y representan un **mercado potencial** de segmentos de personas con preferencias similares, como se indica en la **Figura 13.9**.

**Figura 13.9. Incorporando puntos ideales múltiples en el mapa perceptual.**



Frecuentemente, son utilizadas **2 aproximaciones para determinar los puntos ideales: estimaciones implícitas y explícitas.** Ésta última, parte de las respuestas directas de los sujetos y puede implicar al sujeto:

-Que **califique un ideal hipotético sobre los mismos atributos a partir de los cuales se han calificado otros estímulos.**

-Preguntarle que **incluya entre los estímulos utilizados para recoger los datos de similitudes, un estímulo ideal hipotético (por ejemplo, marca, imagen).**

-Que **conceptualicen un ideal,** normalmente nos encontraremos con problemas., ya que frecuentemente el encuestado **conforma el ideal en los extremos de las calificaciones** específicas utilizadas o como si fueran similares al producto más preferido de entre aquéllos con los que ha tenido alguna experiencia.

-Que **deba pensar no en términos de similitudes, sino en términos de preferencias, lo que a menudo es difícil con objetos relativamente desconocidos.** Con frecuencia, estos problemas de percepción llevan al investigador a utilizar **estimaciones implícitas del punto ideal.**

Por otro lado, **los puntos ideales, pueden ser ubicados dado que existen varios procedimientos para localizar implícitamente** El supuesto básico es que las **medidas derivadas de posiciones espaciales de los puntos ideales son máximamente consistentes con las preferencias de los encuestados individuales.** Srinivasan y Schocker (1973) suponen que **el punto ideal de todos los pares de estímulos se determina de tal forma que incumple con menor daño la restricción de que estará más cerca de los más preferidos en cada par en lugar de ser el menos preferido.**

En conclusión, existen varias formas de aproximar la estimación de los puntos ideales, y no se ha demostrado la existencia de un método mejor. La elección depende de la habilidad del investigador y del procedimiento de análisis multidimensional elegido.

**-Punto ideal y su ubicación.** Para situar implícitamente el punto ideal a partir de los datos de preferencia, puede desarrollar un análisis:

**-Interno,** de los datos de preferencia hace referencia al desarrollo de un mapa espacial conjunto de los estímulos y puntos del sujeto (o vectores) solamente a partir de los datos de preferencia. Debe basarse en ciertos supuestos respecto a la obtención tanto del **mapa perceptual de estímulos, como de puntos ideales**. Las posiciones del objeto se calculan sobre la base de datos de **exposición de preferencias para cada individuo**. Las dimensiones perceptuales se reflejan como **dimensiones perceptuales** que se ven **“alargadas”** y **ponderadas** para **predecir la preferencia**. Otra característica es que normalmente **emplean un vector de representaciones del punto ideal** (representaciones de **vector frente a punto**), mientras que los **modelos externos pueden estimar representaciones tanto de vectores como de puntos**. Se tienen ejemplos de éstas aproximaciones, como el software MDPREF [Chang y Carroll 1969] ó MDSCAL [Kruskal y Carmone 1967], los cuales son 2 de los programas de este tipo más utilizados, permiten al usuario encontrar configuraciones de estímulos y puntos ideales. Al hacerlo, el investigador debe suponer: (1) que no existe ninguna diferencia entre los objetos, (2) configuraciones separadas para cada objeto, o (3) una configuración única con puntos ideales individuales. Al recoger datos de preferencia, **el investigador puede representar tanto estímulos como encuestados sobre un único mapa perceptual.**

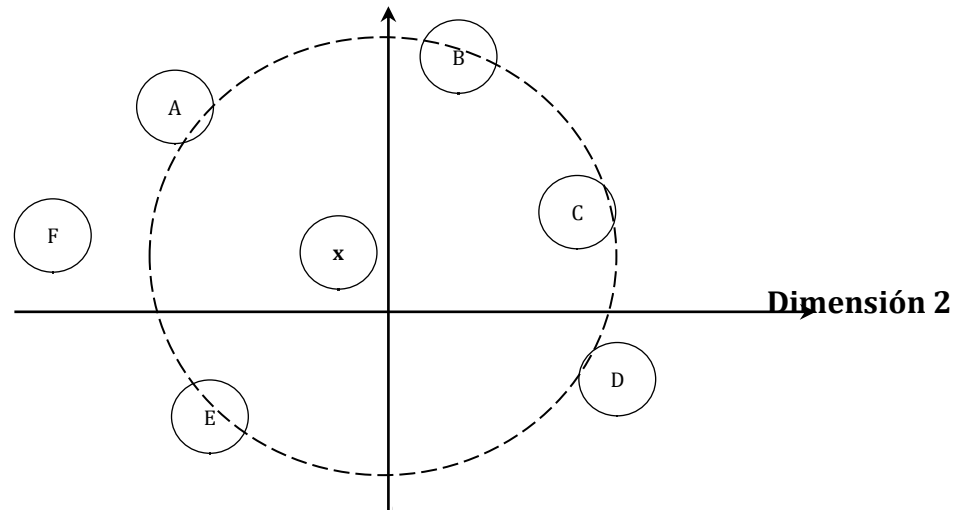
**-Externo** utiliza preferentemente, una **configuración pre-especificada de objetos** para después intentar situar a posteriori los puntos ideales dentro de este mapa perceptual externo. **El análisis externo de los datos de preferencia** se refiere a un **ajuste de puntos ideales** (basados en datos de preferencia) a un **espacio de estímulos desarrollado a partir de datos de similitudes obtenidos de los mismos sujetos**. Por ejemplo, podemos **analizar datos de similitud individualmente**, examinar los mapas individuales para encontrar **percepciones comunes**, y a continuación **analizar los datos de preferencia para cualquier grupo identificado de esta forma**. Así, el investigador tiene que recolectar **tanto datos de preferencia como de similitud para conseguir un análisis externo**. El PREFMAP es un desarrollo para realizar **análisis externos de datos de preferencia**. Dado que la **matriz de similitud** define los objetos en el **mapa perceptual**, el investigador **puede definir a continuación tanto los descriptores de atributos** (suponiendo que el espacio perceptual es el mismo que las dimensiones evaluativas) como puntos ideales para los individuos. El PREFMAP ofrece estimaciones para un número de puntos ideales de diferente tipo, cada uno de ellos basado en distintos supuestos acerca de la naturaleza de las preferencias (es decir, representaciones de vector frente a punto o ponderaciones de dimensión iguales frente a diferentes).

**-Análisis interno y externo. La elección.** Dadas las dificultades de cálculo de los procedimientos de análisis interno y a la confusión de diferencias en preferencia con diferencias en percepción, el tipo de análisis más utilizado, es el **externo**. Además, las salidas sobre las **dimensiones percibidas** pueden cambiar a medida que uno se desplaza **desde el espacio perceptual** (¿son los estímulos similares o no?) **al espacio evaluativo** (¿qué estímulo es el más preferido?).

-Representaciones vector vs. representaciones punto. El método de representación del punto ideal de más fácil comprensión consiste en utilizar la medida de distancia en línea recta (euclídea) de un orden de preferencia desde el punto ideal a todos los puntos que representan los objetos. El método del mapa perceptual de los datos de preferencia hace hincapié en un punto ideal que representa la relación del orden de preferencias de un individuo para un conjunto de estímulos. Estamos suponiendo que la dirección de la distancia desde el punto ideal no es crítica, sólo representa la **distancia relativa**.

En la **Figura 13.10** se muestra un ejemplo, donde el **punto ideal** tal y como se ha posicionado, indica que el **objeto más preferido es C, seguido de B, a continuación E, A, y finalmente F.**

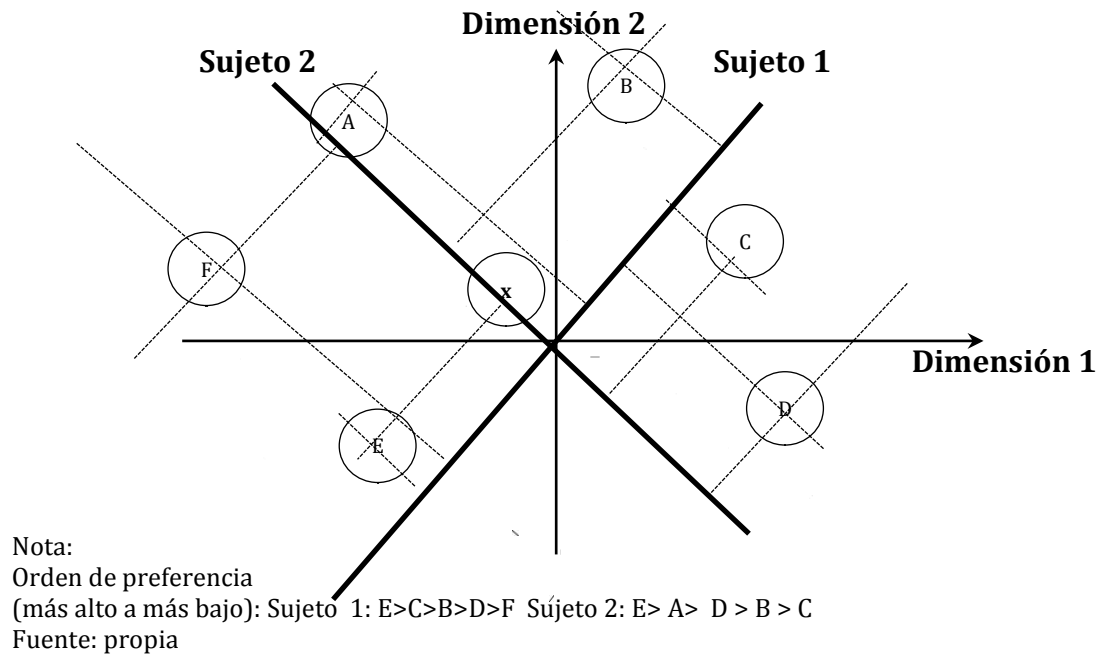
**Figura 13.10. Representación de un punto ideal.**  
Dimensión 2



Nota: orden de preferencia de mayor a menor es: C>B,>E>A>F.  
A,B,C,D,E,F.-Objetos  
X.-Punto ideal  
Fuente: propia

El punto ideal también puede mostrarse como un **vector**. Para calcular las preferencias con esta aproximación, **se dibujan líneas perpendiculares desde los objetos al vector**. La preferencia aumenta en la dirección en que el vector está apuntando. Las preferencias pueden leerse directamente a partir del orden de las proyecciones. La **Figura 13.11** ilustra el enfoque del vector para dos sujetos frente al mismo conjunto de **posiciones de estímulos**.

**Figura 13.11. Representaciones de vector de dos puntos ideales: sujetos 1 y 2.**



Para el **sujeto 1**, el vector tiene la dirección de una preferencia más baja en la esquina inferior izquierda a una preferencia más elevada en la esquina superior derecha. Los valores de preferencia son: **E,C,B,D,A,F**. **Sin embargo, los mismos objetos tienen un orden de preferencia bastante diferente para el sujeto 2.** Para el sujeto 2 el orden de preferencia va desde el más preferido, **A,D,C,E,F,B** al menos preferido, **C**. De esta forma, un solo vector puede representar a cada sujeto. En el enfoque de vector, no hay un único punto ideal pero se asume que el punto ideal está a una distancia infinita externa al vector.

Como **ni la representación de vector ni la de punto puede indicar qué combinaciones de atributos son más preferidas, estas observaciones a menudo NO se ven corroboradas por una experimentación posterior.** Por ejemplo, Raymond (1974) cita un ejemplo en el cual se deducía la conclusión de que la gente prefería los **brownies** debido al grado de humedad y al contenido de chocolate. Cuando los técnicos de alimentación aplicaron este resultado en el laboratorio, encontraron que sus **brownies** hechos para la **especificación experimental eran de chocolate blanco**. Uno **NO** puede **asumir** siempre que las relaciones encontradas **son independientes o lineales, o que se mantienen en el tiempo**, como se apuntó previamente. Sin embargo, el análisis multidimensional es un comienzo en la comprensión de las percepciones y supone un avance considerable en nuestro conocimiento tanto de la metodología como de la percepción humana.

Los datos de preferencia se examinan mejor utilizando el análisis externo como un medio de entender mejor tanto las diferencias perceptuales entre objetos basados en juicios de similitud y elecciones de preferencia hechas dentro de un mapa perceptual de objetos. De esta forma, el investigador puede distinguir entre ambos tipos de evaluaciones perceptuales y entender más precisamente las percepciones de los individuos.

### 13.8. Análisis multidimensional. Paso 5: Interpretación

Conseguido el **mapa perceptual**, observará que los **2 enfoques -composición y descomposición-** divergen de nuevo en su interpretación de los resultados, de tal forma que:

1. **Para los métodos de composición, el mapa perceptual** debe ser validado con otras **medidas de percepción**, dado que las posiciones están totalmente definidas por los atributos especificados por el investigador. Por ejemplo, los resultados del análisis discriminante se pueden aplicar a un nuevo conjunto de objetos o encuestados, **evaluando la capacidad de diferenciar con estas nuevas observaciones.**
2. **Para los métodos de descomposición, el asunto más importante es la descripción de las dimensiones perceptuales y su correspondencia a los atributos.** De hecho, ya hay varias técnicas descriptivas para **“etiquetar”** las dimensiones, así como para integrar preferencias (para **objetos y atributos**) con los **juicios de similitud**. De nuevo, en línea con sus objetivos, los métodos de descomposición proporcionan una visión inicial de las percepciones desde la que pueden surgir perspectivas más formalizadas. Dado que en otros capítulos del texto se han tratado muchas de las técnicas de composición, lo que queda del capítulo se centra en los métodos de descomposición, fundamentalmente las diversas técnicas utilizadas en el análisis multidimensional. Una excepción notable es la discusión de un enfoque de composición análisis de correspondencias que, en cierto grado, salva la brecha entre las dos aproximaciones para su flexibilidad y su método de interpretación.
3. **Identificación de las dimensiones.** Tal y como se discute en el **Capítulo 3**, la identificación de las dimensiones subyacentes es tarea difícil. Las técnicas de análisis multidimensional no tienen un procedimiento específico para identificar las dimensiones. El investigador, una vez desarrollados los mapas con una dimensionalidad seleccionada, puede adoptar diversos procedimientos, tanto subjetivos como objetivos.

**-Procedimientos subjetivos.** La interpretación siempre debe incluir algún elemento de juicio del investigador o del encuestado, y en muchos casos esto se revela adecuado para las cuestiones abiertas. Un método muy simple, aunque efectivo, es que **el encuestado etiquete mediante inspección visual las dimensiones del mapa perceptual.** Se puede preguntar a los encuestados que interpreten subjetivamente la dimensionalidad mediante la inspección de los mapas, o un conjunto de **“expertos”** puede evaluar e identificar las dimensiones. Aunque **no hay un intento de vincular cuantitativamente las dimensiones a los atributos**, este enfoque puede ser el mejor si se cree que las dimensiones son altamente intangibles, afectivas o emocionales, en contenido, de tal forma que no puedan deducirse descriptores adecuados. De forma similar, Usted puede describir las dimensiones en términos de características conocidas (**objetivas**). De esta forma, se hace la correspondencia entre las dimensiones objetiva y perceptual directamente, aunque estas relaciones no son un resultado de una interacción con el encuestado sino del juicio del investigador. **-Procedimientos objetivos.** Como complemento a los procedimientos subjetivos,



se han desarrollado **varios métodos más formalizados**. El método más ampliamente utilizado, PROFIT (PROPERTY FITTING), recoge calificaciones de atributos para cada objeto y a continuación busca la mejor correspondencia de cada atributo para el espacio perceptual derivado. El intento es identificar los atributos determinantes en los juicios de similitud hechos por los individuos. Se dan las medidas de ajuste para cada atributo, así como su correspondencia con las dimensiones. El investigador puede entonces determinar qué atributos describen mejor las posiciones perceptuales y son ilustrativas de las dimensiones. La necesidad de correspondencia entre los atributos y las dimensiones definidas disminuye con el uso de resultados métricos, en la medida que las dimensiones pueden ser modificadas libremente sin cambios en la interpretación.

**-Elección entre procedimientos objetivos y subjetivos.** Tanto para los procedimientos objetivos como para los subjetivos, se debe recordar que aunque la dimensión puede representar un atributo único, habitualmente no lo hace. Un procedimiento común es recoger los datos de varios atributos, asociándolos bien subjetivamente bien empíricamente con las dimensiones donde se aplique, y determinar las etiquetas para cada dimensión utilizando múltiples atributos, de la misma forma que el análisis factorial. Muchos investigadores sugieren que la mejor alternativa consiste en utilizar los datos de atributos para ayudar a etiquetar las dimensiones. **El problema, sin embargo, es posible que Usted no incluya todos los atributos importantes en el estudio** y no estar seguro de que las etiquetas representen todos los atributos relevantes. Tanto los procedimientos subjetivos como los objetivos, ilustran la dificultad de etiquetar los ejes. Esta tarea no debe dejarse para el final, en la medida en que las denominaciones de las dimensiones son esenciales para una ulterior interpretación y uso de los resultados. Usted deberá seleccionar el tipo de procedimiento que mejor se ajusta **tanto a los objetivos de la investigación**, como a la información disponible. Por tanto, debe plantearse tanto la obtención de denominaciones para las dimensiones, como la estimación del mapa perceptual.

### 13.9. Análisis multidimensional. Paso 6: Validación

Debido a la naturaleza altamente inferencial de ésta técnica, este esfuerzo se dirige a asegurar la generalidad de resultados tanto en los objetos como en la población. Pero los esfuerzos de validación son problemáticos y el único resultado que puede ser utilizado a efectos comparativos concierne a las **posiciones relativas de los objetos aunque, las dimensiones subyacentes no tienen fundamento para la comparación**. Si las posiciones varían, Usted no podrá determinar si los objetos se ven de forma diferente, si las dimensiones perceptuales varían o ambas cosas. De hecho, **aún no se han incorporado en los programas informáticos métodos sistemáticos de comparación**. Se deja la libertad a Usted para aplicar métodos que puedan servir como líneas generales, pero que no son específicos de resultados de análisis multidimensional, tales como:

1. **La aproximación más directa es un *split***, o comparación de varias muestras, en la que **o bien se divide la muestra original o se recoge una nueva muestra**. En cada caso, el investigador debe encontrar los medios para **comparar los**

**resultados.** A menudo, la comparación entre resultados se hace visualmente o con una correlación simple de coordenadas. Software, como el FMATCH [Smith, 1989], pero el investigador debe determinar a continuación cuántas de las disparidades se deben a diferencias en las percepciones del objeto, dimensiones diferentes, o ambas.

2. Otro método es **obtener una convergencia de los resultados del análisis multidimensional aplicando tanto métodos de descomposición como de composición a la misma muestra.** Podría aplicarse:
  - En primer lugar **el método de descomposición**, junto con la interpretación de las dimensiones para identificar los atributos claves.
  - Después, podrían aplicarse **uno o más métodos de composición**, en particular los del **análisis de correspondencias**, para confirmar los resultados. **El investigador debe darse cuenta de que ésta no es una verdadera validación de los resultados como generalizables**, pero confirma la interpretación de la dimensión. Desde este punto, podrían considerarse los esfuerzos de validación con otras muestras y otros objetos para demostrar la generalidad de otras muestras.

### 13.10. Análisis multidimensional. Ejemplos

#### Paso 1: Objetivos

**Problema 1:** suponga el caso de estudio donde una empresa fabricante de autos quiere entrar al mercado español (Vila-López, 2013) y requiere hacer un plan de mercadotecnia para la introducción de sus productos/servicios. Para lograrlo, emprendió un estudio en el que involucra **18** empresas de autos, tomando de referencia las que alcanzaron los volúmenes de facturación más elevados España. La decisión se ha tomado a partir de los resultados obtenidos en una fase previa cualitativa.

El propósito de la investigación será explorar la percepción que se tiene de los actuales fabricantes por parte de los clientes en un plan de dos fases de análisis: (1) identificación de la posición competitiva en un mapa perceptual de los principales competidores en el mercado con un estudio de las comparaciones de las dimensiones utilizadas por los potenciales clientes, y (2) evaluación de las preferencias en relación a los principales competidores. Por lo tanto en el curso de las entrevistas, se recogieron 2 tipos de datos: juicios de similitud, y preferencias de cada empresa ante diferentes situaciones de compra.

#### Paso 2: Diseño

Se ha dividido en:

1. **Datos de similitud.** El punto de partida de la recogida de datos fue la obtención de percepciones de los encuestados correspondientes a la similitud o disparidad entre las **18** empresas automotrices competidoras en el mercado. Los juicios de similitud se hicieron con el enfoque de comparaciones pareadas entre objetos. Se presentaron a los encuestados **153 pares de empresas ( $18 \cdot 17 / 2$ )**, e indicaron la similitud de cada par de ellas sobre una escala de 18 puntos, siendo 1:

**“completamente diferente”** y 18 siendo **“muy parecido”**. Nótese que los valores tienen que ser transformados debido a que valores crecientes para calificaciones de similitud indican mayor similitud, lo opuesto de una medida de distancia. Se ha optado por recolectar la información sobre la forma en que se relacionan las **18** empresas categorizándolas. Para lograr lo anterior se solicitó a **211** profesionales del sector que como encuestados ubicaron las percepciones que tenían de las **18** empresas, para asignarlas en tantos grupos como consideraran oportuno, basándose en la competencia que percibe entre ellas. Para tal fin se recurrió al empleo de tarjetas, cada una con el nombre de un competidor diferente. El porcentaje de veces que **2 empresas** han sido agrupadas juntas es lo que se conoce como **coeficiente similitud** o **coeficiente de proximidad**. Así por ejemplo, el **coeficiente de similitud** o **proximidad** entre Audi (**1** en la matriz) y BMW (**2** en la matriz) es de **89.6%**, lo que equivale a afirmar que de los **211** profesionales encuestados, **188** han colocado Audi y BMW en la misma categoría competitiva (**188/211= 89.6%**). Agregando los resultados de toda la muestra de encuestados, se obtuvo **una matriz cuadrada simétrica**, en cuyas celdas se recoge la frecuencia con que **2 competidores** han sido agrupados juntos. **Los datos son recopilados en la base de datos AUTOS Similitud.sav. Ver Figura 13.12, Figura 13.13 y Figura 13.14**

**Figura 13.12. Matriz cuadrada derivada a partir de datos de categorización**

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	100	89.6	3.3	2.8	2.8	3.3	8.5	3.8	85.8	4.7	3.8	2.8	3.3	17.5	3.3	9.5	20.9	57.3
2		100	1.9	3.3	3.3	1.9	9.5	4.7	90.5	3.8	2.4	1.9	1.4	1.8	1.9	11.4	14.2	55.0
3			100	35.1	63.5	76.3	28.9	34.1	3.8	45	69.2	74.4	78.7	33.2	78.2	26.5	42.2	10.4
4				100	50.7	27.5	38.4	83.4	4.3	39.8	24.6	2.8	26.5	30.8	33.6	35.1	16.1	11.8
5					100	59.2	34.6	48.8	5.2	43.6	51.7	51.7	57.3	35.5	61.6	26.5	34.1	10.4
6						100	28.4	26.5	3.8	49.3	80.6	76.3	83.9	32.2	75.4	29.4	47.4	11.4
7							100	44.1	9	54	33.6	35.5	28.9	42.7	29.4	67.3	3.6	23.2
8								100	4.3	39.3	24.2	25.1	25.1	2.8	33.2	3.6	15.6	10.9
9									100	3.3	3.8	3.3	3.3	15.2	3.8	9.5	12.8	56.4
10										100	53.6	47.9	47.4	40.3	48.8	49.8	41.2	14.2
11											100	78.7	76.3	35.1	71.1	32.2	5.4	12.8
12												100	78.2	37.4	72.5	30.3	49.3	11.4
13													100	31.8	79.1	25.6	47.9	9.5
14														100	31.8	44.5	40.8	33.6
15															100	29.9	42.2	9
16																100	37.4	24.6
17																	100	32.7
18																		100

Notas: -Unidad, son proximidades; -Leyenda: Audi: 1 Citroen: 3 Fiat: 5 Honda: 7 Mercedes: 9 Opel: 11 Renault: 13 Seat: 15; Volkswagen: 17; BMW: 2; Daewoo: 4 Ford: 6; Huyndai: 8 Nissan: 10; Peugeot: 12; Rover: 14; Toyota: 16 Volvo: 18. Fuente: Vila-López (2013)

**Figura 13.13. Visor de variables de AUTOS Similitud. Sav.**

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	AUDI_S	Numérico	3	1	similitud	Ninguna	Ninguna	8	Derecha	Escala	Entrada
2	BMW_S	Numérico	3	1	similitud	Ninguna	Ninguna	8	Derecha	Escala	Entrada
3	CITROEN_S	Numérico	3	1	similitud	Ninguna	Ninguna	8	Derecha	Escala	Entrada
4	DAEWOO_S	Numérico	3	1	similitud	Ninguna	Ninguna	8	Derecha	Escala	Entrada
5	FIAT_S	Numérico	3	1	similitud	Ninguna	Ninguna	8	Derecha	Escala	Entrada
6	FORD_S	Numérico	3	1	similitud	Ninguna	Ninguna	8	Derecha	Escala	Entrada
7	HONDA_S	Numérico	3	1	similitud	Ninguna	Ninguna	8	Derecha	Escala	Entrada
8	HYUNDAI_S	Numérico	3	1	similitud	Ninguna	Ninguna	8	Derecha	Escala	Entrada

Fuente : SPSS 20 IBM

**Figura 13.14. Visor de datos de AUTOS Similitud. Sav.**

	AUDI_S	BMW_S	CITROEN_S	DAEWOO_S	FIAT_S	FORD_S	HONDA_S	HYUNDAI_S	MERCEDES_S	NISSAN_S	OPEL_S	PEUGEOT	RENAULT_S	ROVER_S	SEAT_S	TOY
1	1.0	89.6	3.3	2.8	2.8	3.3	8.5	3.8	85.8	4.7	3.8	2.8	3.3	17.5	3.3	20.9
2		1.0	1.9	3.3	3.3	1.9	9.5	4.7	90.5	3.8	2.4	1.9	1.4	1.8	1.9	11.4
3			1.0	35.1	63.5	76.3	28.9	34.1	3.8	45.0	69.2	74.4	78.7	33.2	78.2	26.5
4				1.0	50.7	27.5	38.4	83.4	4.3	39.8	24.6	2.8	26.5	30.8	33.6	35.1
5					1.0	59.2	34.6	48.8	5.2	43.6	51.7	51.7	57.3	35.5	61.6	26.5
6						1.0	28.4	26.5	3.8	49.3	80.6	76.3	83.9	32.2	75.4	29.4
7							1.0	44.1	9	54	33.6	35.5	28.9	42.7	29.4	67.3
8								1.0	4.3	39.3	24.2	25.1	25.1	2.8	33.2	3.6
9									1.0	3.3	3.8	3.3	3.3	15.2	3.8	9.5
10										1.0	53.6	47.9	47.4	40.3	48.8	49.8
11											1.0	78.7	76.3	35.1	71.1	32.2
12												1.0	78.2	37.4	72.5	30.3
13													1.0	31.8	79.1	25.6
14														1.0	31.8	44.5
15															1.0	29.9

Fuente : SPSS 20 IBM

2. **Evaluaciones de preferencia.** Los datos finales sirven para evaluar las preferencias de cada encuestado respecto a las **18 empresas en 15 estrategias diferentes de introducción.** En cada situación, los encuestados clasifican las empresas en orden de preferencia para ese tipo particular de estrategia, recolectando el porcentaje de veces que cada empresa (columna) ha sido asociada con cada propiedad (fila), por los especialistas. Ver **Figura 13.15**

**Figura 13.15. Matriz de preferencias de estrategia por empresa**

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
a	9.8	7	28.5	22	18.2	43.5	8.9	19.2	7	9.3	17.8	15	27.6	7.5	23.4	10.7	11.7	6.5
b	39.7	40.7	15.9	15	15	28	20.1	15.9	45.3	17.3	19.6	15.4	16.8	14.5	16.4	23.8	25.2	20.1
c	7.9	7.9	52.8	25.2	22.4	24.3	9.3	22.9	6.5	12.6	13.6	15.9	16.8	7.5	18.7	8.4	8.9	5.1
d	49.1	50	18.2	15	15.9	22.9	19.6	16.4	52.3	18.7	22	19.2	22	19.2	17.3	22	36.4	29
e	48.1	50.5	15.4	13.6	13.1	17.3	15.9	14	52.3	13.6	16.8	13.6	14	15	15.4	20.6	29.9	22
f	5.7	19.2	23.4	16.4	18.2	26.6	20.6	16.4	18.7	16.4	28	21	22.9	14.5	21	16.4	29.4	16.8
g	22	23.8	26.6	16.8	22.9	33.6	16.4	17.3	24.8	17.8	26.2	22	32.2	15	28	17.8	22	16.4
h	12.6	15.4	28.5	16.8	28	41.1	13.1	16.8	11.7	15.4	33.6	28	32.2	12.6	36.9	12.6	24.3	9.8
i	55.1	64	23	19	19	4.7	6.5	19	72.4	2.3	3.7	2.8	2.3	7	2.3	9.8	11.7	23.8
j	23.8	21	22.4	21.5	22	38.8	20.6	23.8	20.6	18.2	21.5	21	29.4	16.4	21.5	23.8	21.5	15.9
k	14.5	17.3	25.7	18.7	19.2	32.7	15	16.8	15.4	16.8	18.2	16.4	22	14.5	25.2	16.4	14	15
l	23.4	24.3	36.9	16.4	23.4	43.5	16.4	16.4	20.1	19.6	32.2	27.6	46.3	18.2	45.3	15.4	28	16.8
ll	22.4	23.4	38.3	15	23.4	50.5	15.4	15	22	17.8	33.6	29	41.6	19.2	47.2	16.4	25.7	18.7
m	27.6	24.8	37.9	16.4	26.6	48.6	17.8	15.9	24.3	22.4	36.9	32.2	46.3	20.6	44.9	17.3	30.4	21.5
n	28	23.4	33.2	19.6	22.4	43.5	19.2	21	22	21	29	26.6	28.5	20.1	28	20.1	24.8	20.1

Notas: -Unidad, son proximidades; -Leyenda: Audi: 1 Citroen: 3 Fiat: 5 Honda: 7 Mercedes: 9 Opel: 11 Renault: 13 Seat: 15; Volkswagen: 17; BMW: 2; Daewoo: 4 Ford: 6; Huyndai: 8 Nissan: 10; Peugeot: 12; Rover: 14; Toyota: 16 Volvo: 18. Fuente: Vila-López (2013)

Con las estrategias de introducción siguientes de la **Figura 13.16, Figura 13.17 y Figura 13.18.**

**Figura 13.16. Matriz de preferencias de estrategia por empresa**

Estrategia	Descripción
a	Producir muchos coches a la vez para abaratar el coste de fabricación de cada automóvil
b	Invertir mucho en tecnología, por ejemplo en robots, para acortar los tiempos de fabricación
c	Vender sus coches a precios más baratos, haciendo, por ejemplo, ofertas en el concesionario
d	Interesarse por aspectos relacionados con la calidad ofreciendo coches que den buen resultado.
e	Utilizar los mejores componentes: motores potentes, tecnología avanzada.
f	Fabricar coches de bajo consumo, para favorecer el ahorro y responder a las preocupaciones ecológicas
g	Fabricar coches a medida de cada consumidor, teniendo en cuenta sus gustos y poder adquisitivo: color, precio, prestaciones
h	Interesarse por la gente joven, ofreciendo coches para ciudad y económicos acorde con sus necesidades
i	Interesarse por la "clase alta", ofreciendo coches caros pero excelentes que satisfagan sus exigencias
j	Presentar modelos nuevos frecuentemente, ser innovadores en diseño, estilo, color
k	Acortar los plazos de entrega del coche, para que el cliente no tenga que esperar
l	Tener buenas redes de distribución para vender sus coches en España
ll	Tener implantación en España, dando seguridad a los clientes de que la empresa esta cerca y responde
m	Disponer de una amplia red de servicios post venta (reparaciones, talleres, repuestos)
n	Desarrollar técnicas de marketing novedosas que beneficien al cliente: promociones, formas de pago

Fuente: Vila-López (2013)



### Paso 3. Condiciones de aplicabilidad

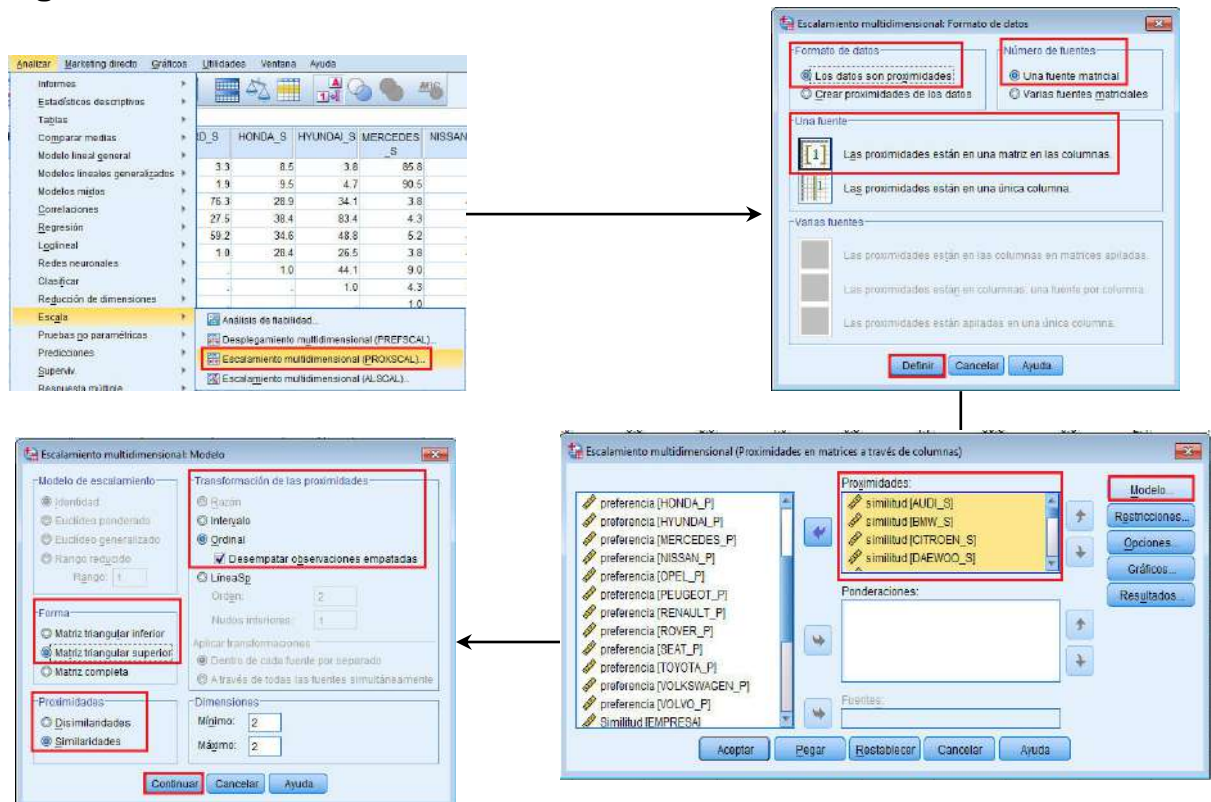
Los supuestos del análisis multidimensional y del análisis de correspondencias se centran principalmente en la **comparabilidad y representatividad** de los objetos que están siendo evaluados y de los encuestados. En relación a la muestra, el plan muestral resaltaba la importancia de obtener una muestra representativa de las empresas competidoras automotrices. Es más, se tomó la precaución de reclutar encuestados de una posición comparable y con conocimiento del mercado, asegurando que las posiciones discrepantes puedan atribuirse a diferencias perceptuales entre los encuestados.

### Paso 4. Estimación y ajuste

#### Caso: **Similitudes**

**Teclear:** Analizar->Escala->Escalamiento multidimensional (PROXSCAL)->Formato de datos: Los datos son proximidades; Número de fuentes: Una fuente matricial; Una fuente: Las proximidades están en una matriz en las columnas->Definir->Proximidades: seleccionar variables métricas de similitud (AUDI\_S, BMW\_S, etc.)->Modelo->Transformación de las proximidades: Ordinal; Desempatar observaciones empatadas->Forma: Matriz triangular superior; Proximidades: Similitudes->Continuar->Aceptar. Ver Figura 13.17

Figura 13.17. Proceso de escalamiento multidimensional PROXSCAL



Fuente: SPSS 20 IBM

### Paso 5: Interpretación

La primera tabla que genera **SPSS** es la de **Resumen de procesamiento de los casos**, la cual reporta casos, fuentes, objetos y proximidades calculadas. Ver **Figura 13.18**.

**Figura 13.18. Tabla Resumen del procesamiento de los casos**

Casos		18
Fuentes		1
Objetos		18
Proximidades	Proximidades totales	153 <sup>a</sup>
	Proximidades perdidas	0
	Proximidades activas <sup>b</sup>	153

- a. Suma de todas las proximidades estrictamente triangulares superiores.
- b. Las proximidades activas incluyen todas las proximidades no perdidas.

