Análisis Factorial:
Una Sencilla y Útil Herramienta de Análisis

Marcelo De Fuentes Merc

Es caso frecuente en un estudio de mercado cuantitativo, el encontrarnos con largas listas de variables o reactivos a través de los cuales evaluamos la importancia de ciertos atributos para la categoría, o mediante los cuales calificamos el desempeño o imagen de un producto/servicio.

Independientemente del problema que esto significa al momento de su aplicación en campo, el proceso analítico se vuelve sumamente complejo y, a veces, poco objetivo. En este sentido, lo que más desea el investigador es poder reducir todas esas variables a tan sólo unas cuantas que "representen", lo mejor posible, lo que la lista completa inicial, y para ello utilizamos en análisis factorial.

De ninguna manera es mi intención hacer creer al lector que la técnica aquí presentada es "el último grito en procesos matemáticos para investigación", ya que se usa en el día a día desde hace muchos años y es, dentro de la estadística multivariada, una de las herramientas más sencillas; simplemente, aprovecho este espacio para dar una breve y rápida explicación de qué es, como se hace y para qué se utiliza.

Es importante aclarar que aquí expondré tan sólo el método más sencillo y comúnmente utilizado (factores ortogonales), mas no por ello necesariamente el único, ni tampoco el mejor existente, ya que depende de cada caso en particular.

CUADRO 1

	Marca "A"	Marca "B"	Marca "C"	Marca "D"
Calidad				
Consistencia				
Sabor en general al probarlo				
Sabor que deja en la boca				
Sabor casero				
Aroma al destapar				
Aroma al preparar				
Aspecto al destapar				
Aspecto al calentar y servir				
Color				
Cantidad de sal				
Nivel de picante				
Empaque				
Diversidad de presentaciones				
Diseño				
Facilidad de encontrarla				
Precio				
Facilidad de almacenamiento				
Facilidad de preparación				
Tiempo de preparación				
Tiempo que dura en la alacena				
Qué tanto gusta a la familia				

Un caso típico en el cual puede ser muy útil aplicar el análisis factorial, es cuando hacemos la evaluación de marcas (ver cuadro).

I. Definición y Aplicación

Comencemos con una definición:

El análisis factorial es un procedimiento mediante el cual se toma un gran número de variables u objetos y se investiga si tienen un pequeño número de factores en común que expliquen su intercorrelación.

Aunque el análisis factorial tiene otra aplicación (que es la identificación de estructuras subyacentes o dimensionalidad de los datos), la más importante, para efectos de este escrito, es la de reducir un elevado número de variables a uno mucho más pequeño que sea más fácil de manejar, lo cual es particularmente útil si los reactivos sobrepasan de 15 (este número es absolutamente arbitrario ya que no existe un mínimo o máximo estadísticamente establecido).

La esencia básica del análisis de factores se basa en que las variables guardan entre sí algún nivel de correlación, no obstante que son totalmente independientes, por lo que podemos agrupar a todas aquellas que se correlacionan fuertemente desde el punto de vista estadístico. Un aspecto fundamental de esta técnica es que trata a todas las variables por igual y no utiliza dependientes e independientes.

II. Matriz de Correlación y Dispersión de Datos

Tradicionalmente, lo primero que hacemos cuando nos enfrentamos a un listado de reactivos de calificación, es obtener la matriz de correlación para ver cuál es el nivel de relación estadística que hay entre las variables.

Una matriz de correlaciones que obtuvimos de un estudio (no mencionamos la marca y nombre del cliente por motivos de confidencialidad) es la que muestro a continuación (debo aclarar que el estudio consistía en un total de 15 reactivos, pero en este momento estoy incluyendo solamente cinco a manera de ejemplo (ver cuadro 2).

Antes de continuar, considero importante recordar lo siguiente:

- Si en la matriz de correlación dos variables distintas muestran un valor cercano a 1 (por ejemplo 0.93) quiere decir que están altamente correlacionadas entre sí y por lo tanto son básicamente lo mismo. Un caso frecuente de esto es cuando pedimos a la gente que evalúe a un hotel en cuanto a los siguientes dos factores:

  - Qué tan amable fue el servicio en general.

  - Qué tan amable fue el personal que lo atendió.