Fuente: SPSS 20 IBM

La siguiente tabla que se muestra como resultado del procesamiento de **SPSS**, es la de **Medidas de ajuste y stress**, de la que se deberá observar que los resultados obtenidos para todos los tipos de Stress sean aproximadamente cercanos a cero (0) y que tanto la Dispersión explicada (D.A.F.) y el **Coefficiente de congruencia de Tucker** sean aproximadamente cercanos a uno (1) a fin de que el modelo propuesto, se considere adecuado en ajuste para continuar el análisis. Ver **Figura 13.19**

**Figura 13.19. Medidas de ajuste y stress**

#### Bondad de ajuste

Stress bruto normalizado	.00231
Stress-I	.04801 <sup>a</sup>
Stress-II	.08758 <sup>a</sup>
S-Stress	.00458 <sup>b</sup>
Dispersión explicada (D.A.F.)	.99769
Coefficiente de congruencia de Tucker	.99885

PROXSCAL minimiza el stress bruto normalizado.

- a. Factor para escalamiento óptimo = 1.002.
- b. Factor para escalamiento óptimo = .999.

Fuente: SPSS 20 IBM

Así también, en el reporte espacio común, se genera tabla **Coordenadas finales** de posicionamiento en 2 dimensiones donde en la Dimensión 1, las marcas más relevantes son: BMW, AUDI, MERCEDES y en la Dimensión 2: Hyundai y Daewoo. Ver **Figura 13.20**.

**Figura 13.20. Coordenadas finales**  
**Espacio común**

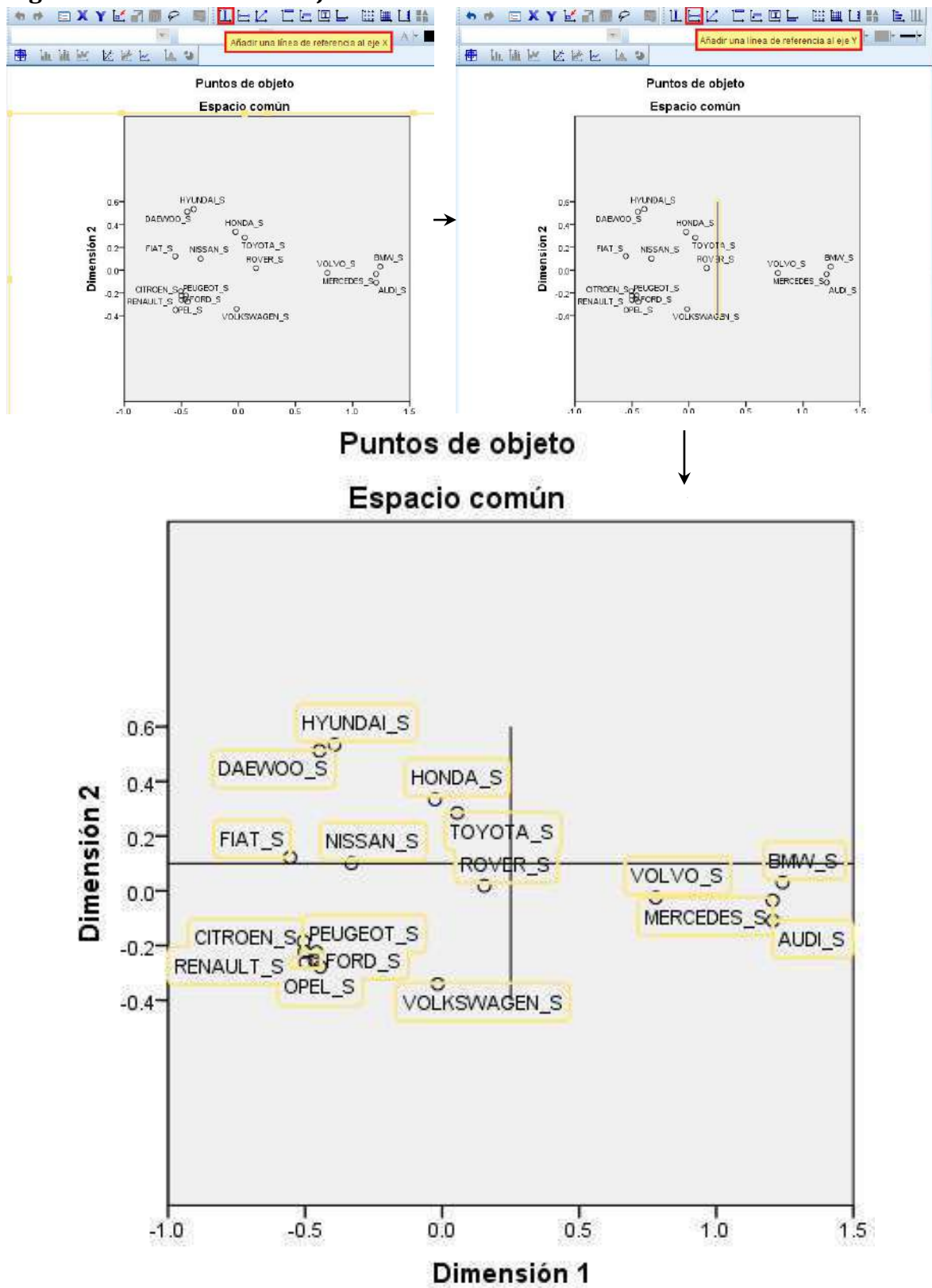
	Dimensión	
	1	2
AUDI	1.206	-.109
BMW	1.242	.030
CITROEN	-.506	-.185
DAEWOO	-.448	.512
FIAT	-.555	.122
FORD	-.465	-.254
HONDA	-.026	.334
HYUNDAI	-.392	.534
MERCEDES	1.206	-.035
NISSAN	-.331	.101
OPEL	-.444	-.278
PEUGEOT	-.459	-.224
RENAULT	-.498	-.262
ROVER	.154	.018
SEAT	-.502	-.223
TOYOTA	.055	.285
VOLKSWAGEN	-.016	-.341
VOLVO	.779	-.025

Fuente: SPSS 20 IBM

Finalmente, **SPSS** genera el espacio común de los **Puntos de objeto**, cual permite gráficamente visualizar los resultados que por similitudes arroja el estudio. Ver **Figura 13.21**.



Figura 13.21. Puntos de objeto



Fuente: SPSS 20 IBM

Como resultado, se observa que existen tres sectores donde se concentran los competidores A partir de las dimensiones 1 y 2 determinadas a nivel de la categorización que realizaron los encuestados especialistas en el área. Con esto, se termina el análisis de similitudes y se iniciará el de preferencias.

**Caso: Preferencias**

Es tomado en cuenta la Figura 13.16. Matriz de preferencias de estrategia por empresa, con los siguientes desglose de la base de datos (ver Figuras 13.22 y 13.23).

**Figura 13.22. Visor de variables de AUTOS Preferencias. Sav. Caso preferencias.**

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
19	ESTRATEGIA_P	Numérico	8	0	preferencia	{1, Producir ...	Ninguna	20	Derecha	Nominal	Entrada
20	AUDI_P	Numérico	3	1	AUDI	Ninguna	Ninguna	8	Derecha	Escala	Entrada
21	BMW_P	Numérico	3	1	BMW	Ninguna	Ninguna	8	Derecha	Escala	Entrada
22	CITROEN_P	Numérico	3	1	CITROEN	Ninguna	Ninguna	8	Derecha	Escala	Entrada
23	DAEWOO_P	Numérico	3	1	DAEWOO	Ninguna	Ninguna	8	Derecha	Escala	Entrada
24	FIAT_P	Numérico	3	1	FIAT	Ninguna	Ninguna	8	Derecha	Escala	Entrada
25	FORD_P	Numérico	3	1	FORD	Ninguna	Ninguna	8	Derecha	Escala	Entrada
26	HONDA_P	Numérico	3	1	HONDA	Ninguna	Ninguna	8	Derecha	Escala	Entrada
27	HYUNDAI_P	Numérico	3	1	HYUNDAI	Ninguna	Ninguna	8	Derecha	Escala	Entrada
28	MERCEDES_P	Numérico	3	1	MERCEDES	Ninguna	Ninguna	8	Derecha	Escala	Entrada
29	NISSAN_P	Numérico	3	1	NISSAN	Ninguna	Ninguna	8	Derecha	Escala	Entrada
30	OPEL_P	Numérico	3	1	OPEL	Ninguna	Ninguna	8	Derecha	Escala	Entrada
31	PEUGEOT_P	Numérico	3	1	PEUGEOT	Ninguna	Ninguna	8	Derecha	Escala	Entrada
32	RENAULT_P	Numérico	3	1	RENAULT	Ninguna	Ninguna	8	Derecha	Escala	Entrada
33	ROVER_P	Numérico	3	1	ROVER	Ninguna	Ninguna	8	Derecha	Escala	Entrada
34	SEAT_P	Numérico	3	1	SEAT	Ninguna	Ninguna	8	Derecha	Escala	Entrada

Fuente: SPSS 20 IBM

**Figura 13.23. Visor de datos de AUTOS Preferencias. Sav. Caso preferencias.**

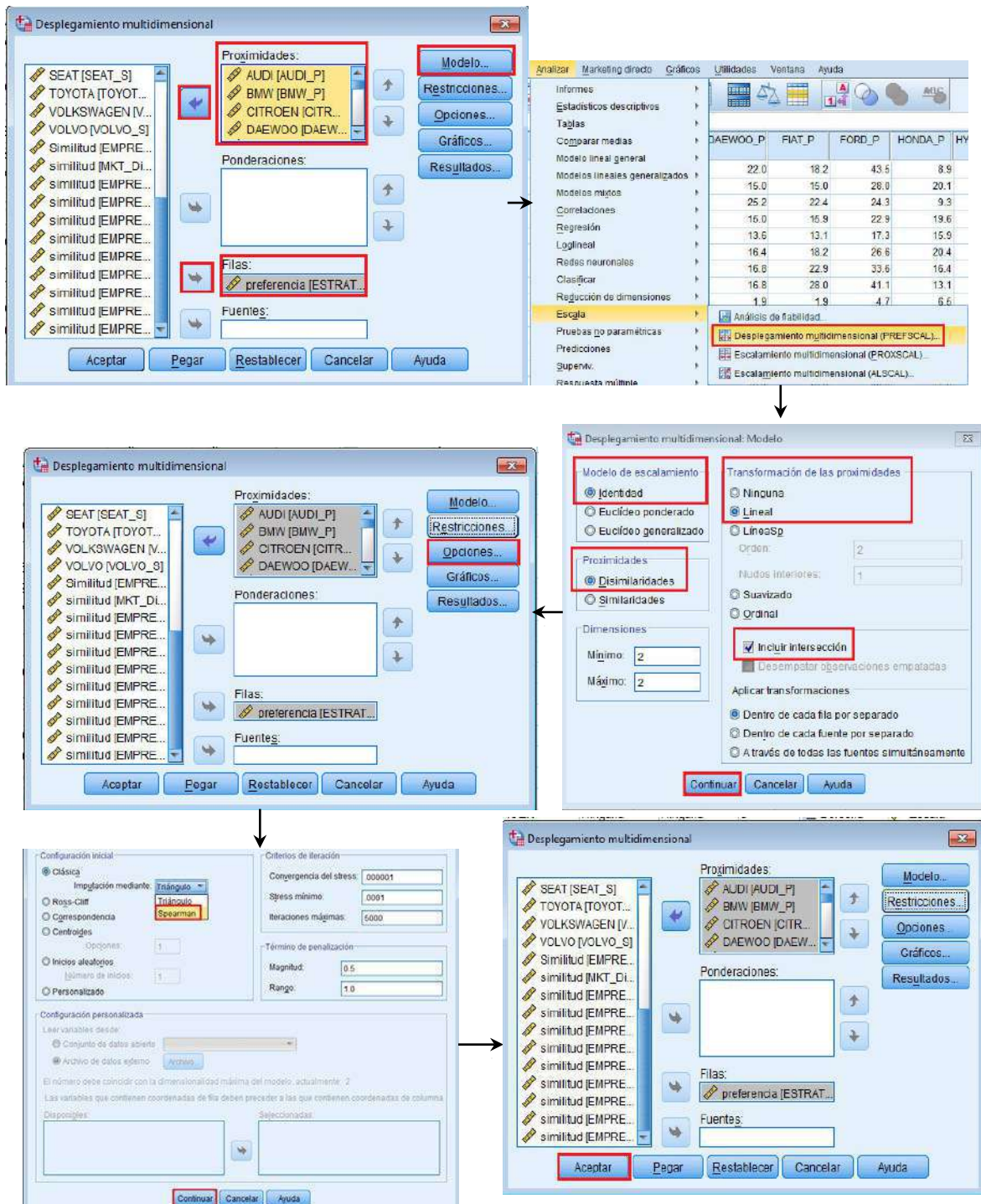
	ESTRATEGIA_P	AUDI_P	BMW_P	CITROEN_P	DAEWOO_P	FIAT_P	FORD_P	HONDA_P	HYUNDAI_P	MERCEDES_P	NISSAN_P	OPEL_P	PEUGEOT_P
1	Producir más autos	9.8	7.0	28.5	22.0	18.2	43.5	8.9	19.2	7.0	9.3	17.8	15.0
2	Tecnología	39.7	40.7	15.9	15.0	15.0	28.0	20.1	15.9	45.3	17.3	19.6	15.4
3	Autos económicos	7.9	7.9	52.8	25.2	22.4	24.3	9.3	22.9	6.5	12.6	13.6	15.9
4	Calidad	49.1	50.0	18.2	15.0	15.9	22.9	19.6	16.4	52.3	18.7	22.0	19.2

Fuente: SPSS 20 IBM

**Paso 4: Ejecución y ajuste**

**Teclar: Analizar->Escala->Despliegamiento multidimensional (PREFSCAL)-> Proximidades: seleccionar variables métricas de similitud (AUDI\_P, BMW\_P, etc.)->Filas: seleccionar la variable categórica (preferencia ESTRATEGIA)-> Modelo->Modelo de escalamiento: Identidad; Proximidades: Disimilaridades; Transformación de las proximidades: Lineal; Incluir intersección->Continuar->Opciones->Configuración inicial ; Clásica; Imputación mediante: Spearman->Continuar->Aceptar. Ver Figura 13.24.**

Figura 13.24. Proceso de escalamiento multidimensional PREFSCAL



Fuente: SPSS 20 IBM



### Paso 5: Interpretación

La primera tabla que genera SPSS, es el **Resumen de procesamiento de casos**, el cual muestra la cantidad de **objetos de columna** (en nuestro caso, la cantidad de **18** marcas automotrices) así como los objetos de fila (en nuestro caso **15** tipos de estrategias) y finalmente 1 fuente de información (el archivo AUTOS Preferencia.sav). Ver **Figura 13.25**.

**Figura 13.25. Tabla Resumen de procesamiento de casos**

Casos	15
Orígenes	1
Objetos de fila	15
Objetos de columna	18

Fuente: SPSS 20 IBM

Como parte de los Diagnósticos del análisis es generada la tabla **Historial de iteraciones**, la que en este caso informa que el algoritmo converge después de **157** iteraciones, con una tensión final penalizada de **0.7448908**. Ver **Figura 13.26**

**Figura 13.26. Historial de iteraciones**

### Diagnósticos del análisis

Iteración	Stress penalizado	Diferencia	Stress	Penalización
0	.9497432		.3965137	2.2748575
157	.7448908	7E-7 <sup>a</sup>	.2439108	2.2748575

a. Diferencia en valores de Stress penalizado consecutivos menor que el criterio DIFFSTRESS.

Fuente: SPSS 20 IBM

Asimismo, se genera la tabla **Medidas**, donde se confirma de nuevo la convergencia del algoritmo después de **157** iteraciones, con una tensión final penalizada de **0.7448908**. Así también se observa que los **coeficientes de variación** y el **índice de Shepard** son suficientemente altos y los **índices de DeSarbo** son suficientemente bajos como para sugerir que no existen problemas de degeneración.

**Figura 13.26. Tabla Medidas**

Iteraciones		157
Valor de función final		.7448908
Partes del valor de función	Parte de Stress	.2439108
	Parte de penalización	2.2748575
Maldad de ajuste	Stress normalizado	.0594401
	Stress-I de Kruskal	.2438035
	Stress-II de Kruskal	1.1695729
	S-Stress-I de Young	.3647414
	S-Stress-II de Young	.4972448
Bondad de ajuste	Dispersión explicada	.9405599
	Varianza explicada	.7336655
	Órdenes de preferencia recuperados	.8579521
	Rho de Spearman	.8366628
	Tau-b de Kendall	.6966524
	Coeficientes de variación	Variación de las Proximidades
Variación de las Proximidades transformadas		.3944354
Variación de las Distancias		.3929293
Índices de degeneración	Suma de cuadrados de los índices de entremezclado de DeSarbo	.1434062
	Índice de no-degeneración aproximada de Shepard	.6313725

Fuente: SPSS 20 IBM

Tablas y gráficos asociados a columnas y filas son generados por el **SPSS**, que permiten apreciar visualmente mejor los resultados; por ejemplo la tabla **Coordenadas de fila finales**, la cual muestra en **2 dimensiones**, para nuestro caso de **cómo las estrategias de introducción**, se definen mejor. Por ejemplo, tomando de referencia valores **>10** en **dimensión 1** observamos que la estrategia **Implantarse en España**, es la que tiene el valor más alto evidenciando la necesidad de la marca por incursionar en dicho país; a su vez, el resto de las estrategias como Producir más autos, Economía, Ecológicos, Gente joven, Presentar nuevos modelos y Plazos de entrega, nos están dando el indicativo a una campaña de administración estratégica de mercadotecnia orientada a los jóvenes con mayor variedad de autos y con mayores facilidades de adquisición, complementada con la dimensión 2 que trata de manera preponderante a la Tecnología y la Calidad que deberá reforzar la campaña hacia el mercado de los jóvenes. **Ver Figura 13.27 y Figura 13.28.**

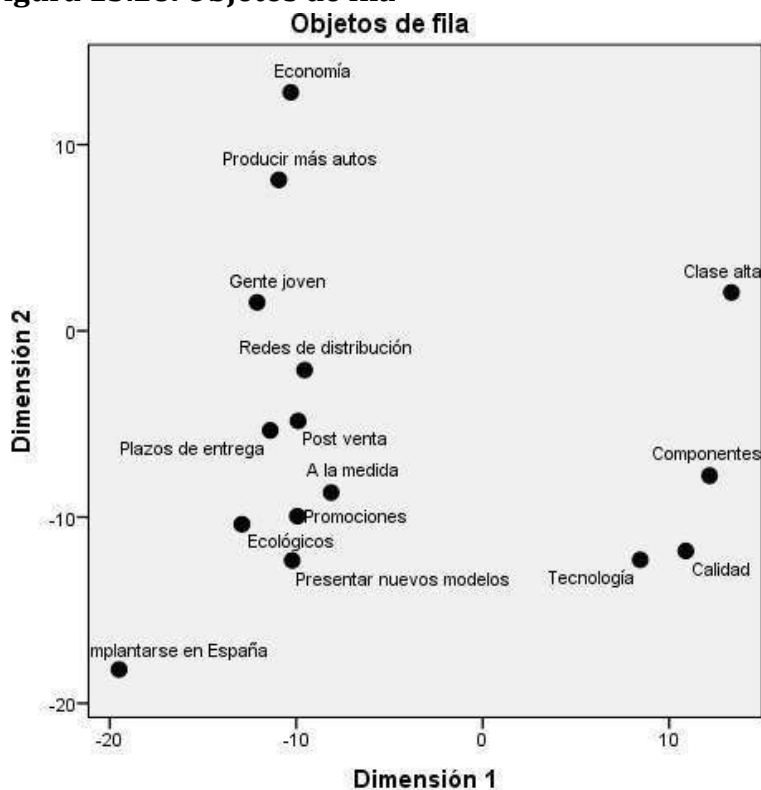
**Figura 13.27. Coordenadas de fila finales**

**Coordenadas de fila finales**

	Dimensión	
	1	2
Producir más autos	-10.935	8.108
Tecnología	8.448	-12.295
Economía	-10.299	12.805
Calidad	10.894	-11.820
Componentes	12.173	-7.776
Ecológicos	-12.915	-10.382
A la medida	-8.132	-8.685
Gente joven	-12.096	1.531
Clase alta	13.335	2.064
Presentar nuevos modelos	-10.215	-12.327
Plazos de entrega	-11.399	-5.344
Redes de distribución	-9.552	-2.100
Implantarse en España	-19.502	-18.186
Post venta	-9.903	-4.841
Promociones	-9.938	-9.947

Fuente: SPSS 20 IBM

**Figura 13.28. Objetos de fila**



Fuente: SPSS 20 IBM

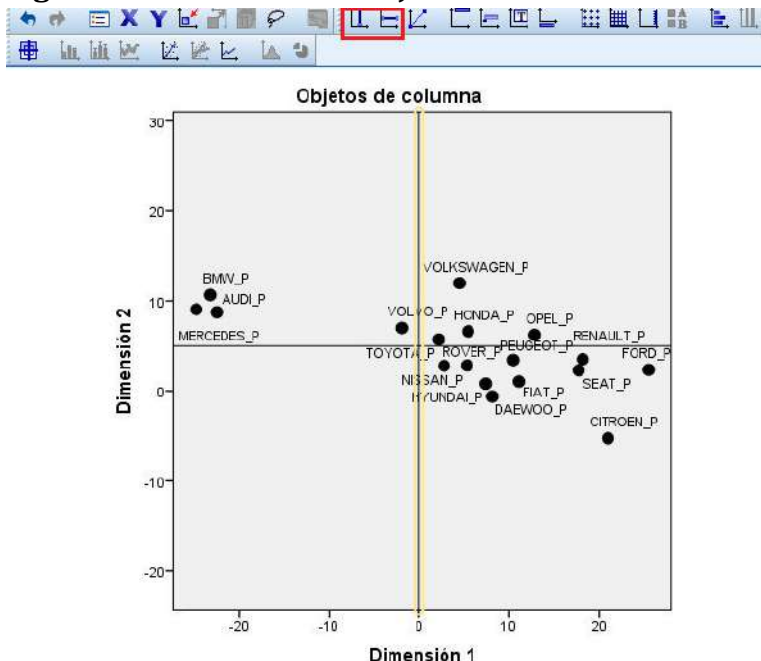
Complementario a lo anterior, tenemos las percepciones de cómo están consideradas competitivamente las marcas, enfatizando sus diferencias para resaltar las preferencias de los encuestados. En este caso, se tienen valores extremos que vale la pena explicar. Por ejemplo, observe como las marcas: AUDI, BMW y MERCEDES se agrupan en la parte izquierda y el resto del grupo a la derecha, resaltando FORD, SEAT y CITROEN. Ver **Figura 13.29** y **Figura 13.30**.

**Figura 13.29. Coordenadas de columna finales**

	Dimensión	
	1	2
AUDI_P	-22.502	8.743
BMW_P	-23.223	10.655
CITROEN_P	20.995	-5.250
DAEWOO_P	8.099	-619
FIAT_P	11.068	1.047
FORD_P	25.524	2.347
HONDA_P	5.445	6.620
HYUNDAI_P	7.396	.809
MERCEDES_P	-24.775	9.054
NISSAN_P	5.314	2.832
OPEL_P	12.848	6.226
PEUGEOT_P	10.447	3.445
RENAULT_P	18.179	3.522
ROVER_P	2.769	2.794
SEAT_P	17.740	2.292
TOYOTA_P	2.145	5.714
VOLKSWAGEN_P	4.482	11.971
VOLVO_P	-1.913	6.994

Fuente: SPSS 20 IBM

**Figura 13.30. Gráfico de objetos de columna**

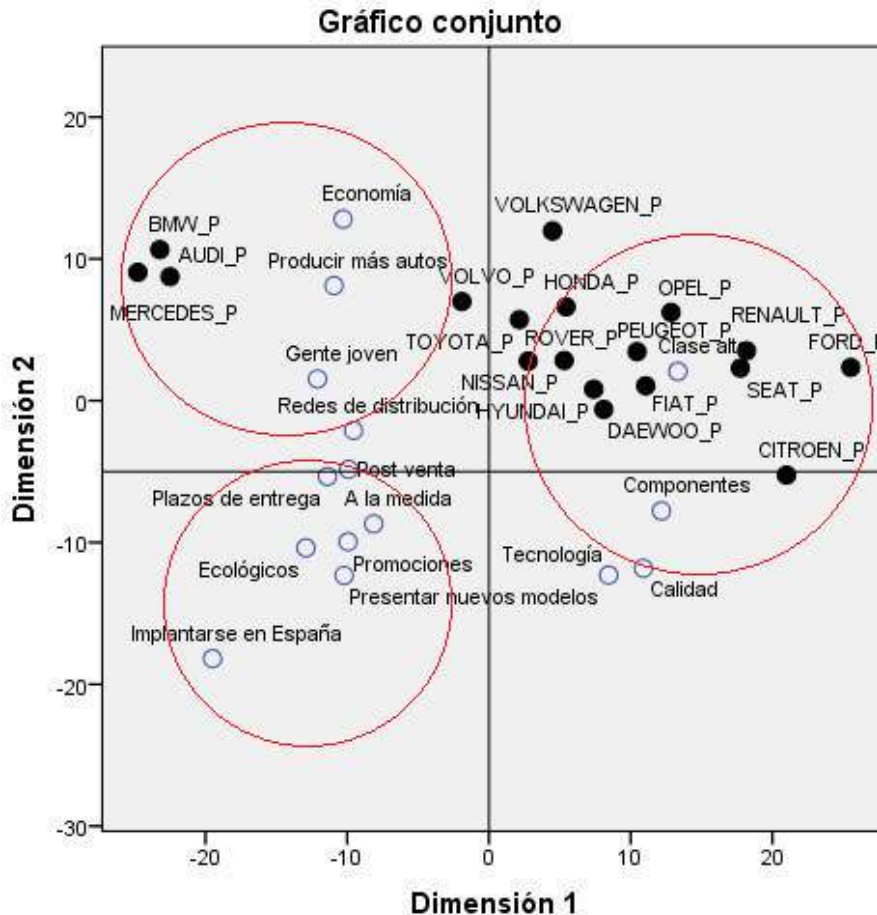


Nota: recuerde colocar el plano de referencia para mejor ubicación de los objetos (hacer doble click en figura y seleccionar íconos de división horizontal y vertical)

Fuente: SPSS 20 IBM

Recuerde que estas tablas y gráficos se realizaron enfatizando las diferencias entre las marcas y las estrategias para hacer evidentes las preferencias, de forma tal que es posible visualizar en un Gráfico de conjunto ambas referencias. Ver **Figura 13.31**.

**Figura 13.31. Gráfico del conjunto del espacio común**



Fuente: SPSS 20 IBM

Tomando en cuenta lo anterior, que las marcas como BMW, AUDI, MERCEDES deberán verificar la pertinencia de incursionar en mercado para gente joven, con autos económicos, produciendo mayor cantidad de autos. O que marcas como FORD, SEAT, CITROEN por ejemplo, hagan lo propio en sectores de clase alta atendiendo los componentes para ese sector. Por último, es importante observar que existe una gran oportunidad para que una nueva marca incursione en España, con estrategias de plazos de entrega, servicio post venta, promociones, ecología, y nuevos modelos a la medida.

### 13.11. Análisis de correspondencias

Al momento, se han discutido las **técnicas de descomposición tradicionales** del análisis multidimensional, pero ¿y las **técnicas de composición**? En el pasado, muchas de las aplicaciones se basan en el **análisis discriminante o el análisis**



**factorial.** Pero desarrollos recientes han combinado aspectos de ambos métodos y del análisis multidimensional para configurar nuevos instrumentos de elaboración de mapas perceptuales. El **análisis de correspondencias** es una técnica de interdependencia que se ha ido haciendo más popular para la reducción dimensional y la elaboración de mapas perceptuales [Carroll et al. 1986, Carroll et al. 1987, Greenacre 1984, Hoffman y George 1986, Lebart y Warwick 1984]. **Es una técnica de composición** debido a que el mapa perceptual se basa en la **asociación entre objetos y un conjunto de características descriptivas o atributos especificados por el investigador.** El **análisis factorial** es el más parecido, entre las **técnicas de composición,** pero el **análisis de correspondencia va más allá del análisis factorial.** Su aplicación más directa es la representación de la **“correspondencia” de categorías de variables, particularmente aquellas medidas en escalas de medida nominales.** Esta correspondencia es **la base del desarrollo de los mapas perceptuales.** Los beneficios del análisis de correspondencias se basan en sus capacidades únicas para representar filas y columnas, por ejemplo, **etiquetas y atributos, en un mismo espacio.**

De acuerdo a Meulman y Heiser (2014), uno de los objetivos del análisis de correspondencias es describir las relaciones existentes **entre 2 variables nominales,** recogidas en una tabla de correspondencias, sobre un espacio de pocas dimensiones, mientras que al mismo tiempo se describen las relaciones entre las categorías de cada variable. Para cada variable, las distancias sobre un gráfico entre los puntos de categorías reflejan las relaciones entre las categorías, con las categorías similares representadas próximas unas a otras. La proyección de los puntos de una variable sobre el vector desde el origen hasta un punto de categoría de la otra variable describe la relación entre ambas variables. El análisis de las tablas de contingencia a menudo incluye examinar los perfiles de fila y de columna, así como contrastar la independencia a través del estadístico de **Chi-cuadrado.** Sin embargo, el número de perfiles puede ser bastante grande y la prueba de **Chi-cuadrado** no revelará la estructura de la dependencia. El procedimiento **Tablas de contingencia** ofrece varias medidas y pruebas de asociación pero no puede representar gráficamente ninguna relación entre las variables. El **análisis factorial** es una técnica típica para describir las relaciones existentes entre variables en un espacio de pocas dimensiones. Sin embargo, el **análisis factorial** requiere datos de intervalo y el número de observaciones debe ser cinco veces el número de variables. Por su parte, el **análisis de correspondencias** asume que las variables son nominales y permite describir las relaciones entre las categorías de cada variable, así como la relación entre las variables. Además, el análisis de correspondencias se puede utilizar para analizar cualquier tabla de medidas de correspondencia que sean positivas. Una tabla de correspondencias es una tabla de doble clasificación cuyas casillas contienen alguna medida de correspondencia entre las filas y las columnas. La medida de correspondencia puede ser cualquier indicación de la similaridad, afinidad, confusión, asociación o interacción entre las variables de fila y de columna. Un tipo muy habitual de tabla de correspondencias es una tabla de contingencia, en la que las casillas contienen las frecuencias. Estas tablas se pueden obtener con facilidad mediante el procedimiento Tablas de contingencia. Sin embargo, una tabla de contingencia no proporciona siempre una imagen clara de la naturaleza de la relación entre las dos

variables. Así ocurre especialmente si las variables de interés son nominales (sin ningún orden o rango inherente) y contienen numerosas categorías. Las tablas de contingencia pueden indicarle que las frecuencias observadas de las casillas difieren considerablemente de los valores esperados en una tabla de contingencia de ocupación y desayuna cereales, pero puede ser difícil determinar qué grupos de ocupaciones tienen gustos similares o cuáles son estos gustos. Continuando con Meulman y Heiser (2014), el análisis de correspondencias permite examinar la relación entre dos variables nominales de manera gráfica en un espacio multidimensional. Se calculan las puntuaciones de fila y de columna y se generan los gráficos basados en las puntuaciones. Las categorías que son similares entre sí aparecen juntas en los gráficos. De esta manera, es fácil ver las categorías de una variable que son similares entre sí o las categorías de las dos variables que están relacionadas. El procedimiento Análisis de correspondencias también permite ajustar puntos suplementarios en el espacio definido por los puntos activos. Si el orden de las categorías de acuerdo con sus puntuaciones no es deseable o se opone a la intuición, se pueden imponer restricciones de orden imponiendo que sean iguales las puntuaciones de algunas categorías. Por ejemplo, supongamos que espera que la variable consumo de tabaco cuyas categorías son ninguno, bajo, medio y alto tengan puntuaciones que correspondan a este orden. Sin embargo, si el análisis ordena las categorías como ninguno, bajo, alto y medio, si obliga a que las puntuaciones de alto y medio sean iguales, se conserva el orden de las categorías en sus puntuaciones. La interpretación del análisis de correspondencias en términos de distancias depende del método de normalización utilizado. El procedimiento Análisis de correspondencias se puede utilizar para analizar tanto las diferencias entre las categorías de una variable como las diferencias entre las variables. Con la normalización por defecto, se analizan las diferencias entre las variables de fila y de columna. El algoritmo de análisis de correspondencias puede realizar muchos tipos de análisis. El centrado de las filas y las columnas y el uso de distancias *Chi-cuadrado* corresponden al análisis de correspondencias típico. Sin embargo, el uso de las opciones de centrado alternativo combinado con las distancias euclídeas permite obtener una representación alternativa de una matriz en un espacio de pocas dimensiones. A continuación, veremos tres ejemplos. El primero utiliza una tabla de correspondencias relativamente pequeña para ilustrar los conceptos inherentes al análisis de correspondencias. El segundo ejemplo muestra una aplicación práctica de marketing. El último ejemplo utiliza una tabla de distancias en una aproximación con escalamiento multidimensional. Una característica importante en la técnica, es la normalización, la cual se utiliza para distribuir la inercia sobre las puntuaciones de fila y de columna. Algunos aspectos de la solución de análisis de correspondencias, como los valores propios, la inercia por dimensión y las contribuciones, no cambian con las diferentes normalizaciones. Las puntuaciones de fila y de columna y sus varianzas si se ven afectadas. El análisis de correspondencias tiene varias maneras de distribuir la inercia. Las tres más habituales incluyen la distribución de la inercia únicamente sobre las puntuaciones de fila, la distribución de la inercia únicamente sobre las puntuaciones de columna y la distribución de la inercia simétricamente sobre las puntuaciones de fila como de columna.

1. **Principal por fila.** En la normalización principal por fila, las distancias euclídeas

entre los puntos de fila aproximan las distancias **Chi-cuadrado** entre las filas de la tabla de correspondencias. Las puntuaciones de fila son la media ponderada de las puntuaciones de columna. Las puntuaciones de columna se tipifican para tener una suma ponderada de los cuadrados de las distancias al centroide de 1. Como este método maximiza las distancias entre las categorías de fila, debe utilizar la normalización principal por fila si está interesado principalmente en ver cómo difieren entre sí las categorías de la variable de fila.

2. **Principal por columna.** Por otra parte, es posible que quiera aproximar las distancias **Chi-cuadrado** entre las columnas de la tabla de correspondencias. En este caso, las puntuaciones de columna deben ser la media ponderada de las puntuaciones de fila. Las puntuaciones de fila se tipifican para tener una suma ponderada de cuadrados de las distancias al centroide de 1. Este método maximiza las distancias entre las categorías de columna y se debe utilizar si está interesado principalmente en ver cómo difieren entre sí las categorías de la variable de columna.
3. **Simétrico.** También puede tratar a las filas y las columnas de manera simétrica. Esta normalización distribuye la inercia de manera idéntica sobre las puntuaciones de fila y de columna. Observe que ni las distancias entre los puntos de fila ni las distancias entre los puntos de columna son aproximaciones de las distancias chi-cuadrado en este caso. Utilice este método si está interesado principalmente en las diferencias y las similitudes entre las dos variables. Normalmente, éste es el método preferido para hacer los diagramas de dispersión biespaciales.
4. **Principal.** Una cuarta opción se denomina normalización principal, en la que la inercia se distribuye dos veces sobre la solución, una vez sobre las puntuaciones de fila y una vez sobre las puntuaciones de columna. Debe utilizar este método si le interesan las distancias entre los puntos de fila y las distancias entre los puntos de columna por separado, pero no en cómo están relacionados entre sí los puntos de fila y de columna. Los diagramas de dispersión biespacial no son apropiados para esta opción de normalización y, por tanto, no están disponibles si se ha especificado el método de normalización principal.

### 13.12. Análisis de correspondencias. Ejemplos

#### Paso 1: Objetivos

**Problema 2:** Se tiene una serie de 23 atributos de imagen de café helado, donde los encuestados seleccionaron todas las marcas que quedaban descritas por cada atributo 6 marcas denotadas como: AA, BB, CC, DD, EE y FF para mantener la confidencialidad.(Caso Kennedy et al. 1996). Ver **Figura 13.32 y Figura 13.33**

**Figura 13.32. Visor de datos de COFEE.sav**

	IMAGEN	MARCA	FRECUENCIA
1	Engorda	Marca AA	82
2	Hombres	Marca AA	96
3	Sur de EUA	Marca AA	72
4	Tradicional	Marca AA	101
5	premium	Marca AA	66
6	Saludable	Marca AA	6
7	caffeine	Marca AA	47
8	Nuevo	Marca AA	1
9	attractive	Marca AA	16
10	Severo	Marca AA	60
11	Popular	Marca AA	137
12	cure	Marca AA	49
13	low fat	Marca AA	3
14	children	Marca AA	24
15	Trabajador	Marca AA	96
16	Dulce	Marca AA	27
17	unpopular	Marca AA	1

Fuente: SPSS 20 IBM

**Figura 13.33. Visor de variables de COFEE.sav**

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	IMAGEN	Númérico	4	0		{1, Engorda}...	Ninguna	8	Derecha	Nominal	Entrada
2	MARCA	Númérico	4	0		{1, Marca A...	Ninguna	8	Derecha	Nominal	Entrada
3	FRECUENCIA	Númérico	4	0		Ninguna	Ninguna	8	Derecha	Escala	Entrada
4											
5											
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											

**Etiquetas de valor**

Etiquetas de valor

Valor:

Etiqueta:

Añadir

Cambiar

Eliminar

1 = "Engorda"  
2 = "Hombres"  
3 = "Sur de EUA"  
4 = "Tradicional"  
5 = "Premium"  
6 = "Saludable"

Ortografía...

Aceptar Cancelar Ayuda

Fuente: SPSS 20 IBM

En principio, nos centraremos en cómo están relacionados los atributos entre sí y cómo están relacionadas las marcas entre sí.

## Paso 2: Diseño

Empleo de la técnica de análisis de correspondencias con previa ponderación mediante el campo **FRECUENCIA**, obtener una solución inicial en **5** dimensiones con normalización principal.

## Paso 3: Condiciones de aplicabilidad

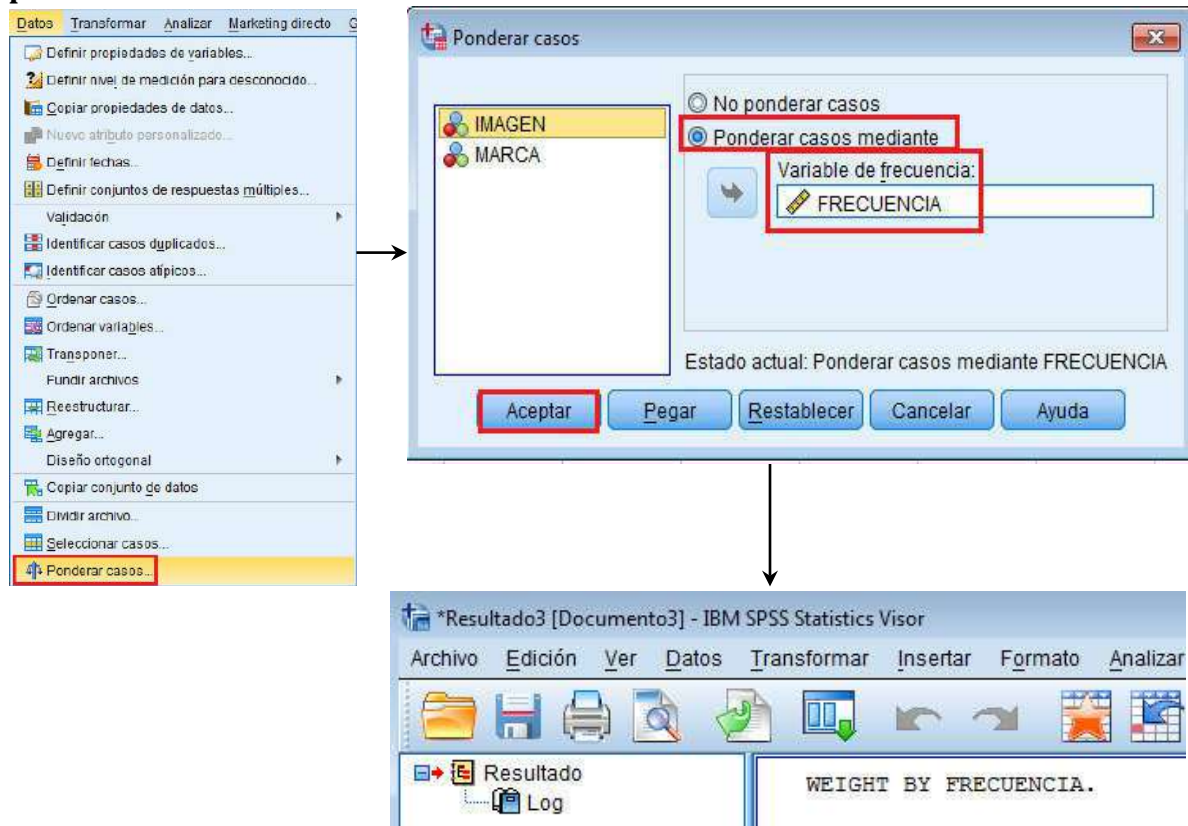
El uso de la normalización principal distribuye la inercia total una vez sobre las filas y una vez sobre las columnas. Aunque esto impide la interpretación del diagrama de dispersión biespacial, es posible examinar las distancias entre las categorías de cada variable.

## Paso 4: Ejecución y ajuste. Caso Normalización principal.

Teclar: **Datos->Ponderar casos->Ponderar casos mediante**; **Variable de frecuencia: FRECUENCIA->Aceptar\***. Ver Figura 13.34

\*Nota: Etapa de ponderación

**Figura 13.34. Proceso de cálculo de análisis de correspondencias. Etapa ponderación**

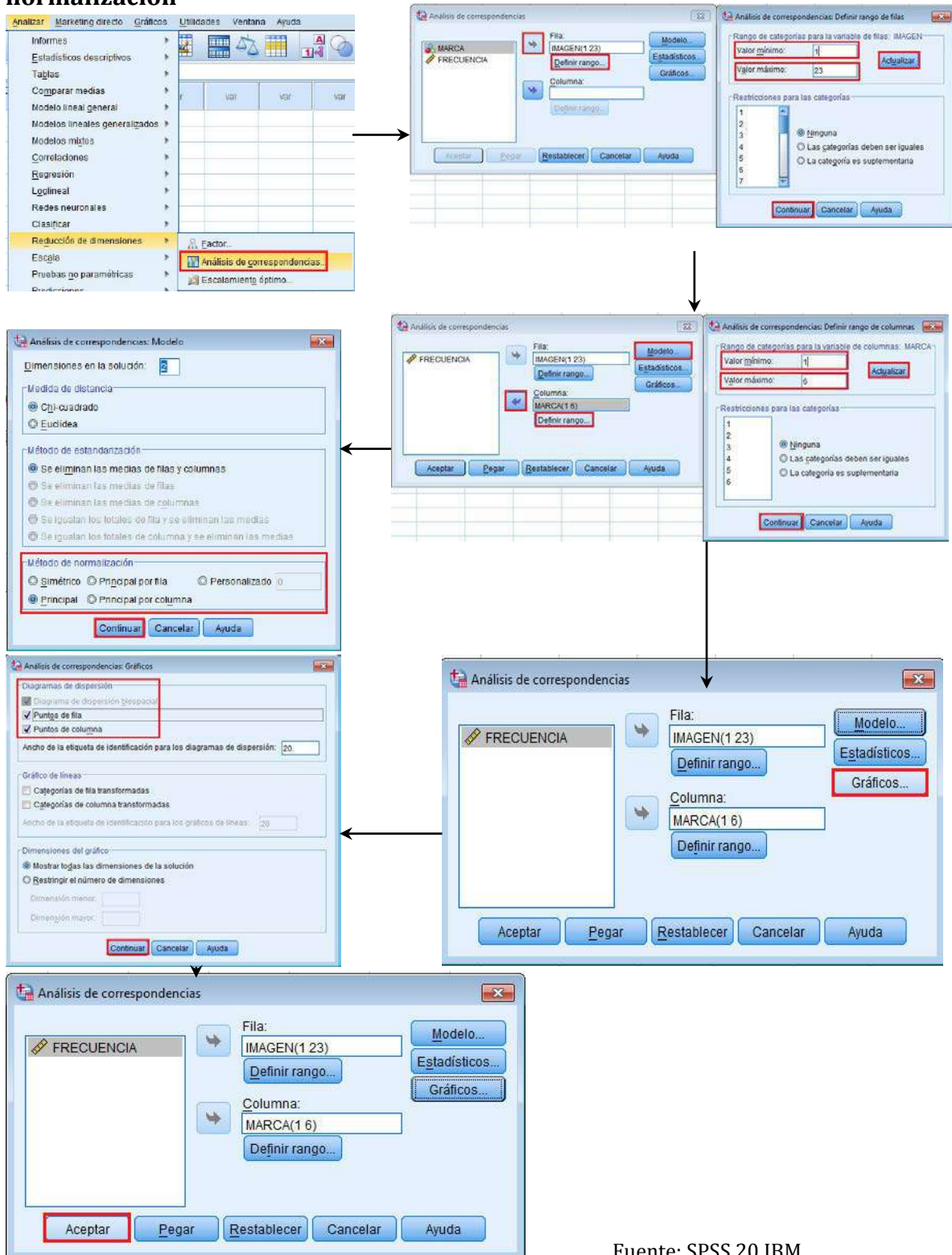


Fuente: SPSS 20 IBM

**Teclear: Analizar > Reducción de dimensiones > Análisis de correspondencias->Fila (IMAGEN)->Definir rango->Rango de categorías para la variable de filas: IMAGEN; Valor mínimo: 1; Valor máximo: 23->Actualizar->Continuar->Columna (MARCA)->Definir rango-> Rango de categorías para la variable de filas: MARCA; Valor mínimo: 1; Valor máximo: 6-> Actualizar->Continuar->Modelo->Método de normalización: Principal->Continuar->Gráficos->Diagramas de dispersión; Puntos de fila y Puntos de columna >Continuar-> Aceptar.Ver Figura 1.35.**



**Figura 13.35. Proceso de cálculo de análisis de correspondencias. Etapa normalización**



Fuente: SPSS 20 IBM

### Paso 5: Interpretación

La primera tabla generada por **SPSS** es la tabla de **Correspondencias**, la cual muestra las frecuencias de los atributos (IMAGEN en este caso) que se tienen registradas por cada una de las marcas. Ver **Figura 13.36**.

**Figura 13.36. Tabla de correspondencias**  
Tabla de correspondencias

IMAGEN	MARCA						
	Marca AA	Marca BB	Marca CC	Marca DD	Marca EE	Marca FF	Margen activo
Engorda	82	78	12	16	76	110	374
Hombres	96	9	0	3	119	11	238
Sur de EUA	72	111	30	13	20	22	268
Tradicional	101	30	1	0	1	53	186
Premium	66	24	14	7	9	76	196
Saludable	6	6	137	93	3	4	249
Cafeína	47	33	14	15	65	43	217
Nuevo	1	11	78	99	15	15	219
Atractivo	16	9	69	55	10	31	190
Severo	60	7	1	2	107	5	182
Popular	137	35	6	4	47	50	279
Remedio	49	10	7	17	26	10	119
Baja grasa	3	2	144	92	0	1	242
Niños	24	44	9	5	9	23	114
Trabajador	96	23	2	3	73	12	209
Dulce	27	21	4	4	25	96	177
No popular	1	18	32	32	21	8	112
Feo	22	32	24	22	20	20	140
Fresco	48	26	27	23	18	25	167
Yuppies	13	14	33	46	8	43	157
Nutritivo	23	17	72	56	7	10	185
Mujeres	19	19	104	73	7	32	254
Menor	3	32	42	73	23	15	188
Margen activo	1012	611	862	753	709	715	4662

Fuente: SPSS 20 IBM

De estos datos, se genera por parte del **SPSS** la tabla Resumen. Ver **Figura 13.37**

**Figura 13.37. Tabla resumen**

Resumen								
Dimensión	Valor propio	Inercia	Chi-cuadrado	Sig.	Proporción de inercia		Confianza para el Valor propio	
					Explicada	Acumulada	Desviación típica	Correlación
								2
1	.711	.506			.629	.629	.009	.132
2	.399	.159			.198	.827	.014	
3	.263	.069			.086	.913		
4	.234	.055			.068	.982		
5	.121	.015			.018	1.000		
Total		.804	3746.968	.000 <sup>a</sup>	1.000	1.000		

a. 110 grados de libertad

Fuente: SPSS 20 IBM



De la tabla Resumen, se observa que la inercia por dimensión muestra la descomposición de la inercia total a lo largo de cada dimensión. Así, **2 dimensiones explican el 83% de la inercia total**. Si se añade una tercera dimensión sólo se añade un **8.6%** a la inercia explicada. Por tanto, puede elegir utilizar una representación en **2 dimensiones**.

Así también, es generada por **SPSS** la tabla **Examen de los puntos de fila** (ver Figura 13.38), la cual aporta una visión general de los puntos de fila muestra las contribuciones de los puntos de fila a la inercia de las dimensiones y las contribuciones de las dimensiones a la inercia de los puntos de fila. Si todos los puntos contribuyen de igual manera a la inercia, las contribuciones serían **0.043**. **Saludable** y **Baja en grasa** contribuyen en una parte importante a la inercia de la **dimensión 1**. **Hombres** y **Severo** contribuyen con las mayores cantidades a la inercia de la **dimensión 2**. **Feas** y **Fresco** contribuyen muy poco a ambas dimensiones (Meulman y Heiser, 2014).

**Figura 13.38. Tabla Examen de los puntos de fila**

**Examen de los puntos de fila<sup>a</sup>**

IMAGEN	Masa	Puntuación en la dimensión		Inercia	Contribución				
		1	2		De los puntos a la inercia de la dimensión		De la dimensión a la inercia del punto		Total
					1	2	1	2	
Engorda	.080	-.514	-.265	.033	.042	.035	.652	.173	.825
Hombres	.051	-.852	.825	.072	.073	.219	.512	.480	.992
Sur de EUA	.057	-.303	-.350	.046	.010	.044	.114	.152	.266
Tradicional	.040	-.703	-.532	.043	.039	.071	.454	.260	.715
Premium	.042	-.444	-.582	.028	.016	.090	.296	.509	.805
Saludable	.053	1.200	.174	.081	.152	.010	.953	.020	.973
Cafeína	.047	-.452	.124	.014	.019	.005	.702	.053	.755
Nuevo	.047	.960	.147	.048	.086	.006	.893	.021	.914
Atractivo	.041	.657	-.056	.019	.035	.001	.911	.007	.918
Severo	.039	-.850	1.002	.070	.056	.246	.404	.560	.964
Popular	.060	-.697	-.042	.038	.058	.001	.771	.003	.774
Remedio	.026	-.389	.266	.009	.008	.011	.446	.209	.655
Baja grasa	.052	1.305	.196	.094	.175	.013	.941	.021	.962
Niños	.024	-.352	-.513	.017	.006	.041	.179	.380	.559
Trabajador	.045	-.785	.477	.040	.055	.064	.693	.255	.948
Dulce	.038	-.519	-.683	.048	.020	.112	.212	.368	.580
No popular	.024	.489	.186	.010	.011	.005	.585	.085	.670
Feo	.030	.006	-.109	.003	.000	.002	.000	.131	.131
Fresco	.036	-.096	-.100	.002	.001	.002	.196	.214	.410
Yuppies	.034	.380	-.301	.012	.010	.019	.392	.246	.637
Nutritivo	.040	.722	.055	.022	.041	.001	.946	.006	.951
Mujeres	.054	.758	-.063	.032	.062	.001	.965	.007	.972
Menor	.040	.579	.063	.023	.027	.001	.593	.007	.600
Total activo	1.000			.804	1.000	1.000			

a. Normalización Principal

Fuente: SPSS 20 IBM

Dos dimensiones contribuyen con una gran cantidad a la inercia para la mayoría de los puntos de fila. Las contribuciones grandes a la primera dimensión de **Saludable**, **Nuevo**, **Atractivo**, **Baja en grasa**, **Nutritivo** y **Mujeres** indican que estos puntos aparecen muy bien representados en una dimensión. Por consiguiente, las

dimensiones superiores contribuyen poco a la inercia de estos puntos, que estarán situados muy cerca del eje horizontal. La **dimensión 2** contribuye sobre todo a **Hombres, Calidad y Severo**. Ambas dimensiones contribuyen muy poco a la inercia de **Sur de EUA y Feo**, por lo que estos puntos aparecen pobremente representados.

Otra tabla que SPSS genera para analizar, es la de **Examen de los puntos de columna**. Ver **Figura 13.39**.

**Figura 13.39. Tabla Examen de los puntos de columna**

**Examen de los puntos columna<sup>a</sup>**

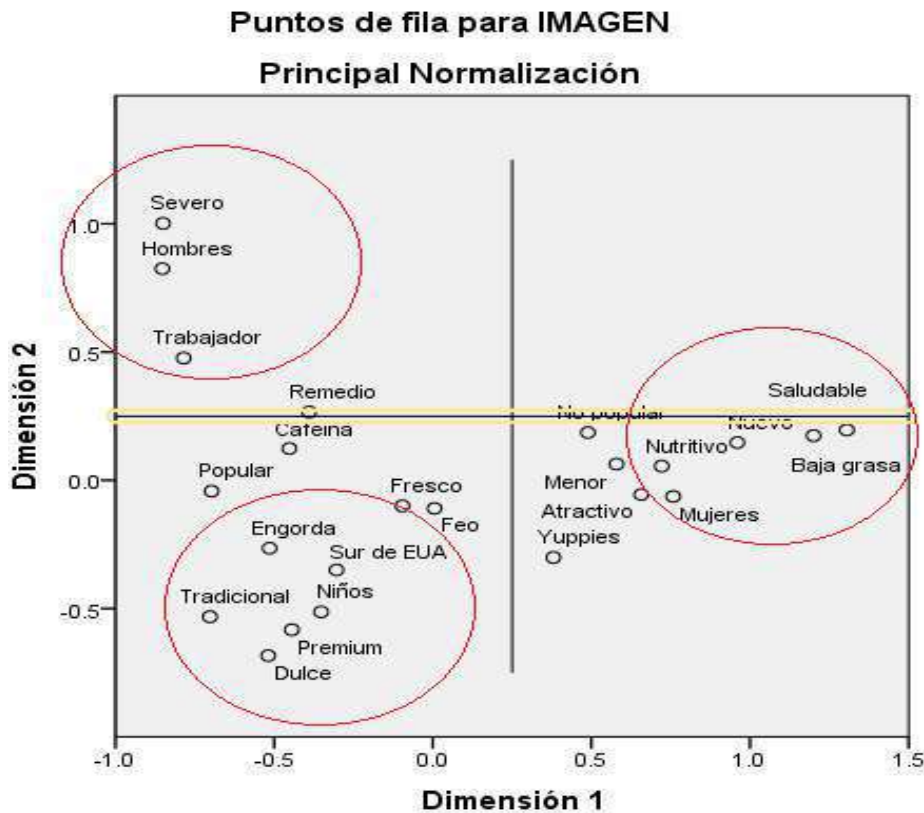
MARCA	Masa	Puntuación en la dimensión		Inercia	Contribución				
		1	2		De los puntos a la inercia de la dimensión		De la dimensión a la inercia del punto		
					1	2	1	2	Total
Marca AA	.217	-.659	.046	.127	.187	.003	.744	.004	.748
Marca BB	.131	-.284	-.404	.078	.021	.134	.135	.272	.407
Marca CC	.185	.996	.076	.193	.362	.007	.951	.006	.957
Marca DD	.162	.915	.101	.146	.267	.010	.928	.011	.939
Marca EE	.152	-.651	.706	.153	.127	.477	.420	.494	.914
Marca FF	.153	-.343	-.618	.107	.036	.369	.169	.550	.718
Total activo	1.000			.804	1.000	1.000			

a. Normalización Principal

Fuente: SPSS 20 IBM

La visión general de los puntos de columna muestra las contribuciones que implican a los puntos de columna. Las **Marcas CC y DD** contribuyen sobre todo a la **dimensión 1**, mientras que las **Marcas EE y FF** explican una gran cantidad de la inercia para la **dimensión 2**. Las **Marcas AA y BB** contribuyen muy poco a ambas dimensiones. En dos dimensiones, todas las marcas salvo **BB** están bien representadas. Las **Marcas CC y DD** están bien representadas en una dimensión. La **dimensión 2** contribuye a las **Marcas EE y FF** con sus mayores cuantías. Observe que la **Marca AA** está bien representada en la **dimensión 1**, pero no tiene una contribución muy alta a dicha dimensión. Cabe destacar, que **SPSS** genera gráficos de soporte para visualizar mejor los resultados, donde por ejemplo el **Gráfico de puntos de fila** (ver **Figura 13.39**), muestra que **Fresco y Feo** están muy cerca del origen, lo que indica que difieren muy poco del perfil de fila medio. Surgen así tres clasificaciones generales. Situado en la parte izquierda superior del gráfico, **Severo, Hombres y Trabajador son similares entre sí**. La parte inferior izquierda contiene **Engorda, Sur de EUA, Niños, Tradicional, Premium, Dulce**. Por el contrario, **Saludable, Baja en grasa, Nutritivo, Mujeres y Nuevo** se agrupan en la parte izquierda del gráfico. Ver **Figura 13.40**.

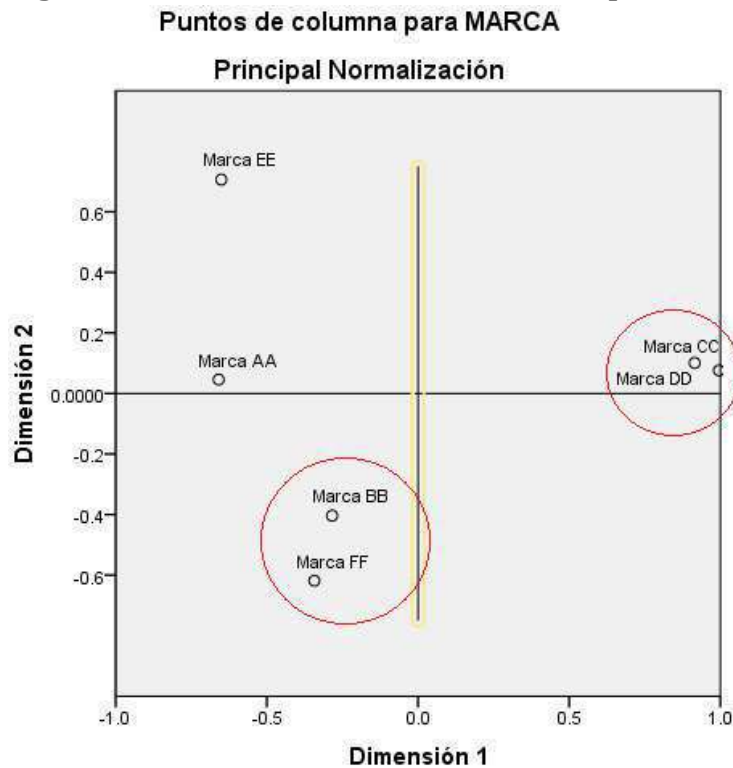
Figura 13.40. Gráfica Puntos de fila para IMAGEN



Fuente: SPSS 20 IBM

La última gráfica generada por SPSS es la de **Puntos columna para MARCA**. Observe que **todas las marcas están lejos del origen, por lo que NO hay ninguna marca que sea similar al centroide global**. Las **Marcas CC y DD** se agrupan juntas a la derecha, mientras que las **Marcas BB y FF** se agrupan en la mitad inferior del gráfico. Las **Marcas AA y EE** **NO** son similares a ninguna otra marca. Ver Figura 13.41.

**Figura 13.41. Gráfica Puntos de columna para MARCA**



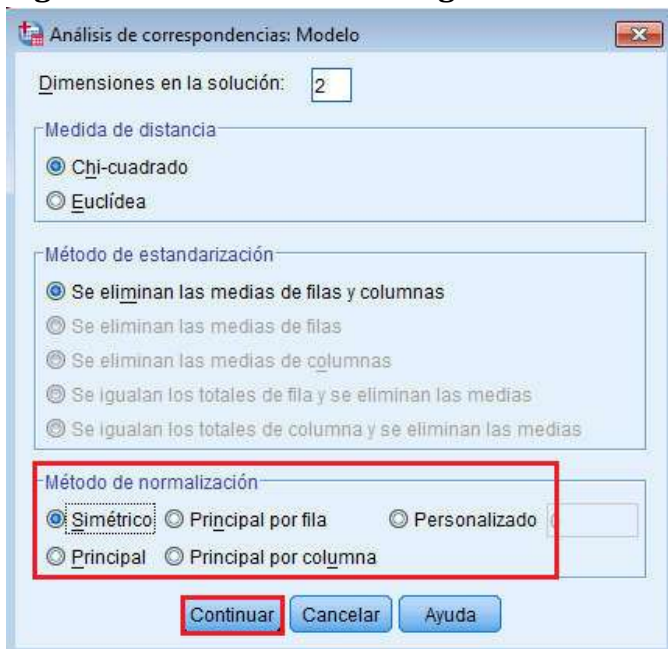
Fuente: SPSS 20 IBM

#### **Paso 4: Ejecución y ajuste. Caso Normalización simétrica.**

¿Cómo están relacionadas las marcas con los atributos de imagen? La normalización principal no puede tratar estas relaciones. **Para centrarnos en cómo están relacionadas las variables entre sí, utilizaremos la normalización simétrica.** En vez de distribuir la inercia **2 veces** (como ocurre en la **normalización principal**), la normalización simétrica **divide la inercia de idéntica manera sobre las filas y las columnas.** Las distancias entre categorías para una única variable **NO** se pueden interpretar, pero las distancias entre las categorías de diferentes variables son significativas.

**Teclear: Analizar > Reducción de dimensiones > Análisis de correspondencias->Fila (IMAGEN)->Definir rango->Rango de categorías para la variable de filas: IMAGEN; Valor mínimo: 1; Valor máximo: 23->Actualizar->Continuar->Columna (MARCA)->Definir rango-> Rango de categorías para la variable de filas: MARCA; Valor mínimo: 1; Valor máximo: 6-> Actualizar->Continuar->Modelo->Método de normalización: Simétrico->Continuar->Gráficos->Diagramas de dispersión; Puntos de fila y Puntos de columna >Continuar-> Aceptar.Ver Figura 13.42.**

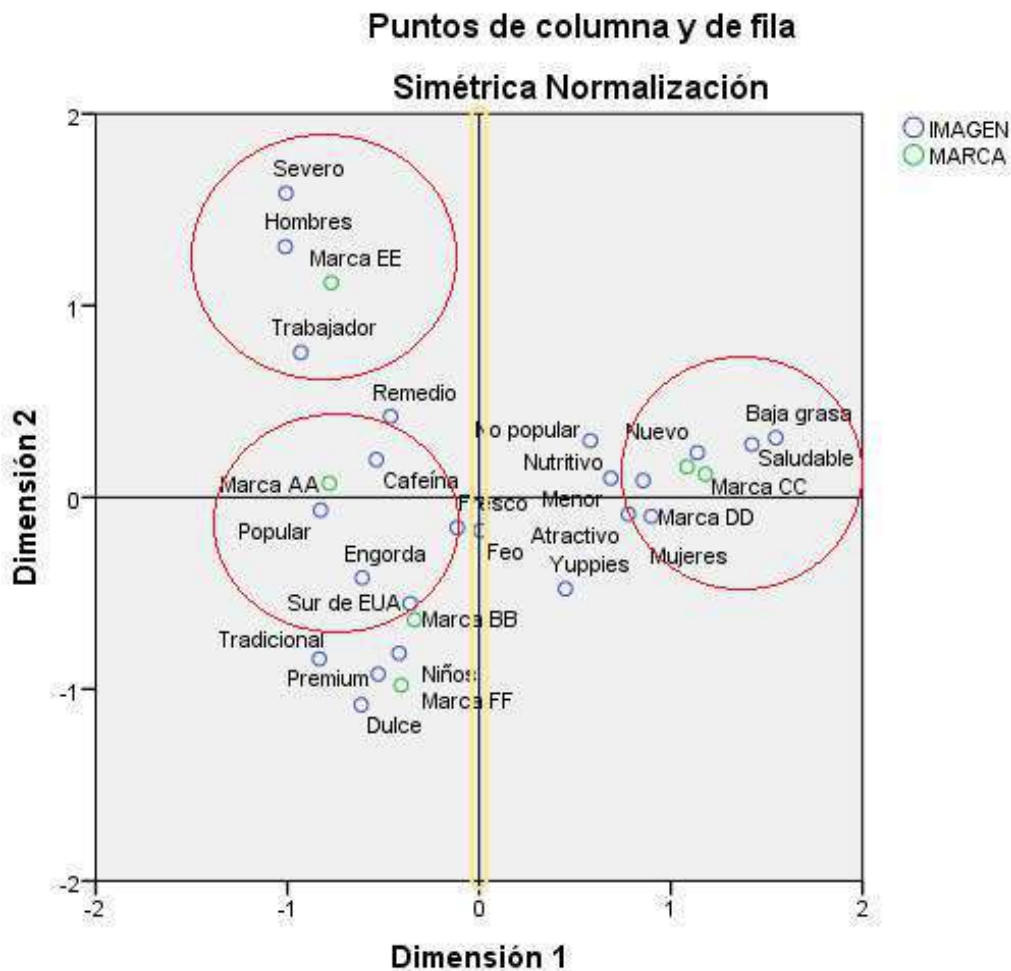
**Figura 13.42. Cuadro de diálogo Modelo**



Fuente: SPSS 20 IBM

En la parte superior del diagrama de dispersión biespacial resultante, la **Marca EE** es la única severa y de la clase trabajadora que resulta atractiva a los hombres. La marca **AA** es la más popular y también se considera la que tiene más cafeína. Las marcas dulces y que engordan incluyen **BB y FF**. Las marcas **CC y DD**, aunque se consideran como nuevas y sanas, también son las menos populares. Ver **Figura 13.43**.

**Figura 13.43.** Diagrama de dispersión biespacial de las marcas y los atributos (normalización simétrica)



Fuente: SPSS 20 IBM

Para continuar con la interpretación, **puede dibujar una línea a través del origen y los 2 atributos de imagen hombres y yupis, y proyectar las marcas sobre esta línea.** Los 2 atributos están opuestos el uno al otro, lo que indica que **el patrón de asociación de las marcas para hombres está invertido en comparación con el patrón de yupis.** Es decir, los hombres son los que están asociados con mayor frecuencia con la Marca **EE** y con menor frecuencia con la marca **CC**, mientras que los **yupis** se asocian con mayor frecuencia con la marca **CC** y con menor frecuencia con la marca **EE**.

## Referencias

- Carroll, J. D., Green, P. E. y Schaffer C. M. (1986), Interpoint Distance Comparisons in Correspondence Analysis. *Journal of Marketing Research* 23 (August): 271-80.
- Carroll, J. D., Green, P. E. y Schaffer, C. M. (1987), Comparing Interpoint Distances in Correspondence Analysis: A Clarification. *Journal of Marketing Research* 24 (November): 445-50.
- Chang, J. J., y Carroll, J. D. (1968), *How to Use PROFIT, a Computer Program for Property Fitting by Optimizing Nonlinear and Linear Correlation*. Unpublished paper, Bell Laboratories, Murray Hill, N.J.
- Chang, J. J., y Carroll, J. D. (1969), *How to use INDSCAL, a Computer Program for Canonical Decomposition of n-way Tables and Individual Differences in Multidimensional Scaling*. Unpublished paper, Bell Laboratories, Murray Hill, N.J.
- Chang, J. J., y Carroll, J. D. (1969), *How to Use MDPREF, a Computer Program for Multidimensional Analysis of Preference Data*. Unpublished paper, Bell Laboratories, Murray Hill, N.J.
- Chang, J. J., y Carroll, J. D. (1972), *How to Use PREFMAP and PREFMAP2-Programs Which Relate Preference Data to Multidimensional Scaling Solution*. Unpublished paper, Bell Laboratories, Murray Hill, N.J.
- Green, P. E. (1975), On the Robustness of Multidimensional Scaling Techniques. *Journal of Marketing Research* 12 (February): 73-81.
- Green, P. E., y Carrnone F. (1969), Multidimensional Scaling: An Introduction and Comparison of Nonmetric Unfolding Techniques. *Journal of Marketing Research* 7 (August): 33-41.
- Green, P. E., Carrnone, F. y Smith, S. M. (1989), *Multidimensional Scaling: Concept and Applications*. Boston: Allyn y Bacon.
- Green, P. E., y Vithala Rao (1972), *Applied Multidimensional Scaling*. New York: Holt, Rinehart and Winston.
- Greenacre, M. J. (1984), *Theory and Applications of Correspondence Analyses*. London: Academic Press.
- Greenacre, M. J. (1989), The Carroli-Grenn- Schaffer Scaling in Correspondence Analysis: A Theoretical and Empirical Appraisal. *Journal of Marketing Research* 26 (August): 358-65.
- Hair, J.F.; Anderson, R.E.; Black, W.C. (1999). *Análisis Multivariante*. 5a. Ed. España:Prentice Hall.
- Hoffman, D. L., y Franke, G. R. (1986), Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research. *Journal of Marketing Research* 23 (August): 213-27.
- Holbrook, M. B., Moore, W. L. y Russell S. W. (1982), Constructing Joint Spaces from Pick- Any Data: A New Tool for Consumer Analysis. *Journal of Consumer Research* 9 (June): 99-105.
- IBM (2011a). *IBM SPSS Statistics Base 20*. EUA. Industrial Business Machines. Recuperado el 20161201 de:



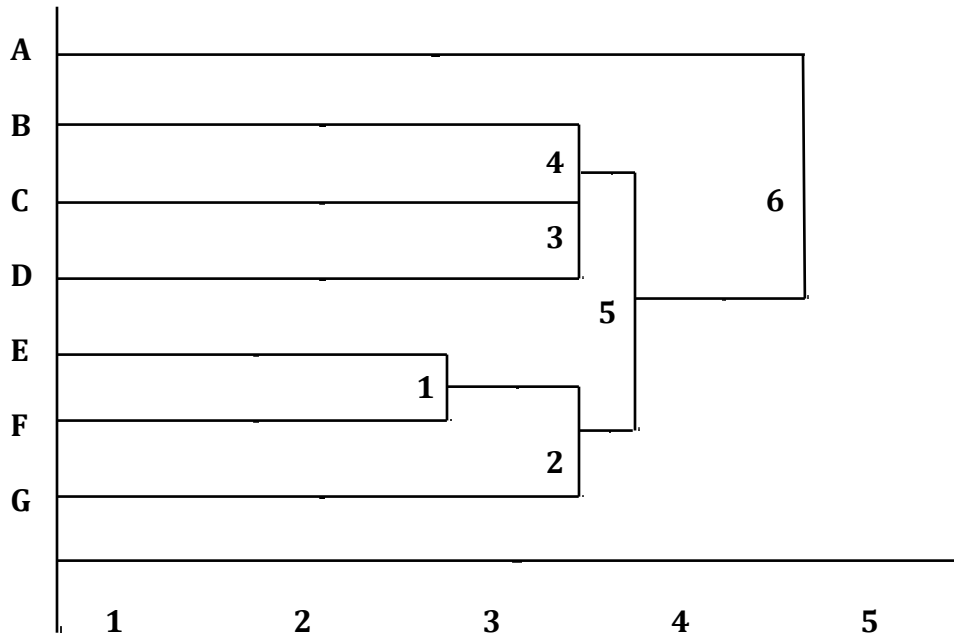
- [ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM SPSS Statistics Base.pdf](ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Base.pdf)
- IBM (2011b). *Guía breve de IBM SPSS Statistics 20*. EUA. Industrial Business Machines.  
Recuperado el 20161201 de:  
[ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM SPSS Statistics Brief Guide.pdf](ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf)
- IBM (2011c). *IBM SPSS Missing Values 20*. EUA. Industrial Business Machines.  
Recuperado el 20161201 de:  
[ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM SPSS Missing Values.pdf](ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Missing_Values.pdf)
- IBM (2012). *IBM SPSS Categories 21*. . EUA. Industrial Business Machines.  
Recuperado el 20161201 de:  
[ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/21.0/es/client/Manuals/IBM SPSS Categories.pdf](ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/21.0/es/client/Manuals/IBM_SPSS_Categories.pdf)
- Kennedy, R., Riquier C. y Sharp. B. (1996). Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing*, 5, 56–70.
- Kruskal, J. B., y Carmone, F. J. (1967), How to Use MDSCAL. Version 5-M, and Other Useful Information. Unpublished paper, Bell Laboratories, Murray Hill, N. J.
- Kruskal, J. B., y Wish, M. (1978), *Multidimensional Scaling*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-0 11, Beverly Hills, Calif.: Sage.
- Levine, J. H. (1979), Joint-Space Analysis of "Pick- Any" Data: Analysis of Choices from an Unconstrained Set of Alternatives. *Psychometrika* 44 (March): 85-92.
- Lingoes, J. C. (1972), *Geometric Representations of Relational Data*. Ann Arbor, Mich.: Mathesis Press.
- Lebart, L., Morineau, A., y Warwick, K. M. (1984), *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. New York: Wiley.
- Maholtra, N. (1987), Validity and Structural Reliability of Multidimensional Scaling. *Journal of Marketing Research* 24 (May): 164-73.
- Market ACTION Research Software, Inc. (1989), MAPWISE: Perceptual Mapping Software. Peoria, Ill.: Business Technology Center, Bradley University.
- Meulman, J.J. y Heiser, W.J. (2014). IBM SPSS Categories 21. Recopilado el 20161213 de: [http://www.sussex.ac.uk/its/pdfs/SPSS Categories 21.pdf](http://www.sussex.ac.uk/its/pdfs/SPSS_Categories_21.pdf)
- Raymond, C. (1974), *The Art of Using Science in Marketing*. New York: Harper y Row.
- Schiffman, S. S., Reynolds, M. L. y Young, F. W. (1981), *Introduction to Multidimensional Scaling*. New York: Academic Press.
- Smith, S. M. (1989), PC-MDS. • *A Multidimensional Statistics Package*. Provo, Utah: Brigham Young University.
- Srinivasan, V., y Schocker, A. D. (1973), Linear Programming Techniques for Multidimensional Analysis of Preferences. *Psychometrika* 38 (September): 337-69.
- Vila-López, N. (2013). *El análisis de escalamiento multidimensional*. Universitat de



València Dpto. de Dirección de Empresas *Juan José Renau Piqueras*. Recopilado el 20161212 de:

<https://wwwyyy.files.wordpress.com/2013/03/escalamiento-multidimensional.pdf>

## Capítulo 14. Análisis Cluster



### 14.1. Análisis cluster ¿qué es?

A menudo, Usted se encontrará con situaciones cuya mejor alternativa de solución, será el determinar **grupos de sujetos que sean homogéneos**, tanto si son objetos, personas u organizaciones, productos/servicios o conductas. En esto, observamos como las ciencias de la administración orientadas a la mercadotecnia, tienen una gran oportunidad al definir las opciones estratégicas, basadas en los grupos determinados de la población, a partir de la **segmentación**. De hecho, segmentar es parte de otras disciplinas como las ciencias naturales (por ejemplo, al crear una taxonomía biológica para la clasificación o reclasificación de varios grupos de animales) a las ciencias sociales (por ejemplo, al analizar varios tipos de perfiles psiquiátricos). Todos estos casos tienen en común, la búsqueda de una estructura "**natural**" que explique a los grupos observados basados en un perfil multivariante. Una de las técnicas para lograrlo y la más utilizada, es el **análisis cluster**, la cual agrupa a los sujetos en **conglomerados**, de forma tal que dichos grupos **son más parecidos entre sí** que los sujetos de otros conglomerados. Lo anterior es un intento de **maximización de la homogeneidad** de los grupos de sujetos la vez que se **maximiza la heterogeneidad** entre los agregados. Se puede afirmar, que se denomina **análisis cluster** a un conjunto de técnicas multivariantes cuyo objetivo principal es agrupar sujetos basándose en las características comunes que los distinguen. La técnica clasifica a los sujetos (por ejemplo, productos/servicios, encuestados, empresas, etc.) de tal forma

que cada sujeto es muy similar a los que hay en el conglomerado, basados en algún criterio de selección definido previamente. Los grupos o conglomerados de sujetos resultantes deben mostrar un **alto grado de homogeneidad interna** (dentro del conglomerado) y un **alto grado de heterogeneidad externa** (entre conglomerados). Con lo anterior, **si la clasificación es acertada, los sujetos dentro de los grupos conglomerados se encontrarán muy próximos entre sí, al representarse gráficamente, y los grupos conglomerados diferentes, estarán muy alejados.**

En todo análisis multivariante, como es este caso, el **valor teórico** es central, aunque en una forma muy diferente del resto de las técnicas multivariantes (Hair et al. 1999). El **valor teórico del análisis cluster** se define como **el conjunto de variables que representan las características utilizadas para comparar sujetos en el análisis cluster**, determinando el “*carácter*” de los sujetos. Es de destacar que **es la única técnica multivariante que no estima el valor teórico empíricamente sino que utiliza el valor teórico especificado por el investigador.** Su objetivo, es la **comparación de sujetos basándose en el valor teórico, no en la estimación del valor teórico en sí misma**, por lo que es factor clave de éxito la definición que dé Usted al valor teórico al usar la técnica. El **análisis cluster** se le conoce también como **análisis de clasificación, construcción de tipología, análisis de clasificación y taxonomía numérica, análisis Q.** Esto se explica en parte, al uso tan extensivo que se tiene como método de agrupación en disciplinas tales como la medicina, la biología, psicología, la ingeniería, sociología, economía, negocios y otras. Todos los métodos tienen una dimensión común: **clasificación de acuerdo a una relación natural** [Bailey, 1994]. Se dice que ésta técnica es comparable al **análisis factorial** en su objetivo de evaluar la estructura, con la diferencia de que **el análisis cluster agrupa sujetos, mientras que el análisis factorial se centra principalmente en la agrupación de variables.** El **análisis clúster** es una herramienta de análisis útil para diferentes situaciones, tales como:

-Una investigación basada en encuestas con un número elevado de observaciones que **no tengan sentido a menos que se clasifiquen en grupos manejables**, en el que el **análisis cluster** puede llevar a cabo objetivamente este procedimiento de **reducción de datos** mediante la **reducción de la información de una población completa o una muestra a información sobre subgrupos pequeños y específicos.**

-Una investigación que recopile las “*actitudes*” de una población mediante la identificación de los principales grupos de la población, entonces es posible hacer reducción de los datos de la población completa a **perfiles de ciertos grupos.** Así, Usted tendrá una descripción más concisa y comprensible de las observaciones, con mínima pérdida de información.