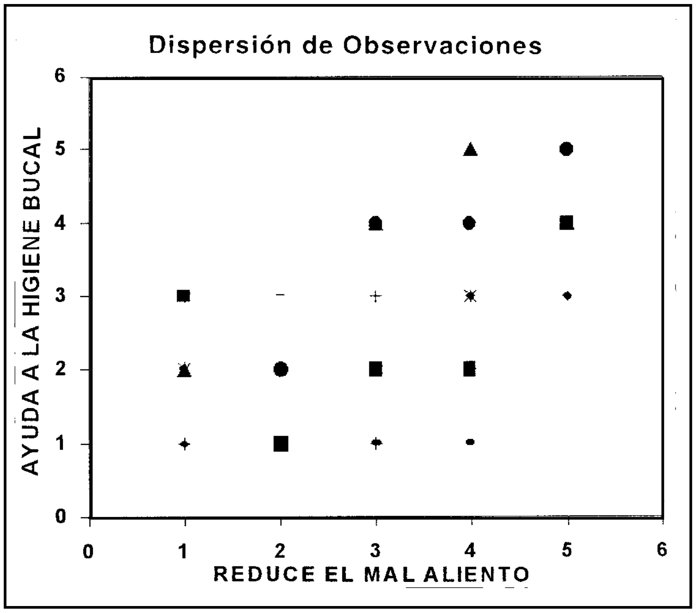
En este caso esperaríamos que la correlación entre ambos reactivos sea muy alta ya que significan prácticamente lo mismo.

- Por el contrario, si el valor es muy cercano a cero (por ejemplo 0.018) significa que entre ellas no existe casi relación estadística alguna.

Hasta ahora ya tenemos y sabemos lo que es la matriz de correlación. Continuando con un ejemplo, en la siguiente gráfica podrán ver los valores observados de dos variables, cualesquiera, en un estudio de mercado. Pensemos que graficamos "reduce el mal aliento" y "ayuda a la higiene bucal", aclarando que para la calificación utilizamos una escala de 5 puntos definida; esto es una gráfica de dispersión de observaciones.

CUADRO 2

	Reduce el el mal aliento	Ayuda a la higiene bucal	Deja la boca fresca	Previene la caries	Previene la gingivitis
Reduce el mal aliento	1.00	0.662	0.584	0.538	0.451
Ayuda a la higiene bucal	0.662	1.00	0.606	0.573	0.522
Deja la boca fresca	0.584	0.606	1.00	0.544	0.493
Previene la caries	0.538	0.573	0.544	1.00	0.627
Previene la gingivitis	0.451	0.522	0.493	0.627	1.00

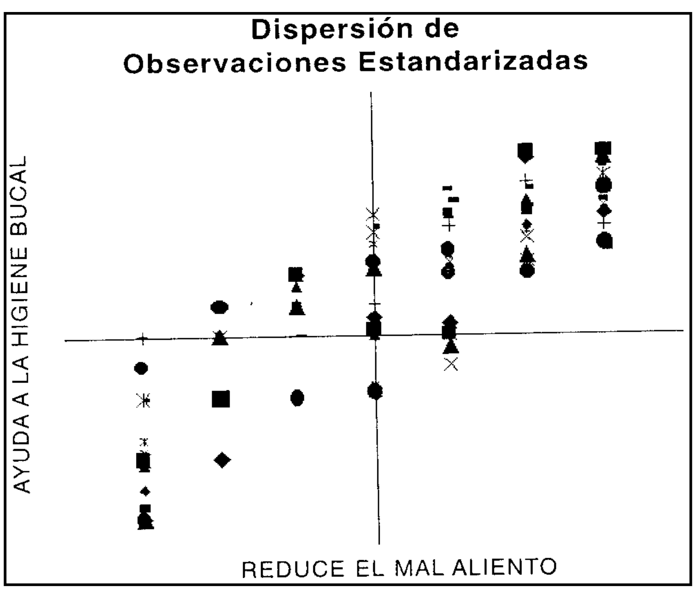


Aclaro que estoy graficando únicamente dos variables para no complicar el ejemplo, pero en la realidad se trata de poner todas las variables que incluimos en el estudio, no importando si fueron diez, veinte o cien.

III. Estandarización de Variables

Para poder analizar los datos de una forma más directa y objetiva, nuestro siguiente paso es estandarizar las variables; es decir, finalizar con variables cuya media es 0 y desviación estándar 1.

Si viéramos esto en la gráfica del ejemplo tendríamos algo como lo siguiente:



El promedio de las observaciones queda en el origen, es decir en el 0 tanto para la variable "reduce el mal aliento" como para "ayuda a la higiene bucal".

IV. Obtención de Factores

El análisis factorial genera "nuevas variables" que resultan de una combinación lineal de las originales; a las nuevas variables se les llama *factores* y a los coeficientes de cada combinación lineal se les conoce como *cargas factoriales*.