-Desarrollar las hipótesis concernientes a la naturaleza de los datos o para examinar las hipótesis previamente establecidas. Por ejemplo, si Usted plantea que las “*actitudes*” hacia la compra de ciertos tipos de smartphones podrían utilizarse para separar a los compradores en segmentos por grupos conglomerados basados en características comunes, que de haberlos, pueden perfilarse mediante diferencias y similitudes demográficas.

Estos son algunos ejemplos a los que puede añadirse clasificaciones psicológicas basadas en la personalidad en situaciones de crisis para determinar distintos tipos de liderazgo, o diferentes situaciones competitivas que generen diferentes modelos

de innovación, todo esto, basados en la clasificación de sujetos, con análisis de similitudes y diferencias entre productos/servicios nuevos y su evaluación del rendimiento de empresas para identificar agrupaciones basadas en las estrategias de las empresas u orientaciones estratégicas. El resultado es una gran cantidad de aplicaciones en casi todas las áreas de investigación, creando no sólo una riqueza de conocimiento en el uso del análisis de conglomerados sino también la necesidad de una mejor comprensión de la técnica para minimizar las **malas prácticas**.

A pesar de los beneficios de la técnica, aún existen algunos inconvenientes ya que el **análisis cluster**:

- Puede caracterizarse como **descriptivo, ateórico y no inferencial**
- No tiene bases estadísticas sobre las cuales deducir inferencias estadísticas para una población a partir de una muestra, y**
- Se utiliza básicamente como una técnica de exploratoria.**
- Las soluciones no son únicas**, en la medida en que la pertenencia al conglomerado para cualquier número de soluciones depende de muchos elementos del procedimiento y se pueden obtener muchas soluciones diferentes variando uno o más de estos elementos.
- Siempre creará conglomerados**, a pesar de la existencia de una “*auténtica*” estructura en los datos.
- La solución por esta técnica es totalmente dependiente de las variables utilizadas** como base para la medida de similitud.
- La adición o anulación de variables relevantes tienen un impacto importante en la solución resultante.** Por tanto, deberá tener un alto nivel de cuidado al evaluar el impacto de cada decisión involucrado en el desarrollo del análisis al usar la técnica.

#### 14.2. Análisis cluster ¿cómo funciona?

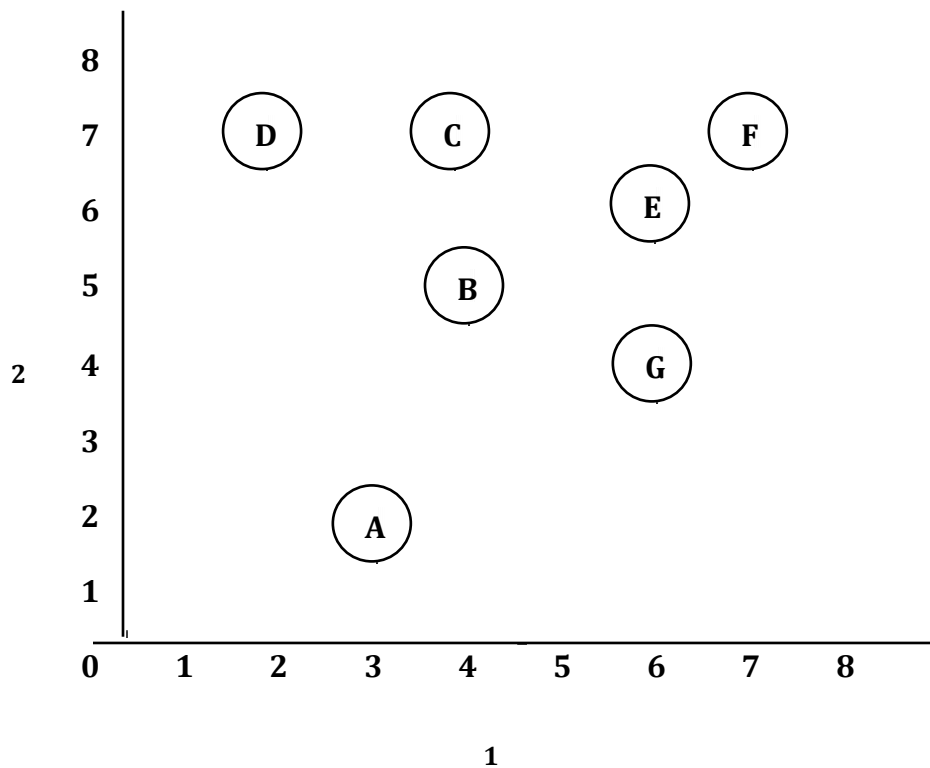
Es posible ejemplificarlo mediante un simple caso bivariante, en el que quizás un investigador está interesado en descubrir los segmentos de mercado en una población reducida basándose en las pautas de diseño: de producto y diseño en la atención de servicio. Así, se procede a seleccionar una muy reducida **muestra de 7 encuestados** como contraste de prueba para verificar cómo se aplica el **análisis cluster**. De esta forma, se miden **2 medidas de diseño: de producto ( $X_1$ ) y servicio ( $X_2$ )** para cada encuestado **en escala de 0-10**. Los valores de cada uno de los siete encuestados se muestran en la **Figura 14.1a y Figura 14.1b**, junto con un diagrama de dispersión representando cada observación en dos variables.

**Figura 14.1a. Valores de datos de 7 observaciones basados en las 2 variables de aglomeración ( $X_1, X_2$ ).**

Variables cluster	Encuestas						
	A	B	C	D	E	F	G
1	3	4	4	2	6	7	6
2	2	5	7	7	6	7	4

Fuente: propia

**Figura 14.1b. Gráfico de dispersión de 7 observaciones basados en las 2 variables de aglomeración ( $X_1$ ,  $X_2$ ).**



Fuente: propia

Considere que el objetivo de la técnica, es hacer una definición de la estructura de datos, buscando y colocando las observaciones más similares en grupos y para lograrlo, se deberá resolver **3 cuestiones básicas**:

1. **¿Cómo medir la similitud?** Para responderlo, determinamos requerir un método de observaciones que sean simultáneamente comparadas sobre las **2 variables de aglomeración ( $X_1$  y  $X_2$ )**. Son posibles varios métodos, como la **correlación entre sujetos**, una **medida de asociación utilizada en otras técnicas multivariantes** o por ejemplo, **medir su proximidad en un espacio bidimensional** de tal forma que la distancia entre las observaciones indica similitud.
2. **¿Cómo formar los conglomerados?** No importa cómo medir la similitud, el procedimiento también debe **agrupar aquellas observaciones que son más similares dentro de un conglomerado** y debe determinar la pertenencia al grupo de cada observación.
3. **¿Cuántos grupos formar?** La tarea fundamental es evaluar la similitud "*media*" dentro de los conglomerados, de forma tal que **a medida que la media aumenta, el conglomerado se hace menos similar**. Usted se enfrentará así, a la siguiente situación: "*pocos conglomerados frente a menos homogeneidad*". Toda estructura simple, al tratarse con **parsimonia**, produce **el menor número de**

**conglomerados posible.** A medida que el número de conglomerados disminuye, la **homogeneidad dentro de los conglomerados necesariamente disminuye.** Es así, que Usted debe **buscar un equilibrio** entre la definición de estructuras más básicas (**pocos conglomerados**) que todavía mantienen el necesario nivel de similitud dentro de los grupos conglomerados. Una vez que resuelva el tener procedimientos para cada asunto, podemos realizar el **análisis cluster.**

### 14.3. Análisis cluster. Medición de la similitud y creación de conglomerados

Para nuestro caso ejemplo de las **7 observaciones** (encuestados A-G) utilizamos procedimientos sencillos para cada uno de los asuntos, de la siguiente forma:

#### 1. Similitud

- La similitud será medida de acuerdo con la distancia Euclidiana (una línea recta) entre cada par de observaciones. La **Figura 14.2** contiene medidas de proximidad entre cada uno de los **7 encuestados.**

**Figura 14.2. Tabla de proximidad de distancia euclidiana entre observaciones**

	Observaciones						
Observaciones	A	B	C	D	E	F	G
A	-						
B	4.162	-					
C	6.099	3.000	-				
D	6.099	3.828	3.000	-			
E	6.000	3.236	3.236	5.123	-		
F	<b>7.403</b>	4.606	4.000	6.000	<b>2.414</b>	-	
G	4.606	3.236	4.606	6.000	3.000	4.162	-

Fuente: propia

- Al aplicar la distancia euclidiana como medida de proximidad, se debe recordar que **las distancias más pequeñas señalan mayor similitud**, de tal forma que las observaciones **E y F** son las más parecidas (**2.414**), y **A y F** son las más diferentes (**7.403**).

#### 2. Conglomerados.

- Existen varios métodos, sin embargo, para nuestro propósito de demostración utilizamos la regla simple: **identificar las dos observaciones más parecidas (cercanas) que no están en el mismo conglomerado y combinar éstas.**
- Aplicamos esta regla repetidas veces**, comenzando con cada observación en su propio grupo "**conglomerado**"
- Combinando 2 grupos conglomerados a un tiempo hasta que todas las observaciones estén en un único conglomerado.** A esto se le conoce como **procedimiento jerárquico** dado que **opera paso a paso** para formar un **rango completo de soluciones cluster.** También se le reconoce como un **método aglomerativo** dado que los conglomerados **se forman por la combinación de los conglomerados existentes.**

Ver la **Figura 14.3** que detalla los pasos del **procedimiento jerárquico**.

**Figura 14.3. Proceso de cluster de conglomerado jerárquico.**

Paso	Proceso de aglomeración		Solución cluster		
	Distancia mínima entre observaciones conjunta (distancias medias no aglomeradas)	Par de observaciones	Pertenencia al conglomerado A-B-C-D-E-F-G	Número de conglomerados 7	Medida de similitud número de conglomerados (dentro del conglomerado)
	Solución inicial				
1	2.414	E-F	A-B-C-D-EF-G	6	2.414
2	3.000	E-G	A-B-C-D-EFG	5	3.192
3	3.000	C-D	A-B-CD-EFG	4	3.144
4	3.000	B-C	A-BCD-EFG	3	3.234
5	3.236	B-E	A-BCDEFG	2	3.896
6	4.162	A-B	ABCDEFG	1	4.420

Fuente: propia

Se observa:

-El **estado inicial** con las **7 observaciones** en conglomerados simples. Posteriormente, se unen los conglomerados en el **proceso aglomerativo** hasta que quede un sólo conglomerado.

-El **paso 1** ubica las 2 observaciones más cercanas (**E y F**) y las combina en un **conglomerado**, reduciendo de 7 a **6 conglomerados**.

-El **paso 2** busca los pares de observaciones más cercanos. En este caso, 3 pares tiene la misma distancia de **3.000 (E-G, C-D, y B-C)**. Al iniciar el análisis con **E-G**. **G es un miembro único de un conglomerado**, pero **E** se combinó en el primer paso con **F**. Así, el conglomerado formado a este nivel tiene **3 miembros: G, E y F**.

-El **paso 3** combina los conglomerados de **miembro único de C y D**

-El **paso 4** combina **B** con el conglomerado de **2 miembros C-D** que se formó en el **paso 3**. Hasta este momento, tenemos **3 conglomerados**: conglomerado 1: (**A**), conglomerado 2 : (**B, C y D**), y conglomerado 3: (**E, F y G**).

- La siguiente distancia más pequeña es **3.236** para 3 pares de observaciones (**E-B, B-G y C- E**). Utilizamos sólo una de estas 3 distancias, sin embargo, en la medida en que cada par de observaciones contenga un miembro de cada uno de los dos conglomerados existentes (**B, C y D frente a E, F y G**).

-El **paso 5** combina los 2 conglomerados de 3 miembros en un único conglomerado de seis miembros.

-El **paso final 6** es combinar la observación **A** con el conglomerado restante (6 observaciones) **en un único conglomerado** a una distancia de **4.162**.

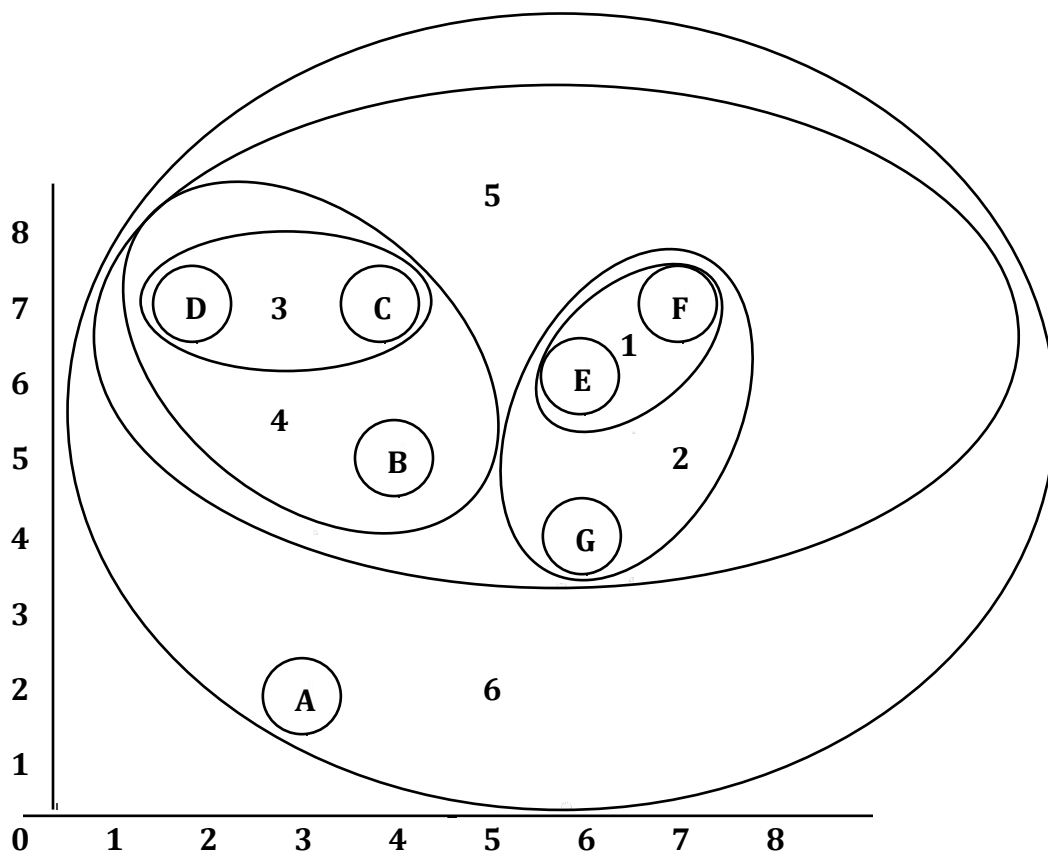
Notará que existen tres distancias iguales o menores a **4.162**, pero que no se utilizan porque están entre los miembros del mismo conglomerado.

El proceso jerárquico de aglomeración puede representarse gráficamente de varias formas. La **Figura 14.4** muestra 2 de tales formas. En primer lugar, dado que el proceso es jerárquico, el proceso de aglomeración puede mostrarse como series de



agrupaciones anidadas (véase **Figura 14.4**). Este proceso, sin embargo, puede representar la proximidad de las observaciones para sólo 2 o 3 variables de aglomeración del gráfico tridimensional o de dispersión.

**Figura 14.4. Agrupación en nido**

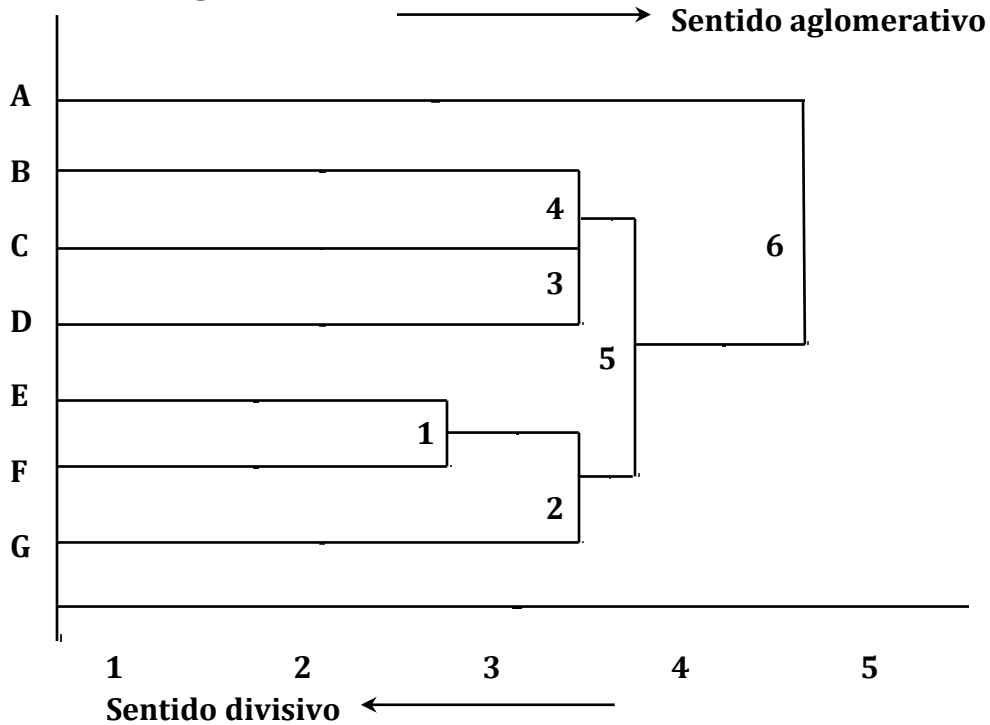


1

Fuente: propia

La **Figura 14.5** muestra una forma gráfica de uso muy habitual: **el dendrograma**, que representa el proceso de aglomeración en un gráfico con forma de **árbol**. El eje horizontal representa el **coeficiente de aglomeración**, en este caso la distancia utilizada en la **unión de aglomerados**. Esta aproximación es particularmente útil en la identificación de **atípicos**, como la observación **A**. También representa el tamaño relativo de los conglomerados que varían, aunque es difícil de manejar cuando aumenta el número de observaciones.

Figura 14.5. Dendograma



Fuente: propia

#### 14.4. Análisis cluster ¿cuántos conglomerados formar?

Un método jerárquico genera un número de soluciones (en este caso van de **1** solución de **1 conglomerado** a 1 solución de **6 conglomerados**), por lo que se produce la pregunta **¿cuál elegir?** Por lo pronto, se sabe que en la medida que nos alejamos de los conglomerados de un único miembro **la homogeneidad disminuye**. Así que, ¿y si nos quedamos con los **7 conglomerados**, que son los más homogéneos posible? El problema es que no hemos definido ninguna estructura con **7 conglomerados**. Así que Usted deberá ver cada solución cluster a partir de la descripción de su estructura compensada con la **homogeneidad de los conglomerados**. En nuestro ejemplo, se usa una medida muy simple de homogeneidad: **las distancias medias** de todas las observaciones dentro de los conglomerados. En la solución inicial con **7 conglomerados**, la medida de similitud conjunto es **0 (ninguna observación está emparejada con otra)**. Así, tenemos:

-En el **paso 1**, para la solución de **6 conglomerados**, la similitud conjunta es la distancia entre las dos observaciones (**2.414**) unidas.

-El **paso 2** forma un conglomerado de **3 miembros (E, F y G)**, de tal forma que la **similitud total** es la **media de las distancias** entre E y F (**2.414**), E y G (**3.000**). y F y G (**4.162**), para una media de **3.192**.

-En el **paso 3**, se forma un nuevo conglomerado de **2 miembros** con una distancia de **3.000**, que provoca que la **media conjunta caiga ligeramente hasta 3.144**. Podemos proceder a formar nuevos conglomerados de esta forma hasta tomar una solución de conglomerado único (**paso 6**), en el que la media de todas las distancias de la matriz

de distancias es **4.420**. Ahora bien, ¿cómo usar esta medida conjunta de similitud para seleccionar una solución cluster? Recuerde que se está intentando conseguir la estructura más simple posible que representa **agrupaciones homogéneas**. Si se controla la medida de similitud conjunta en la medida que disminuye el número de conglomerados, **grandes aumentos en la medida conjunta indican que dos conglomerados no eran tan similares**. Para nuestro caso, aumenta la medida conjunta cuando juntamos dos observaciones en primer lugar (**paso 1**) y a posteriormente, lo hacemos de nuevo cuando construimos el primer conglomerado de **3 miembros (paso 2)**.

-En los pasos (**3 y 4**), no cambia sustancialmente la medida conjunta. Lo anterior indica, que estamos formando más conglomerados con la misma homogeneidad de forma práctica, de los conglomerados existentes.

-**En el paso 5**, que combina los **2 conglomerados de 3 miembros**, se observa un gran aumento. Esto indica que al unir estos **2 conglomerados se obtiene un único conglomerado marcadamente menos homogéneo**. Consideramos la solución cluster del **paso 4** mucho mejor que la del **paso 5**.

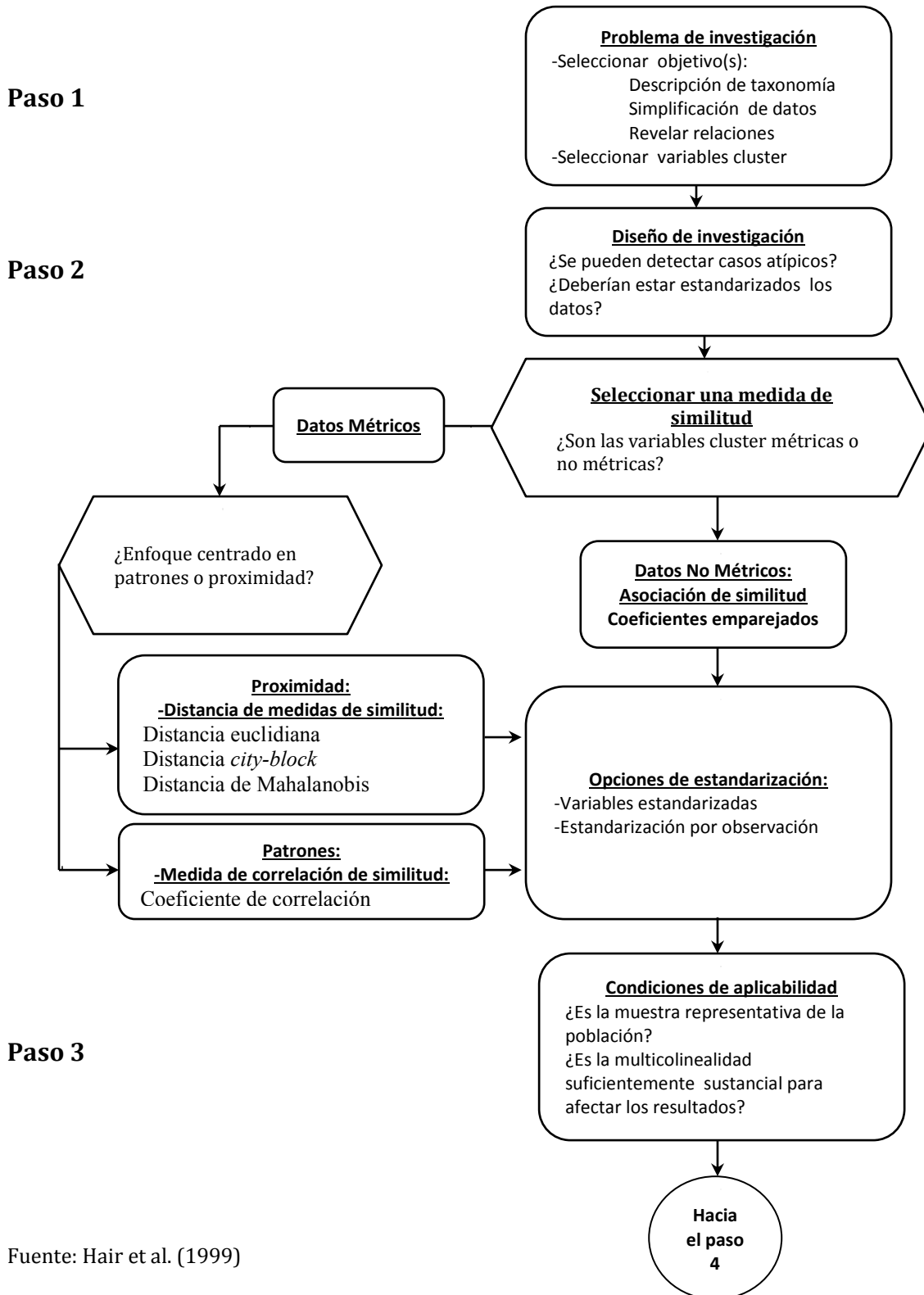
-**En el paso 6** la medida conjunta aumenta de nuevo ligeramente indicando que, incluso aunque la última observación permanezca separada hasta el último paso, **cuando se une cambia la homogeneidad del conglomerado**. Sin embargo, dado el perfil bastante aislado de la observación **A** comparada con el resto, puede ser mejor designar como miembro del **grupo de entropía, aquellas observaciones que son atípicos e independientes de los conglomerados existentes**. Así, al revisar el rango de las alternativas cluster, **la solución de 3 conglomerados del paso 4 se considera como la más apropiada** para una solución cluster definitiva, con **2 conglomerados de igual tamaño y una única observación atípica**.

Como se habrá notado, en la selección de la **solución cluster definitiva**, ésta se deja al libre juicio del investigador, por lo que es considerado por muchos como un **proceso muy subjetivo**. Aún y cuando se han desarrollado métodos más sofisticados para ayudar en la evaluación de las soluciones cluster, **sigue recayendo en el investigador la decisión final del número de conglomerados aceptados en la solución definitiva**. El **análisis cluster** es más simple en un caso como este **bivariante** porque **los datos están en 2 dimensiones**. En las ciencias de la administración de la mercadotecnia, por ejemplo, se miden más de **2 variables** con cada sujeto, y la situación se torna más compleja con muchas más observaciones por evaluar, por lo que es importante emplear procedimientos más sofisticados para tratar con el aumento de la complejidad de las aplicaciones.

Así como el resto de las técnicas multivariantes discutidas anteriormente, la técnica cluster se puede analizar bajo el modelo de los **6 pasos** de Hair et al. (1999) (Vea la **Figura 14.6** y la **Figura 14.7**). Este método inicia con los objetivos de investigación, pudiendo ser tanto confirmatorios como exploratorios, el diseño de un **análisis cluster** que intervenga en la partición del conjunto de datos para formar los conglomerados, la interpretación de los conglomerados y la validación de los resultados. Así, el proceso de interpretación implica entender las características de cada conglomerado y desarrollar un nombre o etiqueta que defina apropiadamente su naturaleza. El proceso final, comprende la evaluación de la validación de la solución

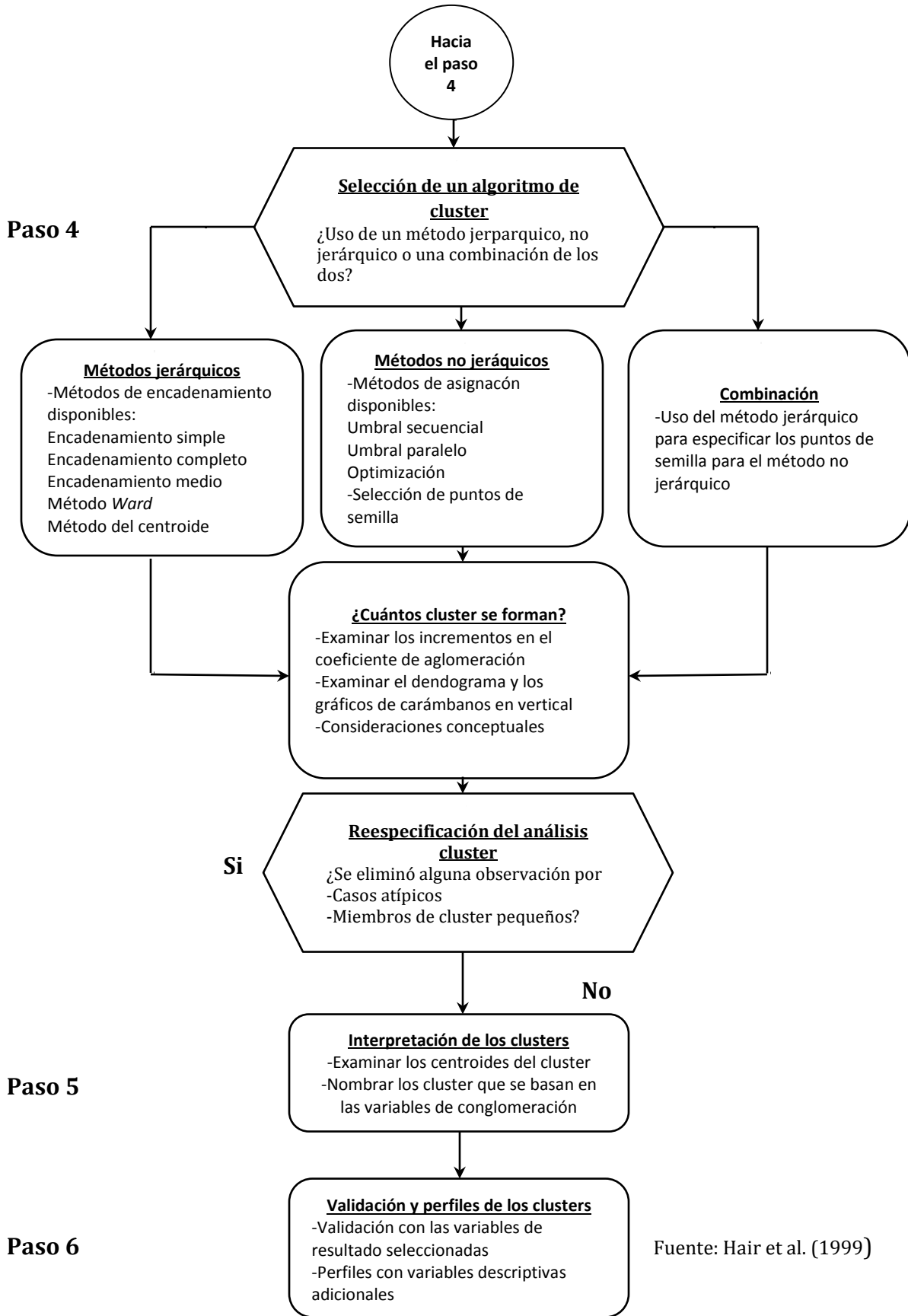
cluster (es decir, determinación de su estabilidad y generalidad), junto con la descripción de las características de cada conglomerado que lo hacen similar y diferente de los otros.

**Figura 14.6. Diagrama de decisión de los pasos 1-3 del análisis de cluster**



Fuente: Hair et al. (1999)

Figura 14.7. Diagrama de decisión de los pasos 4-6 del análisis de cluster



## 14.5 Análisis cluster. Paso 1: Objetivos

Obtener de un conjunto de sujetos dos o más grupos basándose en su similitud para un conjunto de características especificadas (**valor teórico del análisis cluster**), es el objetivo principal de la técnica, en la que, **al formar grupos homogéneos**, Usted puede conseguir:

1. **Creación y descripción de una taxonomía.** Como se explicó al inicio el uso más frecuente de la técnica ha sido basada en propósitos de exploración así como para la formulación de una clasificación de sujetos realizada de forma empírica (taxonomía), debido a su capacidad para la **partición**. También, la técnica es capaz de generar **hipótesis relacionadas con la estructura de los sujetos**. Cabe destacar, que aunque es visto principalmente como una **técnica de exploración**, el **análisis cluster** también se utiliza para **efectos confirmatorios**. Si una estructura propuesta se ha definido para un conjunto de sujetos, es posible aplicar el **análisis cluster**, y comparar su resultado con una tipología propuesta (**clasificación basada en la teoría**).
2. **Reducción de los datos.** Como parte de la obtención de una **taxonomía**, la técnica también **produce una perspectiva simplificada de las observaciones**. Así, las observaciones pueden agruparse para análisis ulteriores, con una estructura definida. Al compararse por ejemplo con el **análisis factorial**, observaremos que este intenta proporcionar “**dimensiones**” o estructuras de variables, mientras **el análisis cluster desarrolla la misma tarea para las observaciones**. Por tanto, todas las observaciones, pueden ser consideradas como miembros de un conglomerado y perfiladas por sus características generales, en vez de considerar todas las observaciones como únicas.
3. **Identificación de relaciones.** Con la estructura subyacente de los datos representada en dichos conglomerados y ya definidos estos, Usted ya tiene un medio para revelar relaciones entre las observaciones que muy probablemente no eran posibles con las observaciones individuales. Mientras se utilizan análisis tales como el **discriminante** para identificar relaciones de forma empírica, o los grupos están sujetos a métodos más bien cualitativos, la estructura simplificada del **análisis cluster** la mayoría de las veces representa **relaciones o similitudes y diferencias no reveladas previamente**.

Los objetivos del **análisis cluster** no deben separarse de la selección de variables que se utilizan para caracterizar los sujetos a agrupar. Ya sea si el objetivo es **exploratorio** o **confirmatorio**, Usted ha restringido efectivamente los resultados posibles por las variables elegidas para su uso. Los conglomerados derivados reflejan la estructura implícita de los datos sólo como definida por las variables. En el **valor teórico del análisis cluster**, la selección de las variables a incluir debe hacerse con relación a **consideraciones teóricas, conceptuales y prácticas**. Cualquier aplicación de la técnica debe descansar en cierta lógica en función de la cual se seleccionan las variables. Tanto si dicha lógica se basa en una teoría explícita, investigación pasada o suposición, Usted debe darse cuenta de la importancia de incluir sólo aquellas variables que:

- a. **Caracterizan los sujetos que se están agrupando, y**
- b. **Se refieren específicamente los objetivos del análisis cluster.**

**Esta técnica, no tiene un medio para diferenciar las variables relevantes de las irrelevantes.** Sólo obtiene los grupos de sujetos más consistentes, aunque diferenciados, para todas las variables. La inclusión de una variable irrelevante **incrementa la posibilidad de que se creen atípicos** sobre estas variables, que pueden tener un efecto importante sobre los resultados. Por tanto, uno **nunca debe incluir variables indiscriminadamente sino que en su lugar, elija las variables utilizando el objetivo de investigación como criterio de selección.** El análisis cluster puede verse drásticamente afectado por la inclusión de una o dos variables inapropiadas o escasamente diferenciadas [Milligan, 1980]. Este procedimiento permite a las técnicas cluster **maximizar los conglomerados** definidos basándose sólo en aquellas variables que exhiban diferencias para todos los sujetos.

#### **14.6. Análisis cluster. Paso 2: Diseño**

Ya definidos los objetivos y seleccionadas sus variables, el investigador debe resolver:

- 1. ¿Qué hacer si hay datos atípicos?**
- 2. ¿La similitud de los sujetos, cómo debería medirse? , y**
- 3. ¿Deben estandarizarse los datos?**

Antes de empezar el proceso de partición. Varios enfoques se pueden utilizar para responder a estas preguntas. A pesar de ello aún no existe la respuesta definitiva ya que ninguno de ellos ha sido suficientemente evaluado en cualquiera de las cuestiones planteadas y, es un hecho desafortunado, que muchas de las aproximaciones arrojan resultados diferentes incluso para el mismo conjunto de datos. Por tanto, el **análisis cluster junto con el análisis factorial, es más un arte que una ciencia** (Hair, et al. 1999). Esta técnica **no puede evaluar todas las particiones posibles** porque, incluso, para un problema relativamente pequeño de partición de **25 sujetos en 5 conglomerados no solapados, existen  $2.4 \times 10^{15}$  particiones posibles** [Anderberg, 1973]. Sólo a partir de las decisiones que realice Usted con base en su experiencia, la técnica le apoyará en identificar a una de las posibles alternativas como “correcta”, lo que hace esta etapa tener quizá un impacto superior al del resto de las técnicas multivariantes. A continuación, se abordará el qué hacer para responder a las 3 interrogantes planteadas

**Particularmente, es importante resolver la cuestión de ¿qué hacer si hay datos atípicos?** La técnica es muy sensible a la inclusión de variables, relevantes o no, dada su naturaleza de buscar una estructura, lo que incluye a los datos atípicos (sujetos que son muy diferentes del resto). Estos pueden provenir de:

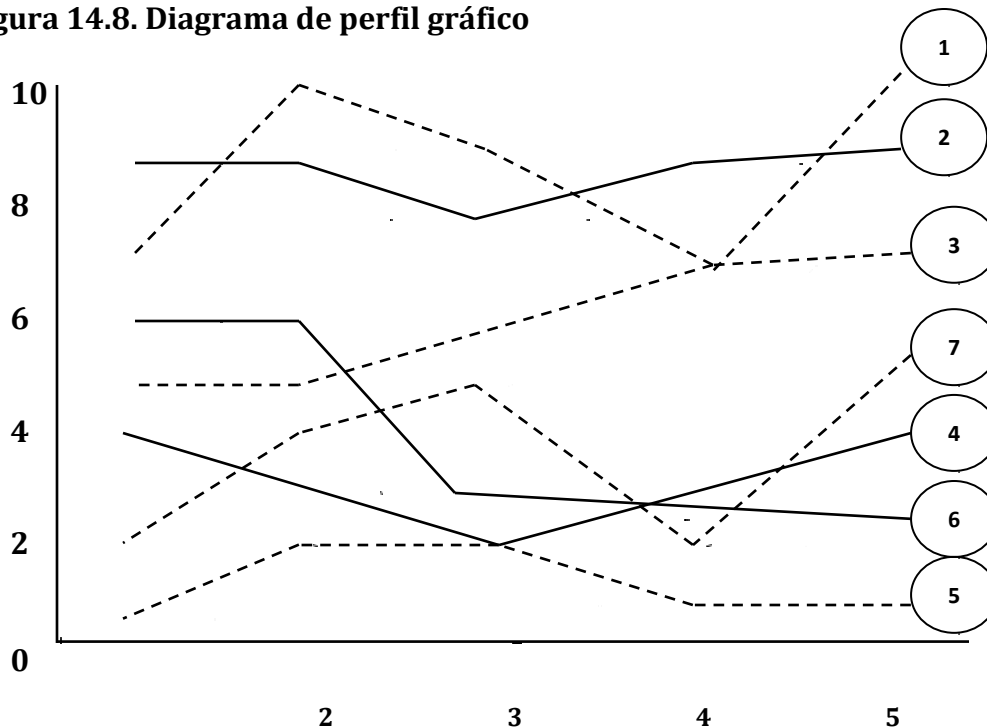
1. Observaciones verdaderamente “*desviadas*” que de la población en general , NO son representativas, o
2. Una reducida muestra del grupo(s) de la población que produce una mala representación del grupo(s) de la muestra.

Los atípicos generan distorsión de la verdadera estructura y hacen que la deducción de los conglomerados no sea representativa de la verdadera estructura poblacional. Esto hace que siempre requiera de una detección preliminar de los



atípicos. Una forma sencilla de lograrlo es realizar un **diagrama de perfil gráfico** (ver **Figura 14.8**).

**Figura 14.8. Diagrama de perfil gráfico**



Fuente: propia

El **diagrama de perfil gráfico** tiene los valores de las variables a lo largo del eje vertical y las variables en el eje horizontal. Cada punto representa el valor de la correspondiente variable, y los puntos están conectados para facilitar la interpretación visual. Son presentados en el gráfico, los perfiles de todos los sujetos, una línea para cada uno. **Los atípicos son aquellos sujetos con perfiles muy diferentes**, la mayoría caracterizados por valores extremos sobre una o más variables. Como se ve, el procedimiento se hace incómodo con un gran número de sujetos (observaciones) o variables. Para las observaciones que se muestran en la **Figura 14.8**, **no hay un atípico obvio que tenga valores extremadamente bajos o altos**. Pero los atípicos pueden definirse también a partir de perfiles únicos que los distinguen del resto de las observaciones. Para estos casos, se pueden aplicar los procedimientos de identificación de atípicos discutidos en el **Capítulo 3** del Tomo I. **También, es posible surjan en el cálculo de similitud**. Cualquiera que sea el medio utilizado, los atípicos puede evaluarse a efectos de su representatividad respecto de la población **y eliminarlos del análisis si se consideran no representativos**. Usted deberá tener precaución en la eliminación de observaciones de la muestra porque **tal eliminación puede distorsionar la estructura efectiva de los datos**. Ahora algunos conceptos que nos emitirán entender mejor la técnica.

**-La similitud**. El concepto de similitud es fundamental para el análisis cluster. La similitud entre sujetos **es una medida de correspondencia**, o parecido, entre sujetos que van a ser agrupados. En el **análisis factorial**, por ejemplo, se crea una **matriz de**

**correlación entre las variables** que fueron utilizadas para agrupar las variables en factores. **Un proceso comparable se produce en el análisis cluster.** En ésta, las características que definen **la similitud** se determinan inicialmente. Posteriormente, las características se combinan en una medida de similitud calculada para todos los pares de sujetos, al igual que lo aplicados en **las correlaciones entre variables en el análisis factorial.** Así, cualquier sujeto es comprado a través de una medida de similitud. El procedimiento del **análisis cluster** procede entonces a agrupar sujetos similares dentro de los conglomerados. La similitud entre sujetos puede medirse de varias formas, aunque **3 métodos dominan:**

- a. **Medidas de correlación,**
- b. **Medidas de distancia y,**
- c. **Medidas de asociación.**

Cada uno de los métodos representa una perspectiva particular de similitud, **dependiendo tanto de sus objetivos como del tipo de datos.** Las medidas de distancia como la correlación exigen datos **métricos**, mientras que las **medidas de asociación** son para datos **no métricos.**

Una de las medidas de similitud entre sujetos muy usada es **el coeficiente de correlación** entre un par de sujetos medido sobre varias variables. Así, en lugar de realizar la correlación entre **2** conjuntos de variables, invertimos la matriz de las **X variables** de los sujetos de tal forma que las columnas representen los sujetos y las filas representan las variables. Por tanto, el **coeficiente de correlación** entre las 2 columnas de números es la correlación (**o similitud**) entre los perfiles de los 2 sujetos. **Correlaciones altas indican similitud y correlaciones bajas indican falta de ella.** Este procedimiento se sigue en la aplicación del **análisis de factor de tipo Q.**

**a. Las medidas de correlación.** Representan la similitud mediante la correspondencia de patrones entre las características (**X variables**). Vea la **Figura 14.8.** Nota: una medida de correlación de similitudes no observa las magnitudes **sino los patrones de los valores.** Ver **Figura 14.9.**

**Figura 14.9. Ejemplo de cálculo de medidas de similitud de distancia y correlación**

Datos originales	Variables				
Caso	X	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>	X <sub>5</sub>
1	7	10	9	7	10
2	9	9	8	9	9
3	5	5	6	7	7
4	6	6	3	3	4
5	1	2	2	1	2
6	4	3	2	3	3
7	2	4	5	2	5

Medida de similitud: correlación	Caso							
	Caso	1	2	3	4	5	6	7
1	<b>1.00</b>							
2	<b>-0.157</b>	1.00						
3	0.000	0.000	1.00					
4	<b>0.097</b>	0.526	<b>-0.834</b>	1.00				
5	<b>0.973</b>	-0.418	0.000	-0.070	1.00			
6	<b>-0.476</b>	0.799	<b>-0.364</b>	0.7	-0.655	1.00		
7	<b>0.991</b>	-0.526	0.175	-0.249	<b>0.973</b>	-0.7	<b>1.00</b>	

Medida de similitud: distancia euclidiana	Caso							
	Caso	1	2	3	4	5	6	7
1	nc							
2	3.42	nc						
3	6.96	6.73	nc					
4	10.34	10.30	6.00	nc				
5	15.88	16.29	10.20	7.17	nc			
6	13.21	13.00	7.38	3.97	3.97	nc		
7	11.37	12.36	6.42	5.20	4.99	4.46	nc	

nc.-No calculada

Fuente: propia

En la **Figura 14.9** que contiene las correlaciones entre **7 observaciones**, se aprecian **2 grupos distintos**:

-Los casos **1, 5 y 7** con patrones similares y por tanto a intercorrelaciones positivas elevadas.

-Los casos **2, 4 y 6** con correlaciones positivas elevadas entre ellos pero con correlaciones bajas o nulas con las otras observaciones. El caso **3** tiene **correlaciones bajas o negativas con el resto de los casos**, por lo que quizá forme un grupo en sí mismo.

De lo anterior, **las correlaciones representan patrones para todas las variables más que las magnitudes**. Es de destacar, que las **medidas de correlación**, sin embargo, **se utilizan rara vez** porque el interés del análisis cluster, **en la mayoría de las aplicaciones no está en los patrones de los valores, sino en las magnitudes de los sujetos**.

**b.Las medidas de la distancia**. Aún y cuando las medidas de correlación son usadas en varias técnicas multivariantes y poseen por tanto, un atractivo intuitivo, **no son las medidas de similitud más utilizadas en el análisis cluster**. Estas corresponden a las **medidas de similitud de distancia**, que representan la similitud como la proximidad de las observaciones respecto a otras variables del valor teórico del análisis cluster. Estas **medidas de distancia** son más bien, **medidas de diferencia**, donde **valores elevados indican una menor similitud**. **La distancia se convierte en medida de similitud utilizando una relación inversa**. Un ejemplo simple de esta

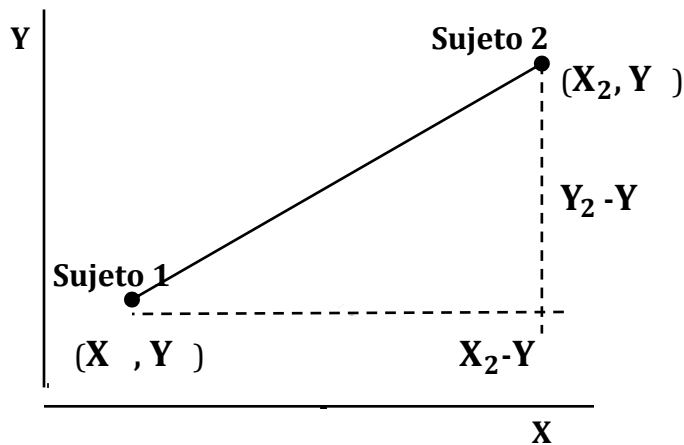
relación, se muestra en la **Figura 14.9** en la que los conglomerados de observaciones se definen de acuerdo a la proximidad entre observaciones y cuando se representa gráficamente las puntuaciones de las observaciones sobre **2** variables.

**Medidas de distancia vs. medidas de correlación.** Tomando de referencia la **Figura 14.8**. Las **medidas de distancia** se centran en la **magnitud de los valores y representan casos similares que están juntos**, pero que tiene pautas muy distintas para todas las variables. La **Figura 14.9** contiene también las **medidas de distancia de similitud** para los **7** casos, apreciándose aglomeraciones muy diferentes de los encontrados con **medidas de correlación**. Con las distancias más reducidas representando similitud, observamos que los **casos 1 y 2 forman un grupo**, y los **casos 4, 5, 6 y 7** otro grupo. Estos grupos representan aquellos con **valores más altos frente a los más bajos**. Un **tercer grupo**, sólo en el **caso 3**, **difiere de los otros en tener valores que tanto bajos como altos**. Aunque los **2** conglomerados usan las **medidas de distancia** tienen diferentes componentes que de los que utilizan las **correlaciones**; siendo el **caso 3 es único** en cada **medida de similitud**. La elección de una **medida de correlación** en lugar de la **medida de distancia (más usada)** requiere de interpretaciones muy diferentes de los resultados por el investigador, considerando:

- Los conglomerados basados en **medidas de correlación pueden no tener valores similares en lugar de tener patrones similares**.
- Los conglomerados basados en la **distancia tienen valores más parecidos para el conjunto de variables, pero los patrones pueden ser bastante diferentes**.

**Medidas de distancia y sus tipos.** Existen varias medidas de distancia, siendo la **más usada, la euclidiana**, en la que un ejemplo de cómo se calcula es la **Figura 14.10**.

**Figura 14.10.** Distancia euclideana entre dos sujetos medidos en las variables (Xn, Yn)



$$\text{Distancia} = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

Fuente: propia

Así, suponga que partimos de **2 puntos con 2 dimensiones**, con coordenadas respectivas de  $(X_1, Y_1)$  y  $(X_2, Y_2)$ . La **distancia euclidiana** entre los puntos es la longitud de la hipotenusa de un **triángulo rectángulo**, calculada por la fórmula bajo la figura, siendo el concepto generalizable para **más de 2 variables**. La distancia euclidiana se utiliza para calcular medidas específicas tales como:

- La **simple distancia Euclídea** (calculada como se explicó) y,
- La **distancia euclidiana cuadrada, o absoluta**, como la suma de las diferencias al cuadrado sin tomar la raíz cuadrada, la que tiene la **ventaja de no tener que tomar la raíz cuadrada lo que acelera notablemente los cálculos**, y es la medida de distancia recomendada para los **métodos de análisis cluster del centroide y Ward**. Existen opciones que no se basan en la distancia Euclídea:
- Una de estas medidas alternativas, muy usada, consiste en **reemplazar la diferencia de los cuadrados por la suma de las diferencias absolutas de las variables, denominada función de la distancia absoluta**. Su enfoque puede resultar apropiado bajo ciertas circunstancias, aunque **puede provocar varios problemas**. Uno es el supuesto de que **las variables no están correlacionadas con el resto; si lo están, los conglomerados no son válidos** [Shephard, 1966].
- En la mayoría de los programas del análisis cluster se encuentran otras medidas que emplean variaciones de las diferencias absolutas o las potencias aplicadas a las diferencias (**que no sean sólo la diferencia de los cuadrados**)

**-Valores de los datos no estandarizados y su impacto.** Las **medidas de distancia** siempre se enfrentan al problema de que el **uso de datos no estandarizados implica inconsistencias** entre las alternativas cluster al cambiar la escala de las variables. Como ejemplo, suponga que **3 sujetos, A, B y C** se miden sobre **2 variables, probabilidad de conversión en el portal X (definido en porcentajes) y cantidad de tiempo gastado viendo el portal X (en minutos o segundos)**. Los valores de cada observación se muestran en la **Figura 14.11**.

**Figura 14.11. Variaciones en las medidas de distancia basadas en distintas escalas de medida**

Datos originales			
Sujeto	Probabilidad de conversión portal X (%)	Segundos	Minutos
A	61	180	3
B	66	210	3.5
C	64	240	4

Medidas de distancia basadas en la probabilidad de conversión y tiempo en MINUTOS de acceso a los productos del portal X						
Par de Sujetos	Distancia euclidiana simple		Distancia euclidiana absoluta o cuadrada		Distancia City-block	
	Valor	Rango	Valor	Rango	Valor	Rango
A-B	5.026	3	25.26	3	5.6	3
A-C	3.163	2	10.01	2	4.1	2
B-C	2.063	1	4.26	1	2.6	1

Medidas de distancia basadas en la probabilidad de conversión y tiempo en SEGUNDOS de acceso a los productos del portal X						
Par de Sujetos	Distancia euclidiana simple		Distancia euclidiana absoluta o cuadrada		Distancia City-block	
	Valor	Rango	Valor	Rango	Valor	Rango
A-B	30.42	2	926	2	36	3
A-C	60.08	3	3610	3	64	2
B-C	30.07	1	905	1	33	1

Medidas de distancia basadas en valores estandarizados de probabilidad de conversión y tiempo en MINUTOS o SEGUNDOS de acceso a los productos del portal X								
Par de Sujetos	Valores estandarizados		Distancia euclidiana simple		Distancia euclidiana absoluta o cuadrada		Distancia City-block	
	Probabilidad de conversión	Minutos/Segundos de acceso a portal X	Valor	Rango	Valor	Rango	Valor	Rango
A-B	-1.17	1.1	2.23	2	4.96	2	3.00	2
A-C	0.94	0.1	2.34	3	5.43	3	3.30	3
B-C	0.14	1.1	1.29	1	1.64	1	1.80	1

Fuente: propia

Así, con los datos de la **Figura 14.11**, es posible calcular las **medidas de distancia**. Por ejemplo, se calculan **3 medidas de distancia para cada par de sujetos: distancia euclidiana simple, la distancia euclidiana absoluta o al cuadrado y la distancia absoluta**. Primero, se debe calcular los valores de las distancias de tomando de base **la probabilidad de conversión de un portal X vs. el tiempo de acceso al portal X**. En la **Figura 14.11** se muestran dichas distancias. **Los valores más reducidos indican mayor proximidad y similitud**, así como su lugar en orden. Los sujetos más parecidos (**menor distancia**) son **B-C**, seguidos de **A y C**, con **A y B siendo los menos parecidos (menos próximos)**. El orden se conserva para las **3 medidas de distancia**, aunque la **similitud relativa o dispersión** entre los sujetos **es más pronunciada en la medida de distancia euclidiana al cuadrado**. Con esto, **el ordenamiento de las similitudes puede cambiar profundamente con sólo un cambio en la escala de una de las variables**. Si medimos el tiempo de acceso al portal en **segundos** en lugar de minutos, entonces el ordenamiento cambia (vea **Figura 14.11**). Los sujetos **B-C** siguen permaneciendo como los más parecidos, pero ahora el par **A-B** es más parecido y casi idéntico a la similitud de **B-C**. Pero cuando tenemos **minutos** de tiempo de visión, el par **A-B** es el menos parecido por un margen substancial. Lo que ha ocurrido es que **la escala de tiempo acceso al portal X como variable ha dominado los cálculos**, generando en la probabilidad de compra que sea **menos significativa en los cálculos**. Al contrario, al medir el **tiempo de acceso al portal X en minutos**, es entonces la **probabilidad de conversión en el portal X** la que **domina en los cálculos**. Usted notará el relevante impacto que la escala de las variables puede tener sobre la solución final. Así, **se recomienda emplear la estandarización de las variables de aglomeración, siempre que sea conceptualmente posible, para evitar casos que originen discrepancias**.

Otra medida de distancia euclídeana muy utilizada que **incorpora directamente un procedimiento de estandarización es la distancia de Mahalanobis [ $D^2$ ]**. Cuyo enfoque no sólo realiza el proceso de estandarización de los datos a escala en términos de las **desviaciones estándar** sino que también evalúa la **varianza-covarianza** unidas dentro del grupo, ajustando las **intercorrelaciones** entre las variables. Conjuntos de variables altamente **intercorrelacionados** del análisis cluster ponderan implícitamente un conjunto de variables en la aglomeración. De lo anterior, se puede decir que el procedimiento de **distancia generalizada de Mahalanobis calcula una medida de distancia entre sujetos comparable al R2 del análisis de regresión** (Hair et al., 1999). A pesar de que en muchas situaciones puede ser apropiado el uso de la **distancia de Mahalanobis [ $D^2$ ]**, **no todos los programas la incluyen como medida de similitud**. En tales casos, Usted deberá seleccionar la **distancia euclídeana al cuadrado**. Al seleccionar una medida de distancia particular, Usted deberá recordar las siguientes **advertencias**:

**-Diferentes medidas de distancia o un cambio en la escala de las variables pueden llevar a diferentes soluciones cluster**. Así que es recomendable utilizar varias medidas y comparar los resultados con **pautas teóricas o conocidas**.

**-Cuando las variables están intercorrelacionadas (positiva o negativamente), la medida de distancia de Mahalanobis es probable que sea la más apropiada** dado que se ajusta para las correlaciones y ponderaciones de todas las variables igualmente.

**-Desde luego, si Usted desea ponderar las variables n forma desigual, existen otros procedimientos [Morrison, 1967; Overall, 1964].**

**c. Medidas de asociación de similitud**. Su uso generalizado es el **comparar sujetos** cuyas características se miden sólo en **términos no métricos (medida nominal y ordinal)**. Por ejemplo, suponga el caso en que los encuestados responden **sí o no** a a cierto cuestionario, en las que se pueden apreciar diferentes **medida de asociación: -Una podría ser evaluar el grado de acuerdo o de acercamiento entre cada par de encuestados**. -Otra forma más simple sería **el porcentaje de veces que existió acuerdo** (ambos dicen sí o ambos dicen no) para el mismo conjunto de cuestiones. Se han desarrollado extensiones de este **simple coeficiente de ajuste para acomodar variables nominales de varias categorías o incluso medidas ordinales**. Muchos programas informáticos, sin embargo, dan un apoyo limitado a las medidas de asociación, y Usted estará forzado en muchas ocasiones a calcular inicialmente, las **medidas de similitud** y a continuación **introducir la matriz de similitud en los programas de análisis cluster**. Hay diversas fuentes en las que se pueden encontrar revisiones de varios tipos de medidas de asociación [Everitt, 1980].

**-Tipificación de los datos**. Con la medida de similitud seleccionada, Usted debe preguntarse: **¿deberían tipificarse los datos antes de calcular las similitudes? . Para resolverlo, debe considerar que a mayoría de las distancias medidas son muy sensibles a las diferentes escalas o magnitudes de las variables** (Recuerde el impacto del caso analizado cuando cambiamos de **minutos a segundos** en una de nuestras variables). Por lo general, **variables con una mayor dispersión**

(grandes desviaciones estándar) tienen más impacto en el valor final de similitud. Por ejemplo, suponga que desea agrupar individuos mediante 3 variables: **disposición a compra, educación, ingresos**. Ahora supongamos que medimos la **disposición a compra con una escala de siete puntos** (entre el agrado y el desagrado), la educación en años de estudio y el ingreso en unidades monetarias (usd). Si graficáramos tridimensionalmente la distancia entre los puntos (y sus similitudes) se observaría que estaría basada casi completamente en las diferencias por los **ingresos**. Las posibles diferencias de **disposición a compra** van de 1 a 7, mientras que los **ingresos** tienen un rango de marcadamente mayor. Por tanto, gráficamente no seríamos capaces de ver ninguna diferencia en la dimensión asociada con la **disposición a compra**. Por esta razón, Usted debe tener cuidado con la ponderación implícita de las variables en función de su **dispersión relativa, que sucede con las medidas de distancia**.

**-Estandarización por variables.** Una forma muy utilizada de estandarización es la conversión de cada variable a unas **puntuaciones estándar (conocidas como puntuaciones Z)** que consiste en **restar la media y dividir por la desviación típica de cada variable**. Es una forma generalizada de la **función de distancia normalizada que utiliza una medida de distancia euclídeana para transformación normalizada de los datos originales**. El proceso hace una conversión de cada puntuación de los datos originales a un **valor estandarizado con una media de 0 y desviación estándar de 1**. La transformación, **elimina el sesgo introducido por las diferencias en las mediciones de varios atributos** o variables utilizadas en el análisis. En la **Figura 14.11** se observan los beneficios de la estandarización en la última sección en la que 2 variables (**probabilidad de conversión del portal X y tiempo de acceso al portal X**) se han estandarizado antes de calcular las 3 medidas de distancia, con las siguientes observaciones:

1. Es mucho más fácil comparar entre las variables en la medida en que están en la **misma escala (una media 0 y desviación estándar 1)**. Los valores positivos están por encima de la media y los valores negativos están por debajo; la magnitud representa el número de desviaciones estándar del valor original a partir de la media.
2. **No existe diferencia entre los valores estandarizados cuando sólo cambia la escala**. Por ejemplo, cuando el **tiempo de acceso al portal X en minutos y en segundos se estandariza, los valores son los mismos**. Así, al usar las variables estándar **se eliminan verdaderamente los efectos debidos a las diferencias de escala** no sólo entre las variables, sino también para la misma variable.
3. Sin embargo, Usted no debería aplicar sistemáticamente la estandarización sin considerar sus consecuencias. **No hay razón para aceptar absolutamente una solución cluster utilizando variables estandarizadas frente a variables no estandarizadas**. Si existe alguna relación "*natural*" reflejada en la escala de las variables, entonces **la estandarización puede no ser apropiada**.
4. La decisión de estandarizar tiene impactos tanto conceptuales como empíricos por lo que es recomendable hacerlo siempre después de una cuidadosa consideración.



**Estandarización por la observación.** Ya analizamos la **estandarización sólo para variables**. Pero **¿qué hay de la estandarización de los encuestados o casos?** ¿Es importante y factible hacerlo? Un ejemplo lo aclarará. Suponga que hizo una encuesta donde se recogió de los encuestados un número de calificaciones sobre una escala de **10 puntos** sobre la importancia de varios atributos que determinan sus decisiones de compra en un producto/servicio. Al aplicar la técnica de análisis cluster y obtener conglomerados, se debe considerar también la probabilidad de obtener grupos con 3 respuestas muy diferenciadoras, como: gente que dice que algo era importante, gente que dice que algo tenía poca importancia, y quizá algunos conglomerados intermedios. Lo que estamos viendo **son los efectos de tipo de respuesta en los conglomerados** que son las pautas sistemáticas de respuesta a un conjunto de preguntas, **tales como los que siempre dicen sí** (respuesta favorable para todas las cuestiones) **o los que siempre dicen no** (respuesta desfavorable para todas las cuestiones). **Si queremos identificar los grupos de acuerdo a su estilo de respuestas, entonces la estandarización no es apropiada.** Sin embargo, en la mayoría de los casos lo que se desea es **la importancia relativa de una variable a otra**. Es decir, **¿es más importante el atributo 1 que el resto de los atributos, y pueden encontrarse en los conglomerados de encuestados pautas de importancia similar?** En este caso, **la estandarización por encuestado estandarizaría cada cuestión no por la media de la muestra sino por la puntuación media del encuestado.** Esta tipificación entre sujetos o tipificación centrada por filas puede ser bastante efectiva al eliminar efectos de respuesta y es especialmente adecuada para muchas formas de **datos de actitud** [Schaninger, et al. 1986]. Esto es similar a una **medida de correlación en la señalización de la pauta para todas las variables**, pero la proximidad de todos los casos todavía determina el valor de similitud. (Hair et al. 1999)

#### **14.7. Análisis cluster. Paso 3: Condiciones de aplicabilidad**

El **análisis cluster**, al igual que el **análisis multidimensional** (vea Capítulo 13), **NO es una técnica de inferencia estadística** en la que de una muestra se analizan los parámetros en la medida en que éstos puedan ser representativos de una población. Al contrario, **el análisis cluster es una metodología de cuantificación de las características estructurales de un conjunto de observaciones muy objetiva.** Como tal, tiene robustas propiedades matemáticas **pero no fundamentos estadísticos. Las exigencias de normalidad, linealidad y homocedasticidad que eran tan importantes en otras técnicas realmente tienen muy poco peso en el análisis cluster.** Así, Usted deberá centrarse, en otros asuntos críticos, **a saber 2:**

- 1. La muestra y su representatividad. Rara vez, tendrá un censo de la población** con el fin de aplicarle ésta técnica. Normalmente, una muestra de casos es la que se obtiene, y de la que se derivan los conglomerados con la **esperanza de que representen la estructura de la población.** Así, Usted deberá confiar en que la muestra obtenida sea **verdaderamente representativa de la población.** Como se explicó, por ejemplo, **los atípicos tienen la posibilidad de ser generados por proceder de una muestra escasa de grupos divergentes que, cuando se descartan, introducen sesgos en la estimación de la estructura.**

Usted deberá darse cuenta de que el **análisis cluster** sólo es tan **eficiente** como lo sea la **muestra en su representatividad**. Por lo tanto, todos los esfuerzos deben dirigirse para **asegurar que la muestra es representativa** a fin de que los resultados sean generalizables para la población a estudiar.

2. **La multicolinealidad y su impacto.** Es tema fundamental de otras técnicas multivariantes a causa de la dificultad de discernir el “*verdadero*” impacto de las variables multicolineales. Sin embargo, en el **análisis cluster** es diferente su efecto porque **aquellas variables que son multicolineales están implícitamente con más fuerza ponderadas**. Por ejemplo, asuma que esta en proceso de agrupar a los encuestados sobre **10** variables, con todas afirmaciones de actitud hacia un producto/servicio. Al analizar la multicolinealidad, observamos que existen realmente **2 conjuntos de variables:**
- El primero constituido de **8 afirmaciones** y
  - El segundo consistente en las **2 restantes**.

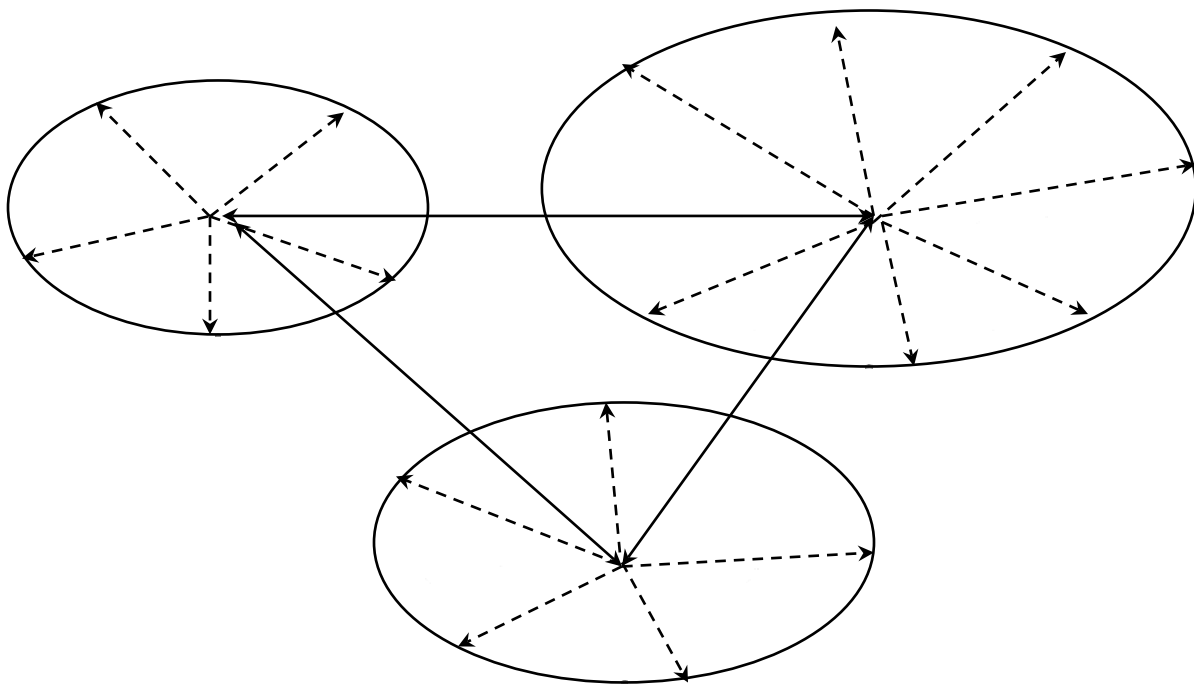
Si nuestro intento es en realidad agrupar a los encuestados a partir de las dimensiones del producto/servicio (representadas por los **2 grupos de variables**), entonces **será un error utilizar las 10 variables iniciales**. Partiendo que **cada variable se pondera igualmente en el análisis cluster**, la **dimensión 1** tendrá como mucho **4 veces** más posibilidad (**8 ítems vs. 2** para afectar a la medida de similitud, y lo mismo ocurrirá con la **dimensión 2**. **De esta forma, como proceso de ponderación no aparente para el observador pero que sin embargo afecta al análisis es como actúa la multicolinealidad**. Por tanto, es importante que analice las variables usadas en el **análisis cluster** en la búsqueda de una **multicolinealidad substancial** y, de encontrarse, **o bien reducir las variables al mismo número en cada conjunto o bien utilizar una de las distancias medidas, como la distancia de Mahalanobis ( $D^2$ )**, que compensa esta **correlación**. Cabe anotar que se ha generado **debate sobre el uso de las puntuaciones de factor en el análisis cluster**. Ciertas investigaciones han mostrado que las variables que verdaderamente discriminan entre los grupos subyacentes **no están bien representados** en la mayoría de las soluciones de factor, de tal forma que, cuando se utilizan **puntuaciones de factor**, es **bastante posible que se obtenga una mala representación de la verdadera estructura de los datos** [Rohlf, 1970]. Por tanto, Usted deberá tratar tanto con la multicolinealidad como con la discriminabilidad de las variables para llegar a la mejor representación de la estructura.

#### **14.8. Análisis cluster. Paso 4: Ejecución y ajuste del conglomerado**

Obtenidas la **selección de variables** y el cálculo de la **matriz de similitud**, inicia el proceso de partición (ver **Figura 14.7**). Con esto, Usted deberá buscar y seleccionar inicialmente, **el algoritmo de aglomeración** utilizado en la formación de conglomerados y, enseguida **tomar la decisión del número de conglomerados** que se van a formar. Las implicaciones son sustanciales en ambas decisiones, no sólo sobre los resultados a generar, sino también sobre la interpretación basada en ellos, como se expone enseguida.

**Algoritmo de conglomerados.** Se parte del cuestionamiento clave, **¿para colocar sujetos similares en grupos o conglomerados, qué procedimiento debería utilizarse?** Esto es, ¿qué conjunto de reglas es más apropiado o mejor dicho, qué algoritmo de obtención de conglomerados usar? Dado que existen cientos de programas informáticos que utilizan diferentes algoritmos, en permanente desarrollo, por lo no es una pregunta sencilla de resolver. Sin embargo, existen criterios esenciales en todos los algoritmos, como el intentar maximizar las diferencias entre los conglomerados relativa a la variación dentro de los conglomerados, tal y como se muestra en la **Figura 14.12.**

**Figura 14.12. Diagrama de conglomerados que muestra la variación dentro y entre conglomerados.**



Fuente: propia

Se hace notar que la razón entre **la variación entre conglomerados y la media de la variación dentro de los conglomerados es comparable (pero no idéntica) a la razón  $F$  del análisis de la varianza.**

Los algoritmos que generan conglomerados, más utilizados se clasifican en **2** grupos:

**1. Jerárquicos.**

Consisten en construir una estructura **arborescente**, con 2 tipos de procedimientos de obtención de conglomerados jerárquicos básicos:

- a. De aglomeración.** En este procedimiento cada sujeto u observación **empieza dentro de su propio conglomerado**. En etapas posteriores, los dos conglomerados más cercanos (**o individuos**) **se combinan en un nuevo conglomerado agregado, reduciendo así el número de conglomerados paso a paso**. En algunos casos, **un tercer individuo se une** a los dos primeros en un conglomerado. En otros, **2 grupos** de individuos formados en un paso

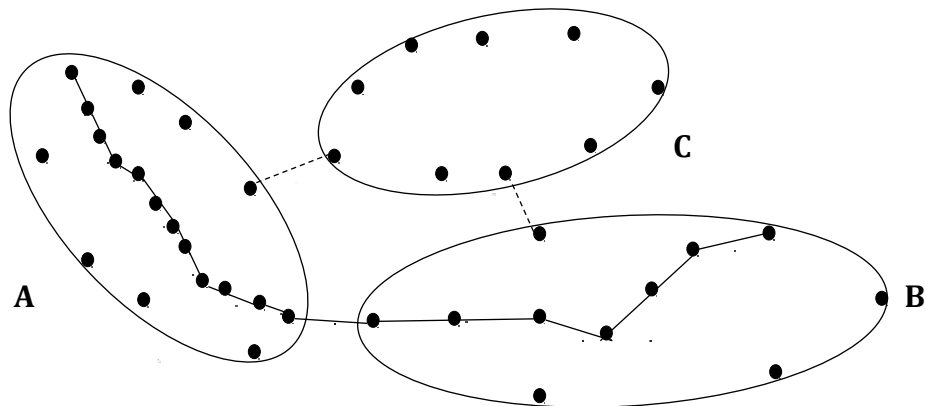
anterior pueden unirse en un nuevo conglomerado. **Finalmente, que todos los individuos se agrupan en un único conglomerado;** por esta razón, este procedimiento de aglomeración se les denomina **como métodos de construcción**. Una característica importante de los **procedimientos jerárquicos** es que los resultados obtenidos en **un paso previo siempre necesitan encajarse dentro de los resultados del siguiente paso, creando algo parecido a un árbol**. Por ejemplo, una solución **de 6 conglomerados** se obtiene uniendo **dos de los conglomerados encontrados** en el paso de **7 conglomerados**, como los conglomerados se forman sólo por unión de los **conglomerados existentes, es posible rastrear incluso hasta su origen de simple observación, cualquier miembro de un conglomerado**. Se muestra este proceso se muestra en la **Figura 14.5**. La representación se denomina **dendrograma o gráfico en forma de árbol**. Otro método gráfico muy habitual es el **diagrama de carámbanos en vertical**. Cuando el proceso de obtención de conglomerados procede en dirección opuesta al **método de aglomeración**, se denomina **método divisivo**.

- b. **Divisivos**. En este método se inicia con un **gran conglomerado que contiene todas las observaciones (sujetos)**. Sucesivamente, **las observaciones que son más diferentes se dividen y se construyen conglomerados más pequeños hasta que cada observación es un conglomerado en sí mismo**. En la Figura 14.13 los **métodos aglomerativos** van de izquierda a derecha y los **métodos divisivos** van de derecha a izquierda. Dado que los programas informáticos más habituales utilizan los **métodos aglomerativos, y los métodos divisivos actúan como métodos aglomerativos al revés**, se recomienda centrarse en los **métodos aglomerativos**.

Los algoritmos más habituales, utilizados para desarrollar conglomerados difieren en cómo se calcula la distancia entre los conglomerados y **son 5**:

1. **Método de encadenamiento simple**. Este se basa en la **distancia mínima**. Busca y encuentra los **2 sujetos separados por la distancia más corta** y los ubica en el primer conglomerado. Posteriormente, se encuentra la distancia más corta, y o bien un **tercer sujeto** se une a los **2 primeros para conformar un conglomerado o se forma un nuevo conglomerado de 2 miembros**. Sucesivamente el proceso continúa **hasta que todos los objetos se encuentran en un conglomerado**. También se le conoce como **procedimiento de enfoque del vecino más cercano**. La distancia entre 2 conglomerados cualquiera es la **distancia más corta** desde cualquier punto en un conglomerado a cualquier punto en el otro, de tal forma que **2 conglomerados se fusionan** en cualquier nivel por el vínculo más corto o más fuerte entre ellos. Sin embargo, se generan problemas **cuando los conglomerados están mal definidos y formar largas y sinuosas cadenas**, donde eventualmente todos los individuos pueden situarse en una cadena. Los individuos que se encuentran en los límites opuestos de una cadena pueden ser muy diferentes. Un ejemplo de este acuerdo se muestra en la **Figura 14.13**.

**Figura 14.13. Encadenamiento simple que une los conglomerados diferentes A y B**

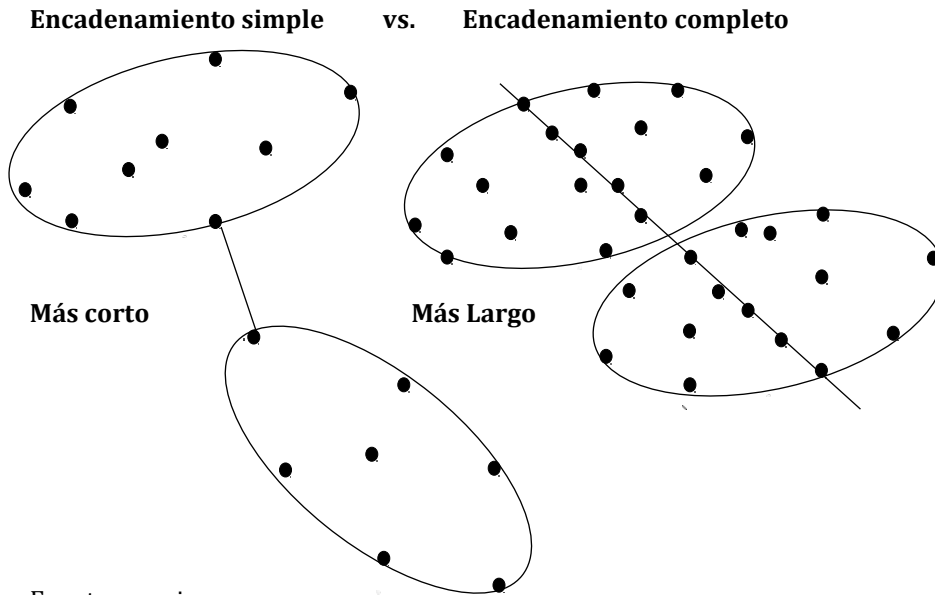


Fuente: propia

Como se observa, están por unirse **3 conglomerados (A,B y C)**. El algoritmo de **encadenamiento simple**, centrándose sólo en los puntos más cercanos de cada conglomerado, **uniría los conglomerados A y B** debido a su distancia más reducida en los **extremos de los conglomerados**. Uniendo los conglomerados **A y B** se crea un conglomerado que rodea al **conglomerado C**. Pero si buscamos la homogeneidad dentro del conglomerado, **sería mucho mejor juntar el conglomerado C con A o B**. Esta es la **principal desventaja del algoritmo de encadenamiento simple**.

2. **Método de encadenamiento completo.** Este procedimiento es similar al del encadenamiento simple **excepto en que el criterio de aglomeración se basa en la distancia máxima**. Por esta razón, se le denomina como **aproximación del vecino más lejano o método del diámetro**. La distancia máxima entre individuos de cada conglomerado representa la esfera más reducida (**diámetro mínimo**) que puede incluir todos los objetos en ambos conglomerados. También se le **denomina encadenamiento completo** porque todos los objetos de un conglomerado se vinculan con el resto a alguna **distancia máxima o por la mínima similitud**. Podemos decir que la similitud dentro del grupo es igual al **diámetro del grupo**. Esta técnica **elimina el problema identificado para el encadenamiento simple**. La **Figura 14.14** muestra cómo las distancias más cortas (**encadenamiento simple**) o las más largas (**encadenamiento completo**) representan la similitud entre grupos.

**Figura 14.14. Comparación de las medidas de distancia del encadenamiento simple y completo.**



Fuente: propia

Ambas medidas reflejan sólo un aspecto de los datos. El uso de la distancia más corta refleja sólo un único par de objetos (**los más cercanos**) y el encadenamiento completo también refleja un único par, esta vez **los 2 más lejanos**. Por tanto, es útil ver las medidas como reflejo de la similitud del par de **sujetos más parecido o del par menos parecido**.

3. **Método de encadenamiento medio.** Este método comienza igual que los métodos de **encadenamiento simple o completo**, sin embargo, fija **el criterio de aglomeración como la distancia media de todos los individuos de un conglomerado con todos los individuos de otro**. Tales técnicas **NO** dependen de los valores extremos, como se hace en el **encadenamiento simple o completo** y la partición se basa en todos los miembros de los conglomerados en lugar de un par único de miembros extremos. Este método, tiende a combinar los conglomerados con variaciones reducidas dentro del conglomerado. También **tiende a estar sesgado** hacia la producción de conglomerados con aproximadamente **la misma varianza**.
4. **Método de Ward.** En este otro método, **la distancia entre 2 conglomerados es la suma de los cuadrados entre 2 conglomerados sumados para todas las variables**. En el desarrollo de cada paso del método, **se minimiza la suma de los cuadrados dentro del conglomerado para todas las particiones** (el conjunto completo de conglomerados disjuntos o separados) obtenida mediante **la combinación de dos conglomerados en un paso previo**. Este procedimiento **tiende a combinar los conglomerados con un número reducido de observaciones**. También está **sesgado hacia la**

**producción de conglomerados con aproximadamente el mismo número de observaciones**

5. **Método del centroide. Centroide.** Aquí, **la distancia entre los 2 conglomerados es la distancia (normalmente euclidiana simple o cuadrada) entre sus centroides.** Los centroides de los grupos representan los valores medios de las observaciones de las variables en el **valor teórico del conglomerado.** Cada vez que se agrupa a los individuos, en este método, un nuevo centroide se calcula. En los grupos, sus centroides cambian a medida que se fusionan conglomerados. Es decir, cada vez que un nuevo individuo o grupo de individuos se añade al conglomerado existente se produce un cambio en un centroide de un grupo. Estos métodos son más populares entre los biólogos, psicólogos, comunicólogos, etc. pero pueden producir resultados desordenados y a menudo confusos que se produce generalmente a causa de los propios cambios, o sea, se generan casos donde la distancia entre los centroides de un par puede ser menor que la distancia entre los centroides de otro par fusionado en una combinación anterior. **La ventaja de este método es que se ve menos afectada por los atípicos que otros métodos jerárquicos.**

## 2. No jerárquicos.

Como contraste estos procedimientos **No implican los procesos de construcción de árboles.** En su lugar, **asignan los sujetos a conglomerados una vez que el número de conglomerados a formar está especificado.** Por tanto, la solución de **6 conglomerados** no es sólo una combinación de **2 conglomerados** desde una solución de **7 conglomerados**, sino que se basa simple y solamente en la búsqueda de la mejor solución de **6 conglomerados.** Para visualizar cómo funciona, suponga que el **paso 1** es seleccionar una **semilla de conglomerado** como **centro de conglomerado inicial**, y todos los **sujetos (individuos)** dentro de una **distancia umbral** previamente especificado **se incluyen dentro del conglomerado resultante.** Entonces, **se selecciona otra semilla de conglomerado** y la asignación continúa sucesivamente hasta que todos los objetos están asignados. Entonces, Los **sujetos** pueden asignarse si están cercanos a otro conglomerado **que no sea el original.** Diferentes aproximaciones existen para **seleccionar las semillas de conglomerado** y asignar sujetos a comentar en la siguiente sección. Los procedimientos de aglomeración no jerarquizados se denominan frecuentemente como aglomeración de **K-medias**, y normalmente utilizan uno de las siguientes **3 aproximaciones** para asignar las observaciones individuales de uno de los conglomerados [Green, 1978]

a. **Umbral secuencial.** Este método empieza al **seleccionar una primera semilla de conglomerado** incluyendo todos los sujetos que caen dentro de una distancia previamente especificada. Al realizarse en su totalidad, se selecciona **una segunda semilla de conglomerado** y se incluyen todos los objetos dentro de la distancia previamente determinada. Enseguida, se selecciona una **tercera semilla**, y el proceso continúa sucesivamente como se ha descrito. Cuando un sujeto se incluye en un conglomerado con una semilla, no se considera a efectos de ulteriores semillas.

- b. **Umbral paralelo.** Como contraste, este método **selecciona varias semillas de conglomerado simultáneamente al principio** y asigna objetos dentro de la distancia umbral hasta la semilla más cercana. A medida que el proceso avanza, **se puede ajustar las distancias umbral para incluir más o menos objetos en los conglomerados.** También, en algunas variantes de este método, los objetos permanecen fuera de los conglomerados si están fuera de la distancia previamente especificada desde cualquiera de las semillas de conglomerado.
- c. **Optimización. O procedimiento de optimización,** es similar a los otros 2 procedimientos mencionados **excepto en que permite la reubicación de los sujetos.** Si en el curso de la asignación de los objetos, un objeto se acercara más a otro conglomerado que no es el que tiene asignado en este momento, entonces **un procedimiento de optimización hace el cambio del objeto al conglomerado más parecido (cercano).**

**Puntos de semilla ¿cómo seleccionar?** En **umbral secuencial** una vez que Usted especifique **el número máximo de conglomerados permitidos**, el procedimiento inicia con la selección de semillas de conglomerados, utilizadas como **supuestos iniciales de las medias de los conglomerados.** La **primera semilla** es la primera observación del conjunto de datos sin valores perdidos. La **segunda semilla** es la siguiente observación completa (sin datos perdidos) que **se separa de la primera semilla** mediante una **distancia mínima especificada.** Una **distancia mínima de cero** es la opción por defecto. Seleccionadas todas las semilla, el programa asigna al conglomerado cada observación con la semilla más próxima. Usted puede especificar que los conglomerados de semillas se revisen (**actualicen**) a partir del **cálculo de medias de los conglomerados de semillas cada vez que se una observación se asigna.** Como contraste, los **métodos del umbral paralelo** (por ejemplo, **QUICK- CLUSTER** en **SPSS**) establecen los puntos de semilla como aportaciones por el usuario o seleccionado de las observaciones en forma aleatoria. Todos los métodos de formación de conglomerados no jerárquicos se enfrentan al principal problema de **cómo seleccionar las semillas de conglomerado.**

**Advertencia:** Con una **opción de umbral secuencial**, por ejemplo, **los resultados del conglomerado inicial y probablemente del final dependerán del orden de las observaciones en el conjunto de datos** y arrastrar el orden de los datos es como afectar a los resultados. Puede reducir este problema, la especificación de las semillas de conglomerado iniciales, como se hace en el procedimiento de **umbral secuencial.** Sin embargo, **incluso la selección aleatoria de las semillas de conglomerado producirá diferentes resultados** para cada conjunto de puntos de semilla aleatorios. Por tanto, Usted deberá estar consciente del impacto del proceso de selección de las semillas de conglomerados en los resultados finales.

**Métodos jerárquicos vs. no jerárquicos. ¿Cuándo usarlos?** No existe una respuesta definitiva debido a dos razones:

1. Existe la tendencia a sugerir un método u otro al problema a investigar en ese momento.



2. Lo que se tiene de un contexto aprendido con la continua aplicación tiende a sugerir un método u otro como el más aconsejable para ese contexto.

**Pros y contras de los métodos jerárquicos.** Inicialmente, las **técnicas jerárquicas de formación de conglomerados eran las más populares**, siendo el **método de Ward y el encadenamiento medio probablemente como los mejores disponibles** [Milligan, 1980]. Su principal ventaja es el de ser **más rápidos y llevar menos tiempo de cálculo**. Pero, esto ya no aplica hoy en día dado el alto poder de cómputo personal de los equipos actuales. Sin embargo, dado que se generan **combinaciones iniciales indeseables que sin cuidado, pueden persistir a lo largo del análisis y llevar a resultados artificiales**, los métodos jerárquicos pueden dar una idea equivocada. **Los atípicos** por cierto, son de interés específico dado su impacto sustancial sobre los **métodos jerárquicos**, particularmente con el método del **encadenamiento completo**. Para reducir esta posibilidad, Usted debe realizar el **análisis cluster de los datos repetidas veces, eliminando los atípicos o las observaciones problemáticas**.

**Advertencia: La destrucción de casos, sin embargo, incluso aquellos que no sean atípicos, puede muchas veces distorsionar la solución.** Por tanto, Usted debe tener un cuidado extremo en la destrucción de las observaciones por la razón que sea.

**¿Problemas con el tamaño muestral?** También, cabe decir que los métodos jerárquicos **No son susceptibles de analizar muestras muy grandes**, dado que los requisitos de almacenamiento de datos aumentan enormemente. Por ejemplo, una muestra de **400 casos** exige el almacenamiento de aproximadamente **80.000 similitudes**, que se incrementa a **125.000** para una muestra de **500**. Usted puede considerar una **muestra aleatoria de las observaciones originales** para reducir el tamaño de la muestra **pero debe cuestionarse ahora la representatividad** de la muestra tomada de la muestra original.

**Aparición de los métodos no jerárquicos.** Estos métodos **han ganado una creciente aceptación** y se aplican cada vez más. Sin embargo, su uso depende de la **capacidad de Usted para seleccionar los puntos de semilla de acuerdo a bases prácticas, objetivas y teóricas**. Los métodos no jerárquicos, en estos casos, tienen varias ventajas sobre las técnicas jerárquicas:

- Los resultados son menos susceptibles a los datos atípicos
- A la medida de distancia utilizada y
- A la inclusión de variables irrelevantes o inapropiadas.

Sin embargo, estos beneficios se obtienen, sólo con el uso de **puntos de semillas no aleatorios (es decir, especificados)**; por tanto, el uso de técnicas no jerárquicas con puntos de semilla aleatorios es **notablemente inferior a las técnicas jerárquicas**. Comenzando incluso con una solución no aleatoria **no se garantiza una formación de conglomerados óptima de observaciones**. De hecho, en muchos casos, el investigador obtendrá una solución final diferente para cada conjunto de puntos de semilla especificados. ¿Cómo va a seleccionar el investigador la respuesta "**correcta**"? **Sólo mediante el análisis y la validación** logrará seleccionar lo que se considera la

“*mejor*” representación de la estructura, teniendo presentes las muchas alternativas que pueden considerarse aceptables.

**Combinando ambos métodos.** Una propuesta de aproximación, es utilizar ambos métodos (**jerárquico y no jerárquico**) para obtener los beneficios de cada uno [Milligan, 1980], de la siguiente forma:

1. Una **técnica jerárquica** establece el número de conglomerados, los perfiles de los centros de conglomerados y la identificación de cualquier atípico obvio. Eliminar los **atípicos**
2. **Un método no jerárquico** entonces entra a analizar las observaciones restantes para ser agrupadas con los centros de conglomerados desde los resultados jerárquicos como los **puntos de semillas iniciales**.

De esta forma, las ventajas de ambos métodos se complementan entre sí con la para “*ajustar*” los resultados permitiendo el cambio de pertenencia a un conglomerado.

**Grupos que deben formarse.** Este puede ser un asunto polémico y desconcertante (**también conocida como regla de parada**). No existe un procedimiento objetivo o estándar desafortunadamente en este momento. Como se explicó anteriormente no se utiliza un criterio estadístico interno para la inferencia, tal como los **test de significación estadística de otros métodos multivariantes**, por lo que los investigadores han desarrollado varios criterios y líneas a seguir para aproximarse al problema. Una de las principales conclusiones es que se han creado procedimientos **ad hoc** que deben ser usados por el investigador, lo que muchas veces implica procedimientos francamente complejos [Aldenderfer y Blashfield (1984); Milligan y Cooper (1985)]. Una clase de reglas de parada que es relativamente **simple examina alguna medida de similitud o distancia entre los conglomerados a cada paso sucesivo, donde la solución cluster se define cuando la medida de similitud excede a un valor especificado o cuando los valores sucesivos entre los pasos dan un salto súbito**. Un ejemplo sencillo ya se utilizó en el caso al principio del capítulo, que buscaba grandes aumentos en la media de la distancia dentro del conglomerado. Al producirse un gran aumento, el investigador selecciona la solución cluster previa en la lógica de que su combinación provocó la sustancial reducción en su similitud. Se ha mostrado que **esta regla de parada ofrece decisiones francamente precisas en los estudios empíricos** [Milligan y Cooper (1985)].

Una segunda clase de reglas de parada **intentan aplicar alguna forma de regla estadística o adaptar un test estadístico, tal como las correlaciones “point-biserial/tau” o el ratio de verosimilitud**. Aunque alguno de estos criterios (como el criterio cúbico de elaboración de conglomerados (CCC) que se encuentra en el SAS) ha mostrado notable éxito, **muchos parecen demasiado complejos** para la mejora que ofrecen sobre muestras simples. Aunque se han propuesto un cierto número de procedimientos específicos, no se ha encontrado ninguno que sea mejor en todas las situaciones.

También, Usted deberá complementar **el juicio estrictamente empírico con cualquier conceptualización de las relaciones teóricas** que pueda sugerir un número natural de conglomerados. Se sugiere iniciar este proceso especificando algún criterio basándose en **consideraciones prácticas**, por ejemplo decir, “*los resultados*

*serán más manejables y más fáciles de comunicar si se logran obtener de tres a seis conglomerados*”, y a continuación resolver para este número de conglomerados y **seleccionar la mejor alternativa después de evaluar todas ellas**. En el análisis final, sin embargo, probablemente sea mejor calcular varias soluciones cluster diferentes (2, 3, 4) y mostrar el proceso de decisión entre las soluciones alternativas utilizando **criterios a priori, juicios prácticos, o mucho mejor fundamentos teóricos** (se sugiere no incluir el *“sentido común”*) Introducir los aspectos conceptuales a la discusión de la solución cluster permitirá que se vean mejoradas los descartes del resto de las alternativas que se generen.

#### **14.9. Análisis cluster. Paso 5: Interpretación de los conglomerados**

Este paso, implica el análisis de cada conglomerado basado en los **términos del valor teórico del conglomerado** o asignar una etiqueta precisa que describa la naturaleza de los conglomerados. Para clarificar este proceso, vamos a referir un ejemplo de *Pantallas TV* y *Smart TVs*. Supongamos que se desarrolla una **escala de actitud** que consista en **afirmaciones** relacionadas con el consumo de *Smart TVs*, tales como *“las Pantallas TV son económicas”, “las Smart TVs son caras”, “las Pantallas TV me entregan el contenido deseado”, “Las Smart TVs tienen acceso a portales especiales de contenidos”, “Las Pantallas TV no requieren acceso a internet”, “las Smart TVs tienen potencial de demandar más servicios de internet”, etc.* Además, suponga que se obtienen también los datos demográficos y de consumo de *Pantallas TV* y *Smart TVs*. Cuando se comienza la interpretación, una medida utilizada frecuentemente es el **centroide del conglomerado**, del que sucede:

1. Si el procedimiento de aglomeración se realizó sobre los datos tal y como se obtuvieron, lo anterior sería una descripción de procedimiento lógico.
2. Pero, **si los datos se estandarizaron o si el análisis cluster se realizó utilizando el análisis factorial (de componentes principales)**, el investigador tendría que retroceder a las puntuaciones dadas por los encuestados de las variables originales y calcular los perfiles medios utilizando estos datos.

Siguiendo con el ejemplo de las *Pantallas TV* y *Smart TVs*, se analiza en esta etapa los **perfiles de las puntuaciones medias sobre las afirmaciones de las actitudes para cada grupo y asignamos una etiqueta descriptiva para cada conglomerado**. Muchas veces el **análisis discriminante se aplica para generar puntuaciones de perfiles**, aunque se debe recordar que las diferencias estadísticamente significativas **no indicarían una solución “óptima” porque se esperan diferencias estadísticas, da dos los objetivos del análisis cluster**. Así, el análisis de los perfiles permite una descripción más abundante de cada conglomerado. Quizás observemos que **2 de los conglomerados** pueden tener **actitudes favorables** sobre *Pantallas TV* y *Smart TVs* y el **tercer conglomerado actitudes negativas**. Además, es posible filtrar en particular cada una de las actitudes de los *Pantallas TV* y *Smart TVs* ya que desde este punto de vista del procedimiento analítico, uno evaluar a las actitudes de cada conglomerado y desarrollaría **interpretaciones sustantivas** para facilitar su clasificación. Por ejemplo, un conglomerado puede calificarse como *“consciente de su potencial de acceso a servicios de internet”* mientras que otro puede calificarse como *“basado en precio de adquisición del aparato”*. Sin embargo, se consigue algo más que

una descripción, basados en los perfiles y la interpretación de los conglomerados, tales como:

1. Proporcionan un medio de evaluar la correspondencia de los conglomerados derivados de aquellos propuestos por una teoría a priori o por la experiencia práctica. Así, los **perfiles del análisis** cluster ofrecen un medio directo de evaluación de la correspondencia De usarse de forma **confirmatoria**.
2. **Evaluaciones de significación práctica**, se obtienen de los perfiles de los conglomerados como opción de realización. Usted puede diseñar y demandar que existan diferencias sustanciales en un conjunto de variables de elaboración de conglomerados y que las **soluciones cluster aumenten hasta que surjan tales diferencias**. Al evaluar tanto su significación práctica o correspondencia, Usted compara los conglomerados derivados con una pre-determinada tipología.

#### **14.10. Análisis cluster. Paso 6: Validación y perfil de grupos**

En esta etapa, y en virtud de que juega de forma importante la **naturaleza subjetiva del análisis cluster** sobre la selección de una solución cluster “*óptima*”, Usted deberá tener mucho cuidado en la validación y asegurarse la relevancia práctica de la **solución cluster definitiva**. Aunque no existe un método único para asegurar la relevancia práctica y su validez, se tiene en consideración diferentes procedimientos que ofrecen cierta base a la evaluación realizada por Usted.

**Solución cluster y su validación.** Los intentos de Usted por la validación incluyen asegurar que la solución cluster sea **representativa** de la población general a fin de que sea **generalizable** a otros objetos y con el tiempo, estable. El procedimiento más directo es realizar **análisis cluster** para muestras distintas. Sin embargo, este procedimiento frecuentemente no es práctico debido a las restricciones de costes o tiempo o a la indisponibilidad de los sujetos (los consumidores) para múltiples análisis cluster. En estos casos, un procedimiento común es dividir la muestra en **2 grupos**. Se analiza cada conglomerado por separado y después se comparan los resultados. Otros procedimientos a considerar, son:

1. Una forma modificada de división de la muestra por la cual se emplean los centros de conglomerados obtenidos desde una solución cluster para definir conglomerados a partir de otras observaciones para comparar después los resultados [McIntyre y Blashfield,1980], y
2. Validación cruzada de forma directa [Punj y Stewart, 1983]. En este procedimiento, Usted puede intentar establecer **alguna forma de criterio o validez predictiva**. Para hacerlo, Usted selecciona una variable o variables no utilizadas para formar los conglomerados pero que se sabe que cambian a lo largo de los conglomerados. En nuestro ejemplo, podemos conocer de la investigación pasada las actitudes hacia los **Smart TVs** por grupos de edad. Con ello, podemos contrastar estadísticamente las diferencias de edad entre aquellos conglomerados que sean favorables a las **Smart TVs** y de aquellos que no lo son. Así el caso, deberían tener un fuerte apoyo teórico o práctico en la medida en que se conviertan en el punto de referencia de selección entre las soluciones de conglomerado, las variables utilizadas para **evaluar la validez predictiva**.

**Solución cluster y los perfiles.** Este paso implica, de cada conglomerado, su descripción de las características que expliquen en que medida las dimensiones relevantes pueden diferir, por lo que es recomendable el uso del **análisis discriminante**. Una vez que se han identificado los conglomerados, inicia el proceso. Usted utiliza datos **previamente no incluidos** en el procedimiento **cluster** para **perfiles las características de cada conglomerado**. Estos datos son normalmente perfiles psicográficos, pautas de consumo, características demográficas, etc. Aunque una razón teórica es posible que no exista, para que difieran entre los conglomerados, tal como requerir la evaluación de la validez predictiva, **si tienen que tener al menos importancia práctica**. Utilizando el análisis discriminante, Usted comparará los perfiles de las puntuaciones medias para los conglomerados. La **variable categórica dependiente** se identifica como el **conjunto de conglomerados previamente identificado** y las **variables independientes son las demográficas, psicográficas, etcétera**. Suponiendo la relevancia estadística, Usted podría concluir, por ejemplo, que el conglomerado de “*consciente de su potencial de acceso a servicios de internet*” del ejemplo previo consiste en profesionistas con altos ingresos y educación superior, consumidores de tecnología. Para finalizar, **el análisis de perfil se centra en la descripción no de lo que directamente se determinan los conglomerados sino de las características de los conglomerados una vez que se han identificado**. Además, se debe hacer hincapié en las **características** que **difieren significativamente** entre los conglomerados de aquellas que podrían **predecir la pertenencia a un conglomerado específico**.

#### 14.11. Análisis cluster. Resumen

El análisis cluster lo equipa a Usted desarrollar una de las tareas más inherentemente humanas: **la clasificación** por medio de un método objetivo y empírico, con el objetivo de explorar, simplificación o confirmar, **ésta técnica tiene un amplio rango de aplicaciones, por ser una potente herramienta analítica**. Por tal motivo, Usted tiene la responsabilidad de aplicar sus principios subyacentes de forma adecuada ya que tiene muchos inconvenientes, que incluso, al investigador experimentado le demanda aplicarlo con precaución. Sin embargo, con un uso adecuado, desarrolla el potencial de revelar estructuras dentro de los datos que no se descubren de otra forma, satisfaciendo la necesidad fundamental de los investigadores en todos los campos, como las ciencias de la administración. Las funciones adicionales de la **sintaxis de comandos CLUSTER**. El procedimiento **conglomerado jerárquico** utiliza la **sintaxis de comandos CLUSTER (IBM, 2011d)**. Con el lenguaje de sintaxis de comandos también podrá:

- Utilizar varios métodos de agrupación en un único análisis.
- Leer y analizar una matriz de proximidades.
- Escribir una matriz de distancias para su análisis posterior.
- Especificar cualquier valor para la potencia y la razón en la medida de distancia personalizada (potencia).
- Especificar nombres para variables guardadas.

### 14.12. Análisis cluster jerárquico. Ejemplos

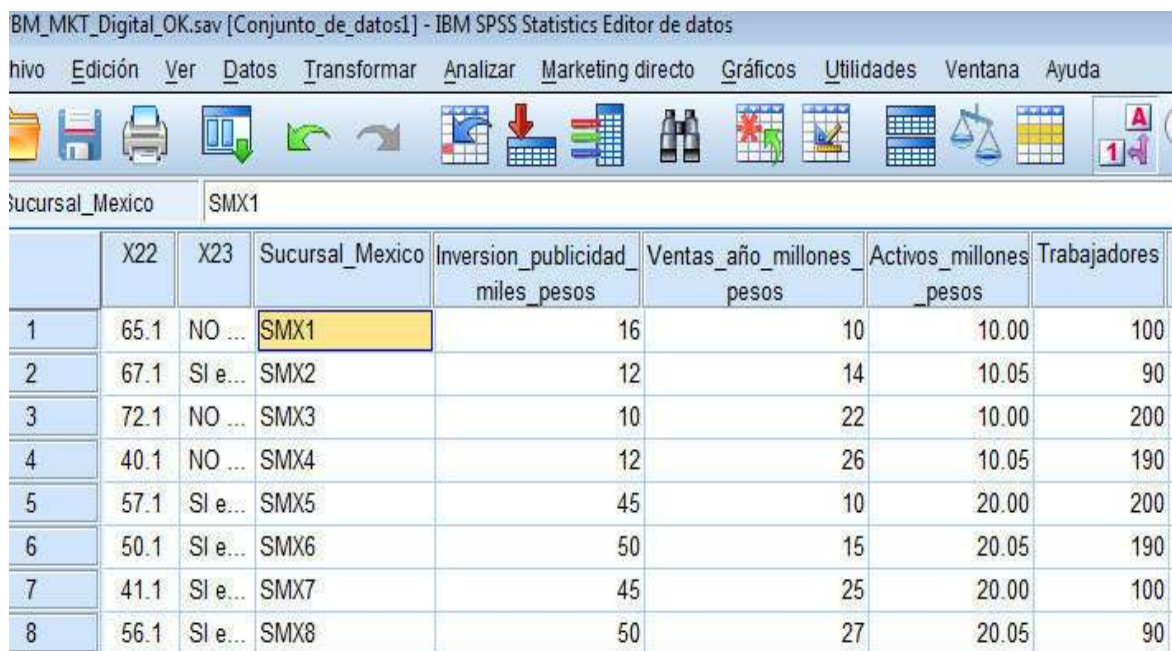
Esta técnica se orienta a identificar **grupos homogéneos en cierto nivel de variables o casos**, con base a características determinadas mediante la propuesta de un algoritmo que inicia con cada variable o caso de un cluster diferente y combina los conglomerados hasta que sólo queda uno. Existe la posibilidad de analizar las variables iniciales o hacer una elección de entre una variedad de transformaciones de estandarización. En este caso, las medidas de similaridad o distancia se generan a partir de la técnica de **Proximidades**. Los estadísticos se deben analizar en cada etapa para ayudar a seleccionar la mejor solución.

#### Paso 1: Objetivos

**Problema 1:** La empresa **MKT Digital** desea saber cómo se encuentran aglomeradas sus 8 sucursales que tiene en México, relacionando la inversión de publicidad con sus ventas. Se sugiere manejar **entre 4 a 5 cluster para análisis máximo**.

Ver **Figura 14.15** y **Figura 14.16**

**Figura 14.15. Visor base de datos BM\_MKT\_Digital.sav**



	X22	X23	Sucursal_Mexico	Inversion_publicidad_miles_pesos	Ventas_año_millones_pesos	Activos_millones_pesos	Trabajadores
1	65.1	NO ...	SMX1	16	10	10.00	100
2	67.1	SI e...	SMX2	12	14	10.05	90
3	72.1	NO ...	SMX3	10	22	10.00	200
4	40.1	NO ...	SMX4	12	26	10.05	190
5	57.1	SI e...	SMX5	45	10	20.00	200
6	50.1	SI e...	SMX6	50	15	20.05	190
7	41.1	SI e...	SMX7	45	25	20.00	100
8	56.1	SI e...	SMX8	50	27	20.05	90

Fuente: SPSS 20 IBM

**Figura 14.16. Visor variables base de datos BM\_MKT\_Digital.sav**

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida
25	Sucursal_Mexico	Cadena	8	0		Ninguna	Ninguna	11	Izquierda	Nominal
26	Inversion_publicidad_miles_pesos	Numérico	2	0		Ninguna	Ninguna	13	Derecha	Escala
27	Ventas_año_millones_pesos	Numérico	2	0		Ninguna	Ninguna	14	Derecha	Escala
28	Activos_millones_pesos	Numérico	3	2		Ninguna	Ninguna	10	Derecha	Escala
29	Trabajadores	Numérico	8	0		Ninguna	Ninguna	8	Derecha	Escala

Fuente: SPSS 20 IBM

## Paso2: Diseño

Se resuelven las preguntas sugeridas, como sigue:

1. **¿Qué hacer si hay datos atípicos?** Se considera que no existen casos atípicos
2. **¿La similitud de los sujetos, cómo debería medirse?** Se propone medir con la técnica de **Vinculación inter-grupos, Intervalo: distancia euclídea al cuadrado.**
3. **¿Deben estandarizarse los datos?** Por lo respondido en los 2 puntos anteriores, no e considera en el caso.

## Paso 3: Condiciones de aplicabilidad

Las exigencias de **normalidad, linealidad y homocedasticidad NO son críticos en el análisis cluster.** Sin embargo, sí lo son:

1. La muestra y su representatividad, se consideran sin problema para el ejemplo ya que parten de un **censo** de las **8 sucursales en México de la empresa MKT Digital.**
2. La multicolinealidad y su impacto, se consideran sin problema para el ejemplo.

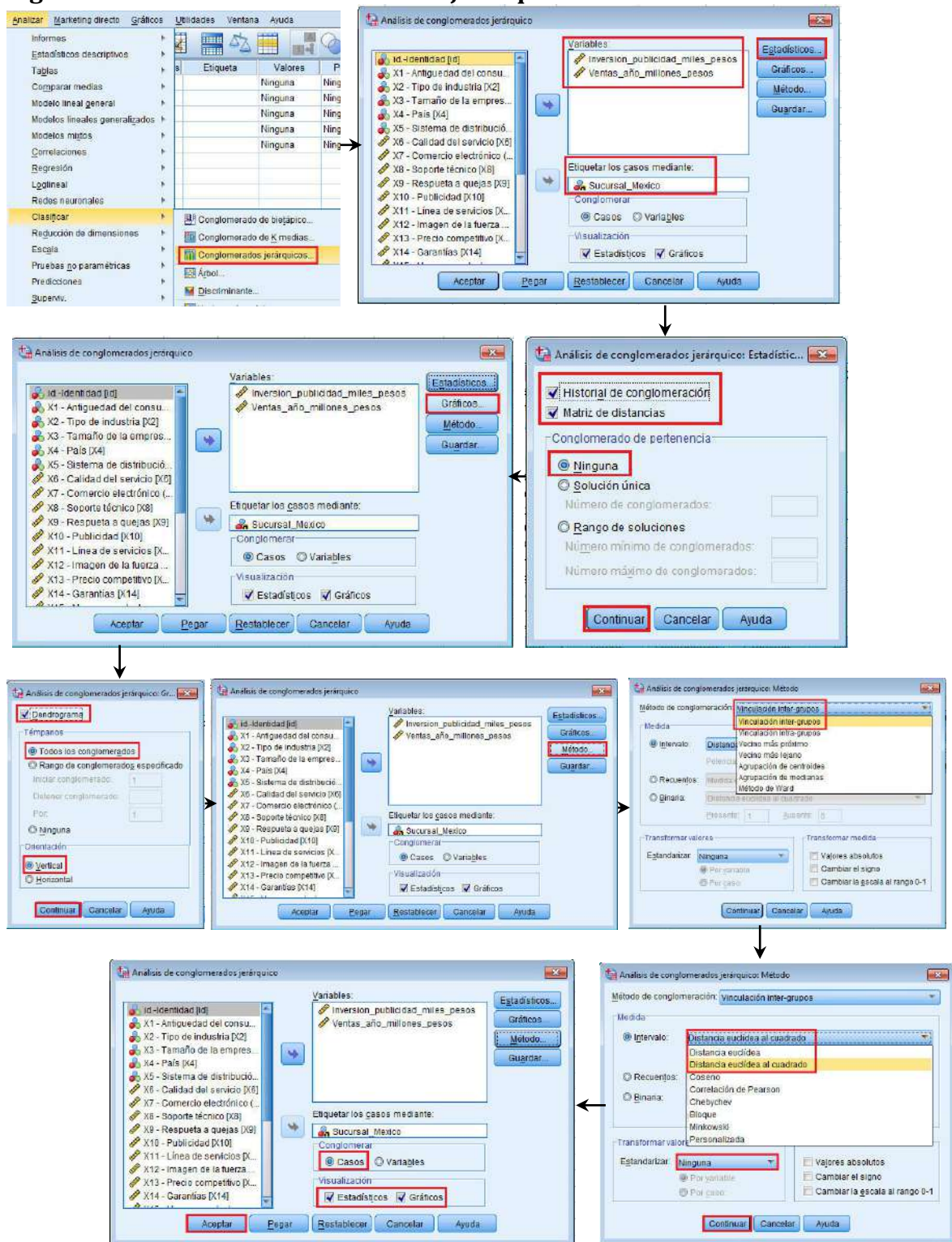
## Paso 4: Estimación y ajuste

Teclar: **Analizar->Clasificar->Conglomerados jerárquicos->Variables (selección métricas, en nuestro caso: Inversión\_publicidad\_miles\_de\_pesos; Ventas\_año\_millones\_de\_pesos); Etiquetar los casos mediante (selección variable categórica, en nuestro caso: Sucursal México)->Estadísticos; seleccionar: Historial de conglomerados y Matriz de distancias; Conglomerados de pertinencia; seleccionar: Ninguna->Continuar->Gráficos; seleccionar: Dendograma; Témpanos: Todos los conglomerados; Orientación: Vertical->Continuar->**Método de conglomeración: Vinculación inter-grupos; Medida; intervalo: Distancia euclídea al cuadrado**->Estandarizar: ninguna->Continuar->Conglomerar: Casos; Visualización: Estadísticos; Gráficos->Aceptar.**

Ver Figura 14.17.



**Figura 14.17 Proceso análisis clúster jerárquico**



Fuente: SPSS 20 IBM



## Paso 5: Interpretación de los conglomerados

SPSS genera la tabla Resumen del procesamiento de casos que indica los que se han incluido en el estudio y los que se hubieran perdido por falta de datos. Ver **Figura 14.18**

**Figura 14.18. Tabla Resumen del procesamiento de los casos**

### Conglomerado

[Conjunto\_de\_datos1] C:\Users\Juan\Desktop\proy libro mc\Bases de datos SPSS

**Resumen del procesamiento de los casos<sup>a</sup>**

Casos					
Válidos		Perdidos		Total	
N	Porcentaje	N	Porcentaje	N	Porcentaje
8	100	0	0	8	100

a. Vinculación promedio (Inter-grupos)

Fuente: SPSS 20 IBM

SPSS enseguida muestra la **tabla Matriz de distancias**, la cual nos da indicios de los grupos más parecidos a medida que sus distancias se hacen más cortas como serían los casos de cruce **SMX2-SMX1= 32.00; SMX3-SMX2= 68.00; SMX4-SMX3=20.00; SMX6-SMX5= 50.00; SMX8-SMX7=29.00**. Ver **Figura 14.19**.

**Figura 14.19. Tabla Matriz de distancias**

**Matriz de distancias**

Caso	distancia euclídea al cuadrado							
	1:SMX1	2:SMX2	3:SMX3	4:SMX4	5:SMX5	6:SMX6	7:SMX7	8:SMX8
1:SMX1	.000	32.000	180.000	272.000	841.000	1181.000	1066.000	1445.000
2:SMX2	32.000	.000	68.000	144.000	1105.000	1445.000	1210.000	1613.000
3:SMX3	180.000	68.000	.000	20.000	1369.000	1649.000	1234.000	1625.000
4:SMX4	272.000	144.000	20.000	.000	1345.000	1565.000	1090.000	1445.000
5:SMX5	841.000	1105.000	1369.000	1345.000	.000	50.000	225.000	314.000
6:SMX6	1181.000	1445.000	1649.000	1565.000	50.000	.000	125.000	144.000
7:SMX7	1066.000	1210.000	1234.000	1090.000	225.000	125.000	.000	29.000
8:SMX8	1445.000	1613.000	1625.000	1445.000	314.000	144.000	29.000	.000

Esta es una matriz de disimilaridades

Fuente: SPSS 20 IBM

Para confirmar lo anterior, **SPSS** presenta la tabla Historial de conglomeración, en la cual, la distancia inicial más corta, representada como **SMX4-SMX3=20.00** reportando de forma consecutiva las siguientes distancias menos cortas y asociándolas en el aglomerado o cluster correspondiente. Observe que cada asociación también indica la

etapa en la que prosigue, por ejemplo, en nuestro caso sigue con la **etapa 5** y así, sucesivamente. Ver **Figura 14.20**.

**Figura 14.20. Tabla Historial de conglomeración**

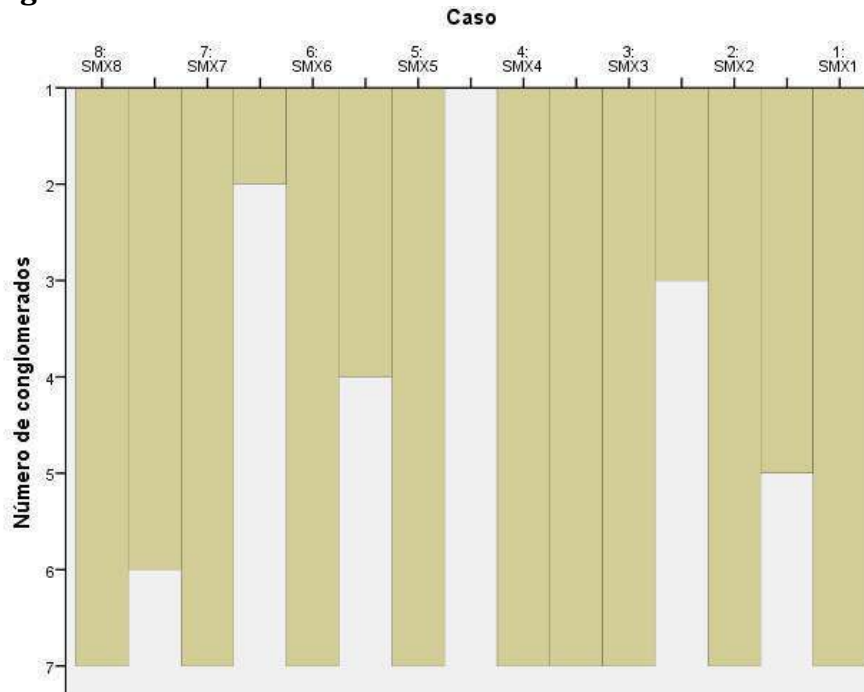
**Vinculación promedio (Inter-grupos)**

Historial de conglomeración						
Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerad o 1	Conglomerad o 2		Conglomerad o 1	Conglomerad o 2	
1	3	4	20.000	0	0	5
2	7	8	29.000	0	0	6
3	1	2	32.000	0	0	5
4	5	6	50.000	0	0	6
5	1	3	166.000	3	1	7
6	5	7	202.000	4	2	7
7	1	5	1326.750	5	6	0

Fuente: SPSS 20 IBM

**SPSS**, genera un gráfico que por su forma se le conoce como **“carámbano”** como el de la **Figura 14.21**.

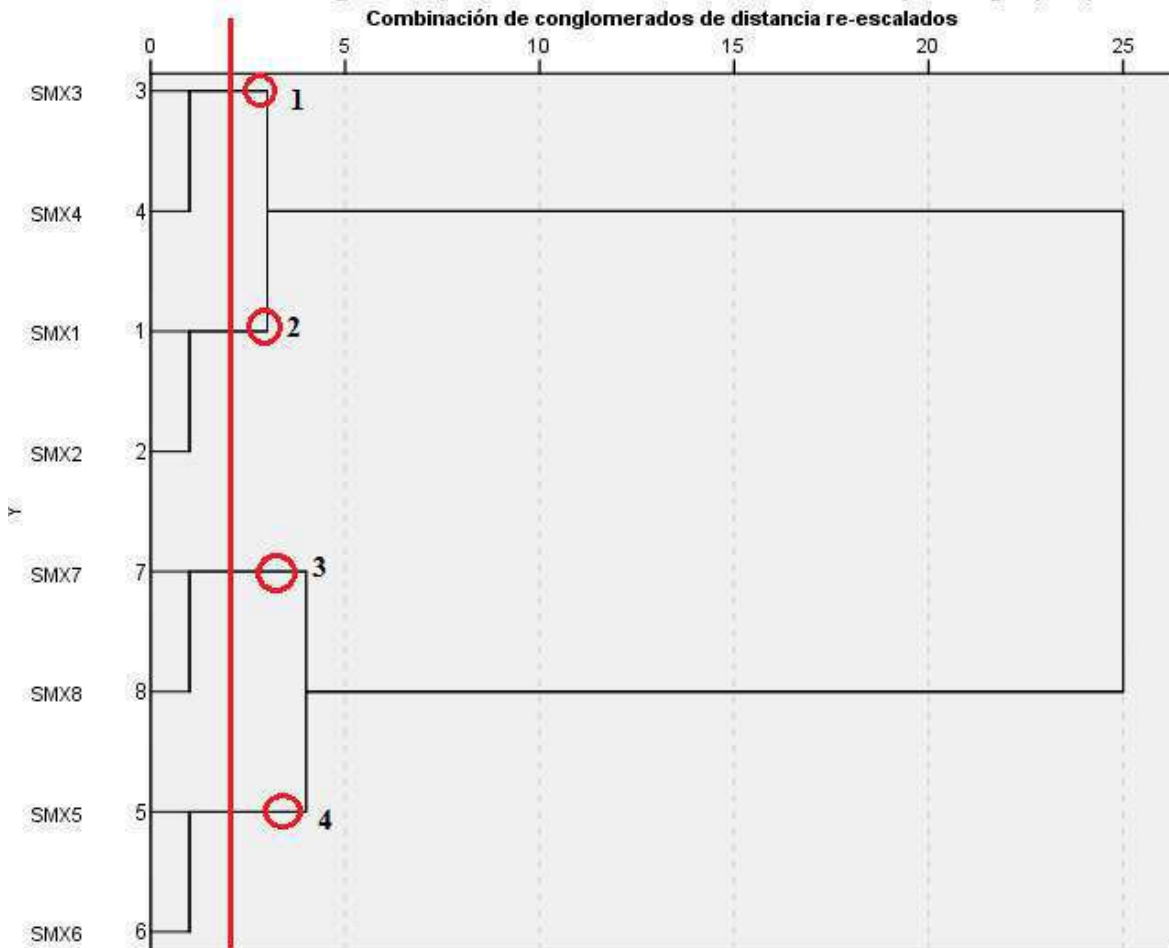
**Figura 14.21. Gráfico de carámbano**



Fuente: SPSS 20 IBM

Sin embargo, el más usado y representativo es el Dendograma de la **Figura 14.22**.

**Figura 14.22.- Dendograma que utiliza la vinculación media (entre grupos)**



Fuente: SPSS 20 IBM

Como se aprecia, las sucursales que se agrupan inicialmente, son:

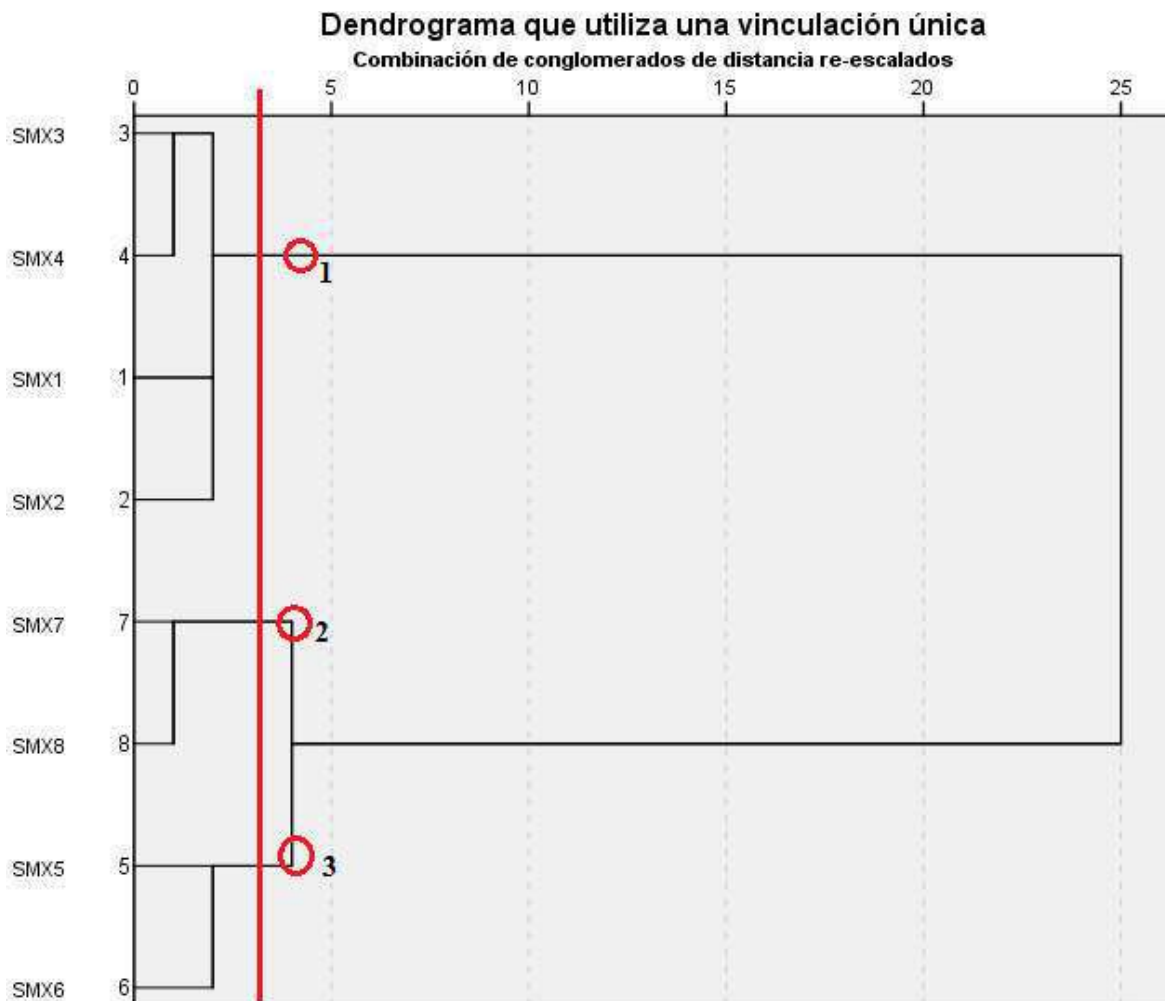
- SMX3 y SMX4**
- SMX1 y SMX2**
- SMX7 y SMX8**
- SMX5 y SMX6**

### **Paso 6: Validación**

Así, trazando una línea vertical que cruce con el agrupamiento en Combinación de conglomerados de distancia re-escalados, con distancias próximas observamos que se generan **4 clusters** de sucursales que presentan similitudes. En este punto habría que recurrir a los antecedentes que se tenga de las tiendas para explicar el porqué dichas sucursales presentan comunales y validar su conglomeración. Se sugiere intentar con los métodos adicionales de conglomeración como: Agrupación de centroides, Vecino más próximo, Vecino más lejano, **Método de Ward**, etc. para verificar éstas técnicas explican mejor las aglomeraciones generadas. Por ejemplo, la **Figura 14.23**.

muestra la opción con el Método de conglomeración Vecino más próximo, el cual arroja **3 clusters**.

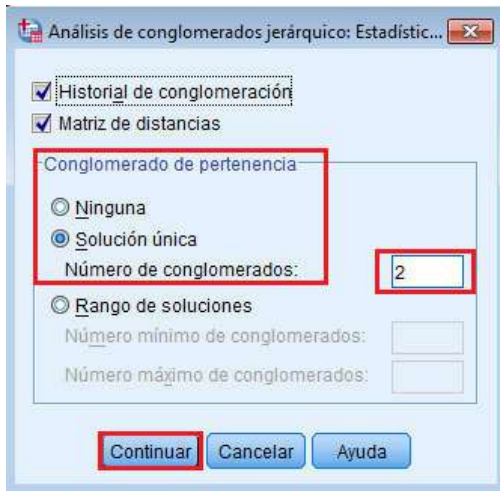
**Figura 14.23. Dendrograma por método de conglomeración vecino más próximo.**



Fuente: SPSS 20 IBM

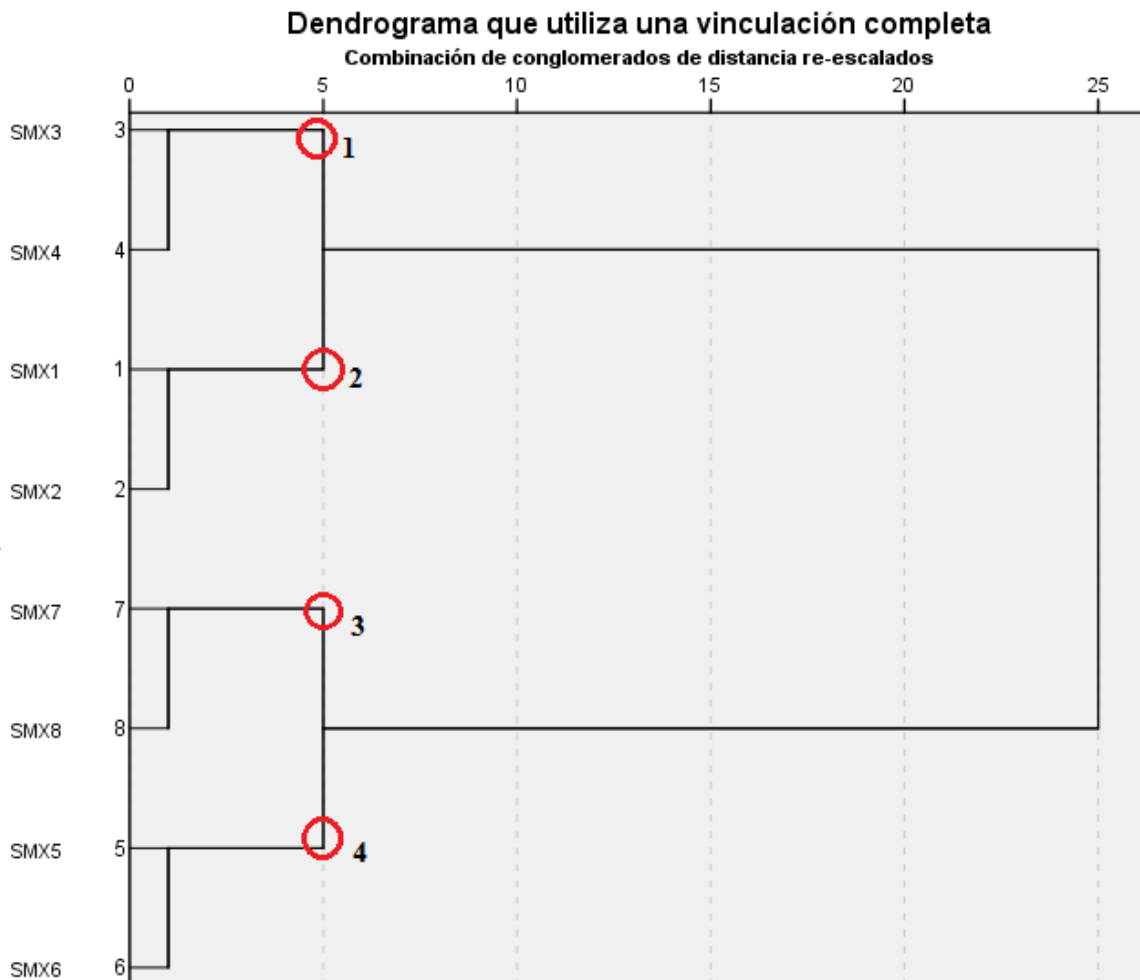
Así también, es posible generar la solución única en la que se podrá escoger la cantidad de conglomerados estimada. El resultado apunta a **4 cluster**. Ver **Figura 14.24, Figura 14.25**

**Figura 14.24. Cuadro de diálogo para solución única**



Fuente: SPSS 20 IBM

**Figura 14.25. Dendrograma por solución única**

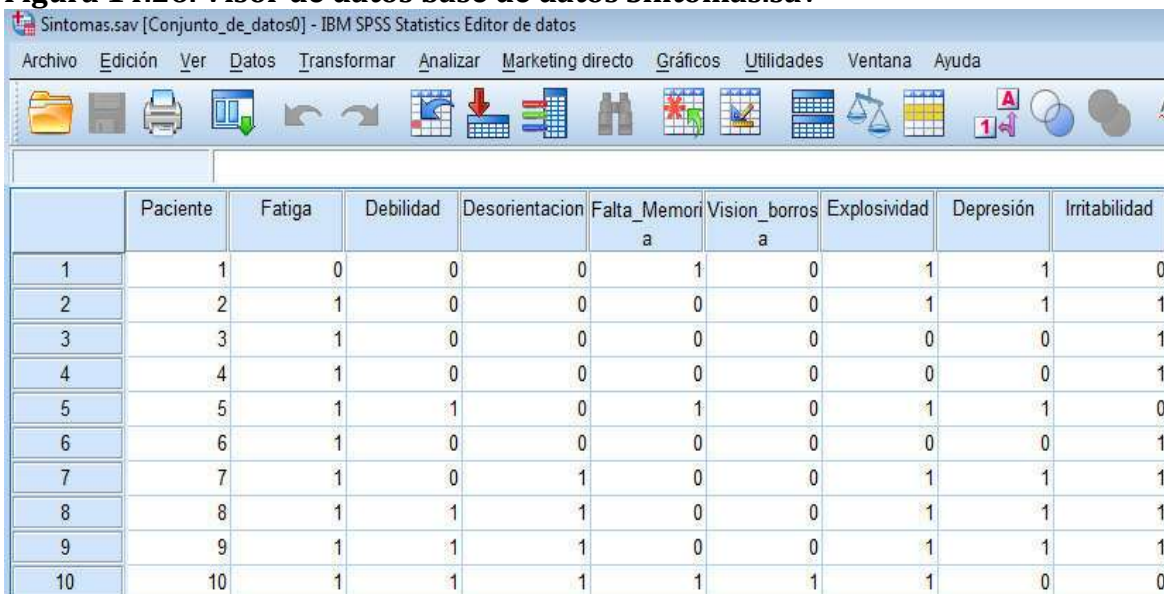


Fuente: SPSS 20 IBM

## Paso 1: Objetivos. Caso datos binarios

Problema 2: las ciencias de la administración, permiten apoyar al sector médico, por ejemplo, al clasificar una serie de síntomas que presentan los pacientes a los que se les suministran ciertos medicamentos y de los que se espera predecir los grupos en los que se puedan analizar con mayor detenimiento. En este caso, se tienen datos binarios. Ver archivo **Sintomas.sav** **Figura 14.26 y 14.27**

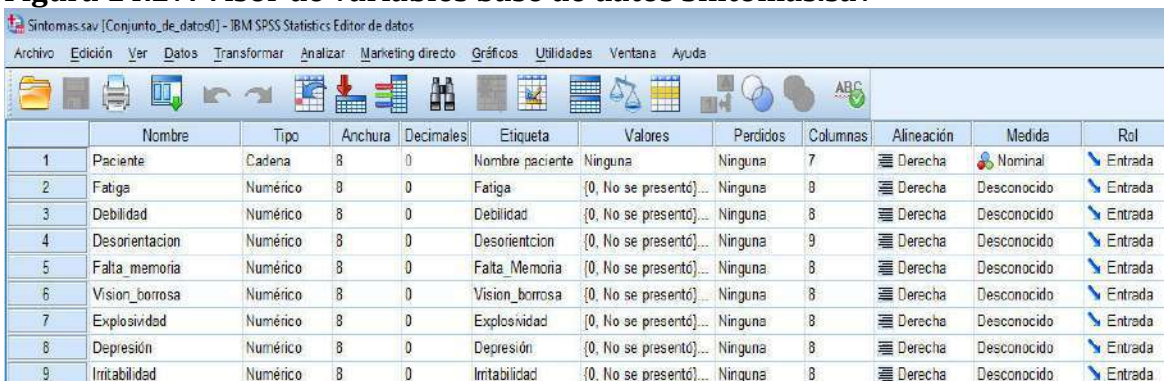
**Figura 14.26. Visor de datos base de datos Sintomas.sav**



	Paciente	Fatiga	Debilidad	Desorientacion	Falta_Memoria	Vision_borrosa	Explosividad	Depresión	Irritabilidad
1	1	0	0	0	1	0	1	1	0
2	2	1	0	0	0	0	1	1	1
3	3	1	0	0	0	0	0	0	1
4	4	1	0	0	0	0	0	0	1
5	5	1	1	0	1	0	1	1	0
6	6	1	0	0	0	0	0	0	1
7	7	1	0	1	0	0	1	1	1
8	8	1	1	1	0	0	1	1	1
9	9	1	1	1	0	0	1	1	1
10	10	1	1	1	1	1	1	0	0

Fuente: SPSS 20 IBM

**Figura 14.27. Visor de variables base de datos Sintomas.sav**



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Percidos	Columnas	Alineación	Medida	Rol
1	Paciente	Cadena	8	0	Nombre paciente	Ninguna	Ninguna	7	Derecha	Nominal	Entrada
2	Fatiga	Numérico	8	0	Fatiga	{0, No se presentó}...	Ninguna	8	Derecha	Desconocido	Entrada
3	Debilidad	Numérico	8	0	Debilidad	{0, No se presentó}...	Ninguna	8	Derecha	Desconocido	Entrada
4	Desorientacion	Numérico	8	0	Desorientacion	{0, No se presentó}...	Ninguna	9	Derecha	Desconocido	Entrada
5	Falta memoria	Numérico	8	0	Falta Memoria	{0, No se presentó}...	Ninguna	8	Derecha	Desconocido	Entrada
6	Vision borrosa	Numérico	8	0	Vision borrosa	{0, No se presentó}...	Ninguna	8	Derecha	Desconocido	Entrada
7	Explosividad	Numérico	8	0	Explosividad	{0, No se presentó}...	Ninguna	8	Derecha	Desconocido	Entrada
8	Depresión	Numérico	8	0	Depresión	{0, No se presentó}...	Ninguna	8	Derecha	Desconocido	Entrada
9	Irritabilidad	Numérico	8	0	Irritabilidad	{0, No se presentó}...	Ninguna	8	Derecha	Desconocido	Entrada

Fuente: SPSS 20 IBM

## Paso2: Diseño

Se resuelven las preguntas sugeridas, como sigue:

1. ¿Qué hacer si hay datos atípicos? Se considera que no existen casos atípicos
2. ¿La similitud de los sujetos, cómo debería medirse? Se propone medir con la técnica de **Agrupación de centroides, Intervalo: de Jaccard.**
3. ¿Deben estandarizarse los datos? Por lo respondido en los 2 puntos anteriores, **NO** se considera en el caso.



### Paso 3: Condiciones de aplicabilidad

Las exigencias de normalidad, linealidad y homocedasticidad **NO** son críticos en el análisis cluster. Sin embargo, sí lo son:

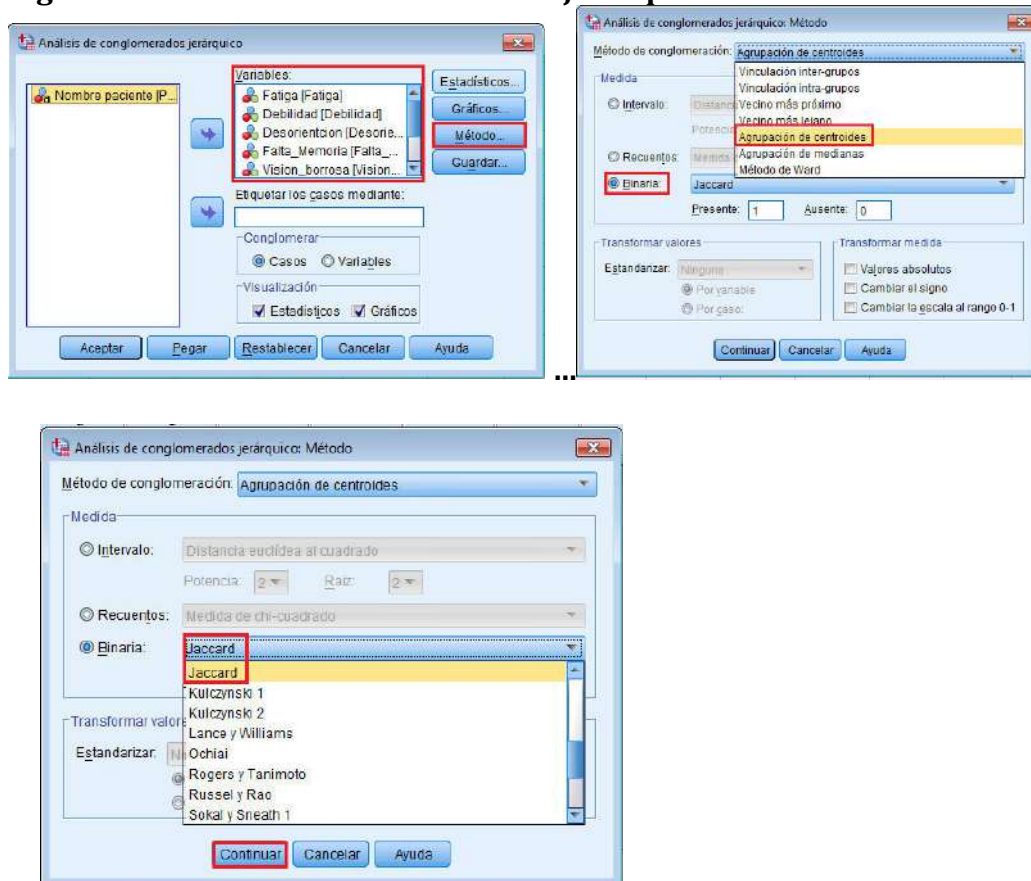
1. La muestra y su representatividad, se consideran sin problema para el ejemplo ya que parten de un **censo de datos de 10 pacientes**.
2. La multicolinealidad y su impacto, se consideran sin problema para el ejemplo.

### Paso 4: Estimación y ajuste

Teclear: **Analizar->Clasificar->Conglomerados jerárquicos->Variables** (selección métricas, en nuestro caso: **Fatiga, Debilidad, Desorientación, Falta\_memoria, Vision\_borrosa, Explosividad, Depresión, Irritabilidad**)->**Estadísticos**; seleccionar: **Historial de conglomerados y Matriz de distancias; Conglomerados de pertinencia**; seleccionar: **Ninguna->Continuar->Gráficos**; seleccionar: **Dendograma**; **Témpanos: Todos los conglomerados**; **Orientación: Vertical->Continuar->Método de conglomeración: Agrupación de centroides; Medida; intervalo: Binaria; Jaccard->Estandarizar: ninguna->Continuar->Conglomerar: Casos; Visualización: Estadísticos; Gráficos->Aceptar.**

Ver Figura 14.28, los cuadros de diálogo a afectar, con base a la Figura 14.17.

**Figura 14.28. Proceso análisis clúster jerárquico. Caso Binario.**

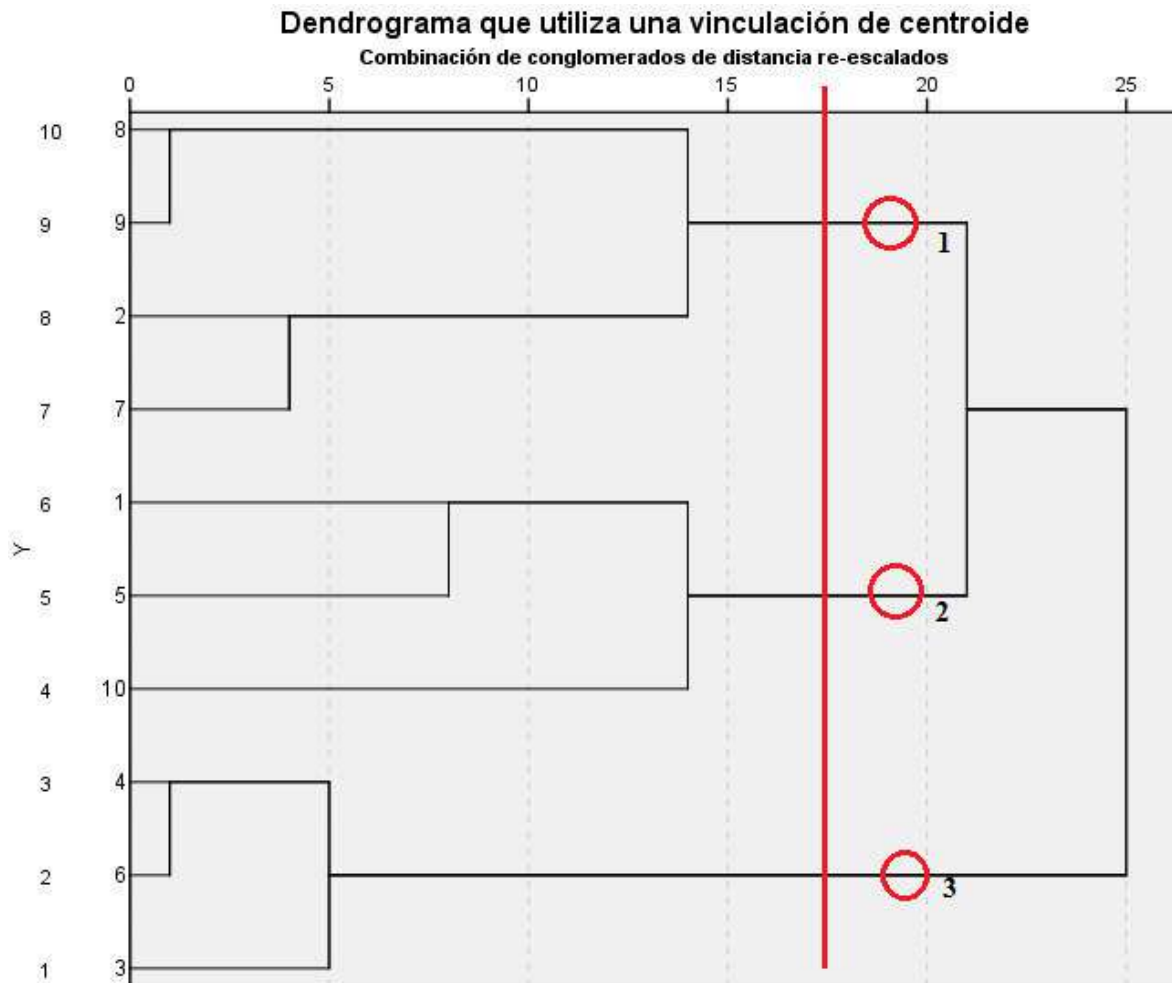


Fuente: SPSS 20 IBM

## Paso 5: Interpretación y Paso 6: Validación

SPSS, genera las tablas vistas en el problema 1, siendo de particular interés el **Dendrograma**. Ver **Figura 14.29**.

**Figura 14.29. Dendrograma caso binario.**



Fuente: SPSS 20 IBM

Así, trazando una línea vertical que cruce con el agrupamiento en Combinación de conglomerados de distancia re-escalados, con distancias próximas observamos que se **3 clusters**, el número **1**, con **4** pacientes (**8,9,2,7**); el número **2**, con **3** pacientes (**2,7,1**) y el número **3**, con **3** pacientes (**4,6,3**). En este punto habría que recurrir a los antecedentes que se tenga de los pacientes para explicar el porqué presentan se agrupan en sus síntomas en la toma de las medicinas de prueba y validar su conglomeración. Se sugiere intentar con los métodos adicionales de conglomeración como: Agrupación de centroides, Vecino más próximo, Vecino más lejano, **Método de Ward**, etc. para verificar éstas técnicas explican mejor las aglomeraciones generadas.



### Paso 1: Objetivos. Solución Mixta (Jerárquica-No Jerárquica)

**Problema 3:** La empresa MKT Digital de su base WEB\_MKT\_Digital.sav requiere identificar los grupos de opinión de sus 100 clientes respecto a 7 características clave de sus servicios de diseño Web como lo son: la tecnología ( $X_1$ ), el precio ( $X_2$ ), el servicio de planeación ( $X_3$ ), la imagen ( $X_4$ ), la experiencia ( $X_5$ ), la calidad ( $X_6$ ), el desempeño ( $X_7$ ). Ver Figura 14.30 y Figura 14.31.

**Figura 14.30. Visor de datos base de datos WEB\_MKT\_Digital**

	ID	V3	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
1	1	0	4.1	.6	6.9	4.7	2.4	2.4	5.2	0	32.0	4.2	1	0	1	1
2	5	0	6.0	.9	9.6	7.8	3.4	4.6	4.5	0	58.0	6.8	1	0	1	3
3	7	0	4.6	2.4	9.5	6.6	3.5	4.5	7.6	0	46.0	5.8	1	0	1	1
4	9	1	5.5	1.6	9.4	4.7	3.5	3.0	7.6	0	63.0	5.4	1	0	1	3
5	12	1	3.9	2.2	9.1	4.6	3.0	2.5	8.3	0	47.0	5.0	1	0	1	2
6	13	1	2.8	1.4	8.1	3.8	2.1	1.4	6.6	1	39.0	4.4	0	1	0	1
7	14	1	3.7	1.5	8.6	5.7	2.7	3.7	6.7	0	38.0	5.0	1	0	1	1
8	15	1	4.7	1.3	9.9	6.7	3.0	2.6	6.8	0	54.0	5.9	1	0	0	3
9	16	1	3.4	2.0	9.7	4.7	2.7	1.7	4.8	0	49.0	4.7	1	0	0	3
10	18	1	4.9	1.8	7.7	4.3	3.4	1.5	5.9	0	40.0	5.6	1	0	0	2
11	19	1	5.3	1.4	9.7	6.1	3.3	3.9	6.8	0	54.0	5.9	1	0	1	3
12	20	1	4.7	1.3	9.9	6.7	3.0	2.6	6.8	0	55.0	6.0	1	0	0	3
13	21	1	3.3	.9	8.6	4.0	2.1	1.8	6.3	0	41.0	4.5	1	0	0	2
14	22	1	3.4	.4	8.3	2.5	1.2	1.7	5.2	0	35.0	3.3	1	0	0	1
15	25	1	5.1	1.4	8.7	4.8	3.3	2.6	3.8	0	49.0	4.9	1	0	0	2
16	26	0	4.6	2.1	7.9	5.8	3.4	2.8	4.7	0	49.0	5.9	1	0	1	3
17	28	0	5.2	1.3	9.7	6.1	3.2	3.9	6.7	0	54.0	5.8	1	0	1	3

Fuente: SPSS 20 IBM

**Figura 14.31. Visor de variables de base de datos WEB\_MKT\_Digital**

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	ID	Númérico	3	0	id	Ninguna	Ninguna	4	Centrado	Escala	Entrada
2	V3	Númérico	1	0	Tamaño de la empresa	{0, Empresa...	Ninguna	25	Derecha	Escala	Entrada
3	X1	Númérico	3	1	Web tecnología	Ninguna	Ninguna	4	Centrado	Escala	Entrada
4	X2	Númérico	3	1	Web precio del servicio	Ninguna	Ninguna	4	Centrado	Escala	Entrada
5	X3	Númérico	4	1	Web planeación estratégica	Ninguna	Ninguna	4	Centrado	Escala	Entrada
6	X4	Númérico	3	1	Web imagen	Ninguna	Ninguna	4	Centrado	Escala	Entrada
7	X5	Númérico	3	1	Web experiencia usuario	Ninguna	Ninguna	4	Centrado	Escala	Entrada
8	X6	Númérico	3	1	Web calidad	Ninguna	Ninguna	4	Centrado	Escala	Entrada
9	X7	Númérico	4	1	Web desempeño	Ninguna	Ninguna	4	Centrado	Escala	Entrada
10	X8	Númérico	1	0	Web analítica	{0, Gratuita}...	Ninguna	3	Centrado	Ordinal	Entrada
11	X9	Númérico	4	1	Web contrataciones de clientes	Ninguna	Ninguna	4	Centrado	Escala	Entrada
12	X10	Númérico	3	1	Web satisfacción	Ninguna	Ninguna	4	Centrado	Escala	Entrada

Fuente: SPSS 20 IBM

## Paso2: Diseño

Se resuelven las preguntas sugeridas, como sigue:

1. ¿Qué hacer si hay datos atípicos? Se considera que no existen casos atípicos. El **dendrograma** y la **tabla historial de conglomeración**, proporcionan un medio de identificación de **atípicos** de la muestra. El **primero**, permite una **inspección visual de los atípicos**, los cuales se presentan como una **“rama”** que no se une hasta muy tarde. Así también, se puede identificar fácilmente con los **conglomerados pequeños, que exhiben grandes “ramas” únicamente para un número reducido de observaciones. Con el segundo**, Usted puede encontrar **conglomerados de un único componente** con facilidad mediante el uso de software estadístico.
2. ¿La similitud de los sujetos, cómo debería medirse? Se propone medir con el **método de Ward con, Intervalo: distancia euclídea al cuadrado**. Se escoge el **método de Ward** para minimizar las diferencias dentro del conglomerado y evitar problemas con el **“encadenamiento”** de las observaciones encontradas en el me todo de encadenamiento simple.
3. ¿Deben estandarizarse los datos? Por lo respondido en los 2 puntos anteriores, **NO** se considera en el caso.

## Paso 3: Condiciones de aplicabilidad

Las exigencias de normalidad, linealidad y homocedasticidad **NO** son críticos en el análisis cluster. Sin embargo, sí lo son:

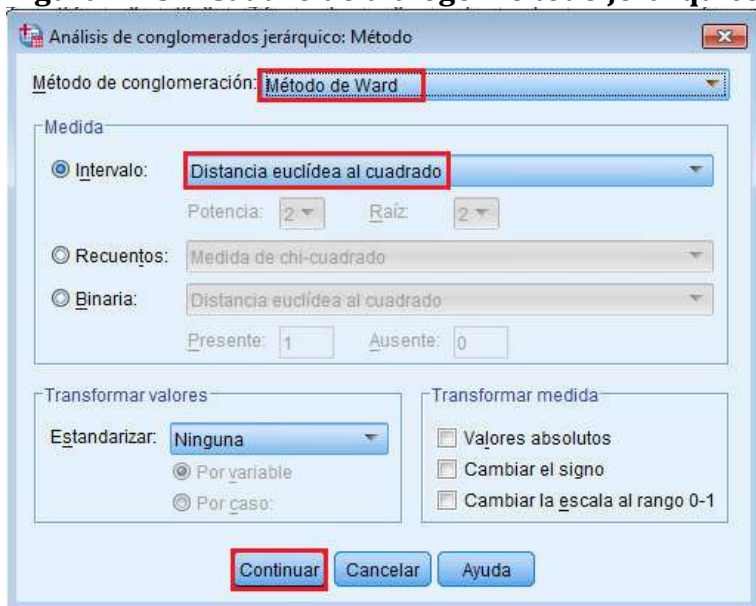
1. La muestra y su representatividad, se consideran sin problema para el ejemplo ya que parten de un **censo de datos de 100 clientes**
2. La multicolinealidad y su impacto, se consideran sin problema para el ejemplo.

## Paso 4: Estimación y ajuste. Etapa Jerárquica

Teclar: **Analizar->Clasificar->Conglomerados jerárquicos->Variables (selección métricas, en nuestro caso: Fatiga, Debilidad, Desorientación, Falta\_memoria, Vision\_borrosa, Explosividad, Depresión, Irritabilidad)->Estadísticos; seleccionar: Historial de conglomerados y Matriz de distancias; Conglomerados de pertinencia; seleccionar: Ninguna->Continuar->Gráficos; seleccionar: Dendograma; Témpanos: Todos los conglomerados; Orientación: Vertical->Continuar->Método de conglomeración: **Método de Ward; Medida de intervalo: Distancia euclídea al cuadrado**->Estandarizar: ninguna->Continuar->Conglomerar: Casos; Visualización: Estadísticos; Gráficos->Aceptar.**

Ver Figura 14.32, los cuadros de diálogo a afectar, con base a la Figura 14.17.

**Figura 14.32. Cuadro de diálogo mé todo jerárquico de Ward.**



Fuente: SPSS 20 IBM

### Paso 5: Interpretación

El SPSS genera la **tabla Historial de conglomeración y el dendograma. Ver Tabla 14.33 y Figura 14.34**

**Tabla 14.33. Historial de conglomerados Vinculación de Ward**

Historial de conglomeración						
Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerad o 1	Conglomerad o 2		Conglomerad o 1	Conglomerad o 2	
1	8	12	.000	0	0	60
2	2	22	.005	0	0	94
3	60	61	.010	0	0	74
4	26	33	.020	0	0	78
5	11	17	.040	0	0	60
6	37	47	.070	0	0	39
7	66	70	.105	0	0	45
8	29	43	.140	0	0	72
9	10	48	.175	0	0	65
10	19	34	.210	0	0	63

.....Continúa

34	4	41	3.141	0	0	55
35	73	78	3.431	0	0	57
36	56	65	3.766	0	0	43
37	16	32	4.116	0	0	64
38	27	50	4.536	0	0	81
39	3	37	5.006	0	6	77
40	6	13	5.526	0	0	51
41	69	75	6.051	0	0	53
42	86	95	6.576	0	0	91
43	56	62	7.108	36	0	50
44	36	45	7.643	0	0	59
45	66	88	8.200	7	26	70
46	14	30	8.760	0	0	71
47	54	82	9.420	0	0	57
48	71	90	10.250	0	0	53
49	68	97	11.090	0	0	68
50	56	74	11.976	43	0	56
51	6	20	13.036	40	0	71
52	28	38	14.479	17	16	65
53	69	71	15.981	41	48	73
54	7	21	17.570	0	15	59
55	4	31	19.225	34	0	67
56	56	63	21.272	50	0	58
57	54	73	23.527	47	35	88
58	56	67	25.880	56	22	75
59	7	36	28.256	54	44	80

....Continúa

70	55	66	64.375	32	45	83
71	6	14	68.591	51	46	90
72	1	29	73.077	23	8	84
73	69	96	77.881	53	19	85
74	53	60	82.779	62	3	82
75	52	56	88.128	30	58	79
76	10	23	93.517	65	13	92
77	3	8	98.971	39	60	86
78	9	26	104.829	61	4	90
79	52	77	111.619	75	0	91
80	4	7	118.524	67	59	81
81	4	27	126.001	80	38	86
82	51	53	134.767	28	74	85

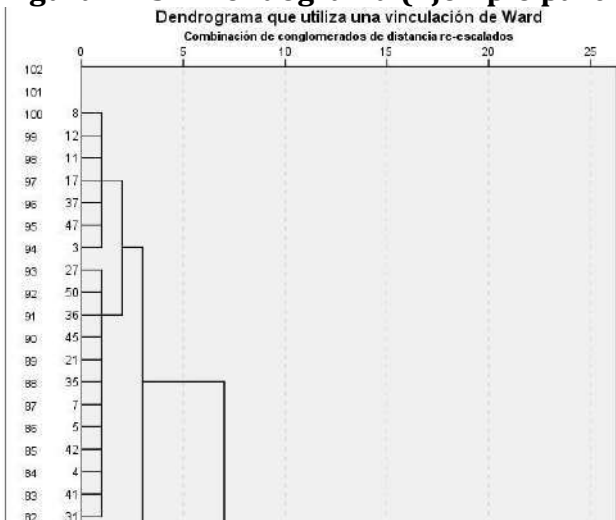
.... Continúa

90	6	9	258.717	71	78	95
91	52	86	281.415	79	42	97
92	1	10	305.053	84	76	95
93	54	57	333.106	88	87	96
94	2	3	364.923	2	86	98
95	1	6	398.113	92	90	98
96	51	54	446.314	89	93	97
97	51	52	523.012	96	91	99
98	1	2	614.935	95	94	99
99	1	51	994.718	98	97	0

Fuente: SPSS 20 IBM

En esta tabla, se muestra en el lado izquierdo del **coeficiente de aglomeración** los conglomerados que se están combinando. En las columnas de la mano derecha, se anotan los pasos en que cada conglomerado se va formando. **Una observación que nunca se ha unido a un conglomerado tiene un nivel 0.** En los primeros **38 pasos**, las observaciones únicas se van uniendo. Sólo en el **paso 39** el análisis cluster consigue unir un conglomerado formado en otro paso. Esta información también se usa para identificar observaciones únicas que **se unen tardíamente** al proceso de elaboración de **conglomerados potenciales atípicos**. Mirando hacia el **Paso 99**, vemos que en el **Paso 94** (6 conglomerados) se unió un conglomerado **formado** en el **Paso 2**. Esto significa que si seleccionamos una solución de **7 conglomerados**, uno de los conglomerados tendría sólo **2 observaciones**. Podemos ver también que el conglomerado de miembro único se unió en el **Paso 79**. Por tanto, si el análisis se confina en un número reducido de conglomerados (digamos, **10** o menos), entonces sólo tiene un problema potencial (el conglomerado de **2** miembros) que tratar. En este caso, la selección de menos de **7 conglomerados** elimina la necesidad de cualquier vuelta de especificación del análisis cluster. Del Dendograma (**Figura 14.33**), se aprecian **4 clusters importantes para analizar**.

**Figura 14.34. Dendograma (Ejemplo parcial)**



Fuente: SPSS 20 IBM



#### Paso 4: Estimación y ajuste. Etapa No Jerárquica (Uso de K-Medias)

Teclear: Analizar->Clasificar->Conglomerado de K medias->Variables ( $X_1$ ,  $X_2$ ,  $X_3$ ,  $X_4$ ,  $X_5$ ,  $X_6$ ,  $X_7$ ); Número de conglomerados: 4->Iterar->Iteraciones máximas: 10->Continuar->Opciones->Estadísticos: Centros de conglomerados iniciales; Tabla de Anova; Información del conglomerado para cada caso->Continuar->Aceptar. Ver Figura 14.35.

Figura 14.35. Proceso análisis cluster. No jerárquico. Uso de K Medias



Fuente: SPSS 20 IBM

## Paso 5: Interpretación.No jerárquica. Uso de Kmedias

SPSS genera la **tabla Centros iniciales de los conglomerados**. Ver **Figura 14.36**.

**Figura 14.36. Tabla Centros iniciales de los conglomerados**

**Centros iniciales de los conglomerados**

	Conglomerado			
	1	2	3	4
Web tecnología	1.0	6.0	4.9	3.0
Web precio del servicio	1.9	.9	4.4	3.8
Web planeación estratégica	7.1	9.6	7.4	5.5
Web imagen	4.5	7.8	6.9	4.9
Web experiencia usuario	1.5	3.4	4.6	3.4
Web calidad	3.1	4.6	4.0	2.6
Web desempeño	9.9	4.5	9.6	6.0

Fuente: SPSS 20 IBM

La cual reporta los centros que se generan en los **4 clusters** propuestos y en las **10** iteraciones máximas en la idea de identificar, óptimamente, los grupos que se encuentren con las distancias más contrastadas o diferenciadas. Esto es, Una vez que el **SPSS** ha calculado los centros iniciales de los conglomerados, entonces se asigna cada caso al cluster más cercano dependiendo su distancia al centro de los conglomerados. De esta forma se reinicia el cálculo con los elementos en torno a cada cluster y se detiene hasta que se alcanzan las **10** iteraciones indicadas o hasta que se eficientiza el valor de las distancias de cada cluster.

El **SPSS** genera la **tabla Historial de iteraciones**. Ver **Figura 14.37**.

**Figura 14.37. Historial de iteraciones**

**Historial de iteraciones<sup>a</sup>**

Iteración	Cambio en los centros de los conglomerados			
	1	2	3	4
1	2.160	3.256	2.434	2.503
2	.480	.424	.730	.801
3	.326	.130	.228	.653
4	.141	.384	.603	.528
5	.101	.387	.154	.495
6	.000	.327	.000	.266
7	.000	.215	.000	.155
8	.000	.241	.000	.133
9	.000	.134	.000	.072
10	.000	.000	.000	.000

a. Se ha logrado la convergencia debido a que los centros de los conglomerados no presentan ningún cambio o éste es pequeño. El cambio máximo de coordenadas absolutas para cualquier centro es de .000. La iteración actual es 10. La distancia mínima entre los centros iniciales es de 5.286.

La cual muestra cómo van cambiando los centros de los conglomerados a medida que se ingresan los casos por cluster y también, en la medida que se realizan las iteraciones hasta llegar a estabilizarse a partir de llegar al máximo de iteraciones (en nuestro caso **10**)

SPSS genera la **tabla Pertenencia a los conglomerados**. Ver **Figura 14.38**.

**Figura 14.38. Tabla Pertenencia a los conglomerados.**

**Pertenencia a los conglomerados**

Número de caso	Conglomerado	Distancia
1	4	2.093
2	2	3.690
3	2	1.848
4	2	1.475
5	2	2.312
6	4	2.267
7	2	1.576
8	2	1.309
9	4	1.495
10	4	1.617
11	2	.891
12	2	1.309
13	4	1.729
14	4	2.925
15	4	1.968
16	4	2.027
17	2	.888
18	4	2.104
19	4	2.309
20	4	2.872
...		
95	3	3.166
96	1	1.984
97	1	2.556
98	1	1.784
99	1	2.066
100	4	2.196

Fuente: SPSS 20 IBM

En esta tabla se muestra cada uno de los **100** casos de la base de datos **WEB\_MKT\_Digital.sav**, a qué conglomerado se agrupan y la distancia al centroide del grupo que representan.



SPSS genera una de las tablas más importantes que es la de **Centros de los conglomerados finales**. Ver Figura 14.39.

**Figura 14.39. Tabla Centros de los conglomerados finales.**

	Conglomerado			
	1	2	3	4
Web tecnología	2.0	4.9	3.4	4.1
Web precio del servicio	2.7	1.5	4.0	1.6
Web planeación estratégica	7.0	9.4	6.6	8.6
Web imagen	5.2	5.8	6.2	4.4
Web experiencia usuario	2.3	3.2	3.7	2.8
Web calidad	2.6	3.3	3.2	2.1
Web desempeño	8.2	7.0	8.0	5.3

Fuente: SPSS 20 IBM

En ésta tabla se puede observar, de manera muy directa la cantidad de casos que se agrupan en cada cluster. Cabe destacar que la base de datos se diseñó en una escala de respuestas de **1-10** representando los valores extremos de calificación desde pésimo hasta excelente, como apreciación de los clientes de la empresa **MKT Digital**, a cada uno de los **7** características clave de sus servicios presentados. Por ejemplo, en el **caso 1** de **Web tecnología**, en el **cluster 1**, los agrupados en dicho cluster tienen una muy baja calificación (**2.0**) de la tecnología empleada por la empresa **MKT Digital** a sus servicios. Por otro lado, en el **caso 3** Web planeación estratégica, en el **cluster 2**, los agrupados en dicho cluster tienen una muy alta calificación (**9.2**) de los servicios de planeación en el diseño web de la empresa **MKT Digital** a sus servicios.

Así, se puede afirmar, tomando de referencia los valores **>6** (como ejemplo), que los clientes de la empresa **MKT Digital**, **consideran de manera positiva los servicios que presta, de acuerdo a su agrupación en:**

- a. **Cluster 1**, los servicios de planeación estratégica (**7.0**) y desempeño (**8.2**).
- b. **Cluster 2**, los servicios de planeación estratégica (**9.4**) y desempeño (**7.0**).
- c. **Cluster 3**, los servicios de planeación estratégica (**6.6**), imagen (**6.2**) y desempeño (**8.0**).
- d. **Cluster 4**, el servicio de planeación estratégica.

Como se observa, la cercanía de los grupos implica que la baja diferenciación presente las mismas variables de caso entre los grupos como lo son los **servicios de planeación estratégica y desempeño**.

Lo anterior, permite establecer oportunidades para trabajar con el resto de las variables de caso que son percibidas con bajo valor por los clientes de la empresa **MKT Digital** y lograr una mayor diferenciación de los grupos.

SPSS genera la **tabla Distancias entre los centros de los conglomerados finales**. Ver Figura 14.40.

**Figura 14.40. Tabla Distancias entre los centros de los conglomerados finales**  
**Distancias entre los centros de los conglomerados finales**

Conglomerado	1	2	3	4
1		4.300	2.646	4.230
2	4.300		4.191	2.779
3	2.646	4.191		4.803
4	4.230	2.779	4.803	

Fuente: SPSS 20 IBM

La cual muestra que en este caso, los centros están muy cercanos entre sí y que por lo tanto los grupos **NO** son tan diferentes entre ellos. Una forma de solucionar es variar la cantidad de cluster a generar, con el fin de identificar mayor diferenciación de los grupos de variables de caso entre los cluster.

SPSS genera una **tabla ANOVA**. Ver **Figura 14.41**.

**Figura 14.41. Tabla ANOVA**

	Conglomerado		Error		F	Sig.
	Media cuadrática	gl	Media cuadrática	gl		
Web tecnología	37.108	3	.639	96	58.055	.000
Web precio del servicio	28.530	3	.583	96	48.960	.000
Web planeación estratégica	39.267	3	.755	96	51.985	.000
Web imagen	15.527	3	.835	96	18.598	.000
Web experiencia usuario	7.487	3	.348	96	21.509	.000
Web calidad	8.223	3	.355	96	23.132	.000
Web desempeño	53.222	3	.928	96	57.330	.000

Las pruebas F sólo se deben utilizar con una finalidad descriptiva puesto que los conglomerados han sido elegidos para maximizar las diferencias entre los casos en diferentes conglomerados. Los niveles críticos no son corregidos, por lo que no pueden interpretarse como pruebas de la hipótesis de que los centros de los conglomerados son iguales.

Fuente: SPSS 20 IBM

Donde los **coeficientes F** con la **sig. (p value <0.005)** que nos reporta que dichas variables contribuyen a las diferencias entre los cluster generados

SPSS, por último produce la **tabla Número de casos en cada conglomerado**. Ver **Figura 14.42**.

**Figura 14.42. Tabla Número de casos en cada conglomerado**

Número de casos en cada conglomerado		
Conglomerado	1	29.000
	2	19.000
	3	19.000
	4	33.000
Válidos		100.000
Perdidos		.000

Fuente: SPSS 20 IBM

En la cual, se resume la cantidad de casos a qué conglomerado pertenece.

**Conclusión.**

A partir de la **Figura 14.38. Tabla Centros de los conglomerados finales**, se establece que en **4** agrupaciones o **cluster**, **3/7** servicios son apreciados por los clientes de la empresa MKT Digital, a saber: **servicios de planeación estratégica, desempeño e imagen**. A pesar de que todas las variables enunciadas en la **Figura 14.40 de la Tabla ANOVA**, aportan en la diferenciación de los grupos, aún ésta es muy baja, por lo que se sugiere que la empresa trabaje en lograr una mayor diferenciación de sus servicios en las variables restantes como son: tecnología, precio del servicio (por cierto es la que presenta los valores más bajos en los clusters), experiencia en el usuario y calidad.

**14.13. Análisis cluster. Observaciones finales**

Con lo anterior, podemos resumir el método mixto aplicado, como se aprecia en la **Figura 14.43**.

**Figura 14.43. Método mixto**

<b>Análisis cluster jerárquico</b>	<b>Análisis clúster no jerárquico</b>
Técnica de carácter exploratorio que permite estudiar el método de aglomeración de las observaciones a través del gráfico de dendograma.	El algoritmo de kmedias, permite optimizar con respecto a la variabilidad intra-grupos, las soluciones cluster obtenidas aplicando las técnicas jerárquicas.
No se necesita conocer el número de cluster <i>a priori</i>	Las observaciones podrían ser reasignadas a otros grupos en distintas fases del método.
Un inconveniente se encuentra en la dificultad para discernir el número de grupos en algunos casos	El Inconveniente de kmedias es que el resultado obtenido depende de la elección inicial de centroides considerada y puede no proporcionar una solución óptima global.
Una vez que una observación es asignada a un grupo ya no es posible reasignarla a otro en una iteración posterior	-----

Fuente: propia

**El análisis cluster**, al igual que **el análisis factorial**, es muy útil como técnica de **reducción de datos**. Como se observa, algunos lo consideran más arte que ciencia,

por lo que tiende a sobreestimarse y abusar de ella fácilmente, provocando resultados indeseables. Diferentes medidas y algoritmos entre los sujetos **pueden afectar a los resultados**. En la mayoría de los casos, tanto consideraciones objetivas como subjetivas intervienen en la selección del conglomerado de la solución de cluster final. Así, se recomienda prudencia y considerar estos temas siempre que evalúe el impacto de todas las decisiones. Junto con el **análisis multidimensional, el análisis cluster debido a su falta de base estadística para inferir de la población, tiene una mayor necesidad de aplicarse varias veces bajo condiciones cambiantes**. De proceder con el cuidado necesario, el **análisis cluster** resulta ser un instrumento muy valioso en la identificación de factores subyacentes mediante realizar **agrupaciones (conglomerados)** sugeridas, de sujetos que no son identificables mediante otras técnicas multivariantes.

## Referencias

- Aldenderfer, M.S. y Blashfield R.K.(1984), *Cluster Analysis*. Thousand Oaks, Calif.: Sage Publications.
- Anderberg, M. (1973). *Cluster Analysis for Applications*. New York: Academic Press
- Bailey, K. D. (1994). *Typologies and Taxonomies: An Introduction to Classification Techniques*. Thousand Oaks, Calif.: Sage Publications.
- Everitt, B. ( 1980), *Cluster Analysis*, 2d ed. New York: Halsted Press.
- Green, P. E. (1978). *Analyzing Multivariate Data*. Hinsdale, Ill.: Holt, Rinehart y Winston.
- Hair, J.F.; Anderson, R.E.; Black, W.C. (1999). *Análisis Multivariante*. 5a. Ed. España:Prentice Hall.
- IBM (2011a). *IBM SPSS Statistics Base 20*. EUA. .Industrial Business Machines. Recuperado el 20161201 de:  
[ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM SPSS Statistics Base.pdf](ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Base.pdf)
- IBM (2011b). *Guía breve de IBM SPSS Statistics 20*. EUA.Industrial Business Machines. Recuperado el 20161201 de:  
[ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM SPSS Statistics Brief Guide.pdf](ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf)
- IBM (2011c). *IBM SPSS Missing Values 20*. EUA. .Industrial Business Machines. Recuperado el 20161201 de:  
[ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM SPSS Missing Values.pdf](ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Missing_Values.pdf)
- IBM (2011d). *IBM SPSS Statistics 20 Command Syntax Reference*. EUA. .Industrial Business Machines. Recuperado el 20161201 de:  
[ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/en/client/Manuals/IBM SPSS Statistics Command Syntax Reference.pdf](ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/en/client/Manuals/IBM_SPSS_Statistics_Command_Syntax_Reference.pdf)
- IBM (2012). *IBM SPSS Categories 21*. . EUA.Industrial Business Machines.

Recuperado el 20161201 de:

[ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/21.0/es/client/Manuals/IBM SPSS Categories.pdf](ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/21.0/es/client/Manuals/IBM_SPSS_Categories.pdf)

- McIntyre, R. M., y Blashfield, R.K ( 1980). ANearest Centroid Technique for Evaluating the Minimum- Variance Clustering Procedure. *Multivariate Behavioral Research*. 15: 225-38.
- Milligan, G. (1980). An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. *Psychometrika* 45 (September): 325-42.
- Milligan, G.W., and Cooper, M.C. (1985). An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika* 50(2): 159-79.
- Morrison, D. (1967). Measurement Problems in Cluster Analysis. *Management Science* 13 (August): 775-80.
- Overall, J. (1964). Note on Multivariate Methods for Profile Analysis. *Psychological Bulletin*. 61 (3): 195-98.
- Punj, G. y D. Stewart (1983), Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Journal of Marketing Research*. 20 (May): 134-48.
- Rohlf, F. J. (1970), «Adaptive Hierarchical Clustering Schemes. *Systematic Zoology* 19: 58.
- Schaninger, C. M., y W. C. Bass (1986). Removing Response-Style Effects in Attribute-Determinance Ratings to Identify Market Segments. *Journal of Business Research* 14: 237-52
- Shephard, R. (1966), Metric Structures in Ordinal Data. *Journal of Mathematical Psychology* 3: 287-315.

## Apéndice. Matriz de pruebas estadísticas sugeridas

¿Que desea realizar?	Número de variables /condiciones	Diseño	Paramétrico/No Paramétrico	Prueba Estadística Recomendada	Procedimiento Estadístico		
Búsqueda de diferencias entre condiciones	Una variable: 2 condiciones	Mediciones Independientes	Paramétrico	Prueba t de Muestras Independientes	Comparación de Medias	Prueba t de Muestras Independientes	
		Mediciones repetidas	Paramétrico	Prueba t Relacionada	Comparación de Medias	Prueba t de Muestras Pareadas	
		Mediciones Independientes	No Paramétrico	Mann-Whitney U	Pruebas No Paramétricas	Prueba Dos Muestras Independientes	
		Mediciones repetidas	No Paramétrico	Wilcoxon	Pruebas No Paramétricas	Prueba de Dos Muestras Relacionadas	
	Una variable: más de dos condiciones	Mediciones Independientes	Paramétrico	ANOVA de Un Factor de Mediciones Independientes	Modelo General Lineal	Univariado	
		Mediciones repetidas	Paramétrico	ANOVA de Un Factor de Mediciones Repetidas	Modelo General Lineal	Medidas Repetidas	
		Mediciones Independientes	No Paramétrico	Kruskal-Wallis	Pruebas No Paramétricas	K Muestras Independientes	
		Mediciones repetidas	No Paramétrico	Friedman	Pruebas No Paramétricas	K Muestras Relacionadas	
	Dos variables	Mediciones Independientes en ambas variables	Paramétrico	ANOVA de Dos Factores Independientes	Modelo General Lineal	Univariado	
		Mediciones una Independiente y otra de Medidas Repetidas	Paramétrico	ANOVA de Dos Factores de Mediciones Repetidas	Modelo General Lineal	Mediciones Repetidas	
		Mediciones Independientes en ambas variables	Paramétrico	ANOVA de Dos Factores de Diseño Combinado	Modelo General Lineal	Mediciones Repetidas	
	Más que una variable dependiente	Mediciones Independientes	Paramétrico	MANOVA Independiente	Modelo General Lineal	Multivariado	
		Mediciones Repetidas	Paramétrico	MANOVA Repetida	Modelo General Lineal	Mediciones Repetidas	
	Comparar conteo de frecuencias (Categorías)	NA	Asociación	No Paramétrico	Chi-Cuadrada	Descriptivo	Cruce-Tabular
	Correlación de variables	Dos variables	Correlacional	Paramétrico	Pearson	Correlacionar	Bivariado
				No Paramétrico	Spearman	Correlacionar	Bivariado
No Paramétrico				Kendall tau-b	Correlacionar	Bivariado	
	Más de dos variables	Correlacional	Paramétrico	Regresión Múltiple	Regresión	Lineal	
Reducción de datos	Muchas variables	Correlacional	Paramétrico	Análisis Factorial	Reducción de datos	Factor	
		Correlacional	Paramétrico	Análisis de confiabilidad	Escala	Análisis de confiabilidad	

Fuente: Hinton, P.R.; Brownlow, Ch.; McMurray, I y Cozens, B. (2004). *SPSS Explained*. USA: Routledge, Taylor y Francis Group

Zapopan, Jal. a 30 de Agosto de 2022

## **Dictamen de Obra AMIDI.DO.20220830.EMT2**

Los miembros del equipo editorial de la Academia Mexicana de Investigación y Docencia en Innovación (**AMIDI**), ver:

<https://www.amidibiblioteca.amidi.mx/index.php/AB/about/editorialTeam>

se reunieron para atender la invitación a dictaminar el libro:

### ***ESTADÍSTICA MULTIVARIANTE TOMO II. Técnicas Interdependientes con SPSS en las Ciencias Sociales***

Cuyo autor de la obra es el **Dr. Juan Mejía Trejo**

Dicho documento fue sometido al proceso de evaluación por pares doble ciego, de acuerdo a la política de la editorial, para su dictaminación de aceptación, ver:

<https://www.amidibiblioteca.amidi.mx/index.php/AB/procesodeevaluacionporparesen ciego>

Los miembros del equipo editorial se reúnen con el curador principal del repositorio digital para convocar:

1. Que el comité científico, de forma colegiada, revise los contenidos y proponga a los pares evaluadores que colaboran dentro del comité de redacción, tomando en cuenta su especialidad, pertinencia, argumentos, enfoque de los capítulos al tema central del libro, entre otros.
2. Se invita a los pares evaluadores a participar, formalizando su colaboración.
3. Se envía así, el formato de evaluación para inicio del proceso de evaluación doble ciego a los evaluadores elegidos de la mencionada obra.
4. El comité científico recibe las evaluaciones de los pares evaluadores e informa a el/la (los/las) autor(es/as), los resultados a fin de que se atiendan las observaciones, el requerimiento de reducción de similitudes, y recomendaciones de mejora a la obra.
5. La obra evaluada, consta de:

### **Introducción, 3 capítulos, referencias y apéndice en 194 páginas**

Av. Lázaro Cárdenas 3454 int. 6,  
Col. Jardines de los Arcos, C.P. 44500,  
Guadalajara, Jalisco, México  
Tel. Oficina. 33 3560 7860/ Cel. 3312809887  
editorial@scientiaetpraxis.amidi.mx



6. El desglose de su contenido, de describe a continuación

Capítulo	Páginas
Introducción	5
Capítulo 12. Análisis Factorial	6-73
Capítulo 13. Análisis Multidimensional y de Correspondencias	74-135
Capítulo 14. Análisis Cluster	136-191
Referencias	192-193
Apéndice. Matriz de pruebas estadísticas sugeridas	194

7. Una vez emitidas las observaciones, el requerimiento de reducción de similitudes, y recomendaciones de mejora a la obra por los evaluadores y todas ellas resueltas por el/la (los/las) autor(es/as), el resultado resalta que el contenido del libro:

- a. Reúne los elementos teóricos actualizados y prácticos desglosados en cada uno de sus capítulos.
- b. Los capítulos contenidos en la obra, muestran claridad en el dominio del tema, congruencia con el título central del libro, y una estructura consistente
- c. Se concluye finalmente, que la obra dictaminada, puede fungir como libro de texto principal o de apoyo tanto para estudiantes de licenciatura como de posgrado.

8. Por lo que el resultado del dictamen de aceptación de la obra fue:

**FAVORABLE PARA SU PUBLICACIÓN.**

Sirva la presente para los fines que a los Interesados convengan.

Atentamente

Dr. Carlos Omar Aguilar Navarro.

ORCID: <https://orcid.org/0000-0001-9881-0236>

Curador Principal AMIDI.Biblioteca

AMIDI