Hay un importante número de métodos para la obtención de factores, pero el más conocido es el *método de componentes principales*, el cual tiene como principio el derivar nuevos factores que no tienen la más mínima relación entre sí.

Si bien un poco más adelante podrán ver una representación gráfica de este método aplicado a nuestro ejemplo, permítanme primero dar la expresión matemática de los factores:

$$F = B_1X_1 + B_2X_2 + B_3X_3 + \dots + B_mX_m$$

En donde:

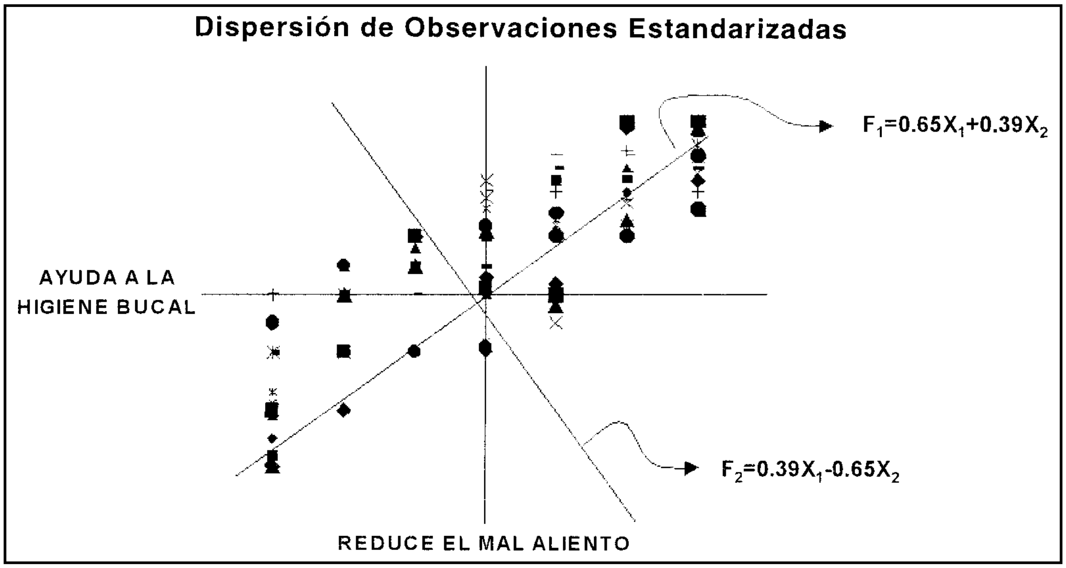
F = Es la combinación lineal que recibe el nombre de componente principal o factor.
B = Son los coeficientes de la combinación lineal que se conocen como cargas factoriales.
X = Son los valores originales observados.

Ustedes pueden ver que la ecuación anterior es la representación de una línea recta y para nuestro ejemplo de tan sólo dos variables se vería de la siguiente forma:

$F = B_1$(reduce el mal aliento) + B_2(ayuda a la higiene bucal)

Ya que conocemos la expresión matemática del nuevo factor, lo que procede es calcular las cargas factoriales (o B's) para poder encontrar las nuevas coordenadas del mismo. En este sentido es válido utilizar como ejemplo a la regresión lineal, la cual busca encontrar una línea recta que explique la mayor cantidad de varianza de las observaciones. Exactamente esto es lo que encontraremos con el nuevo factor al determinar los valores de las cargas factoriales, pero con la importante diferencia que en el análisis factorial no tendremos una variable dependiente y el resto independientes que explicarán a la primera, sino que todas son tratadas por igual.

Para mayor claridad, en nuestro ejemplo de tan sólo dos variables, gráficamente los resultados se verían de la siguiente forma:



El F_1 es la combinación lineal entre X_1 y X_2 que explica la mayor cantidad de varianza en la matriz de correlación y el factor 2 es el que resulta perfectamente perpendicular al primero, es decir, entre ambos factores no hay correlación alguna; a estos factores se les conoce como *ortogonales*.

Si en nuestro ejemplo substituimos un punto cualquiera (que es una combinación entre "reduce el mal aliento" y "ayuda a la higiene bucal") en el factor 1, tendríamos lo siguiente:

Punto: Reduce el mal aliento=3
Ayuda a la higiene bucal=2

F_1 = 0.65(3) + 0.39(2) = 2.73
F_2 = 0.39(3) – 0.65(2) = -0.13

Estos nuevos factores se encuentran en los ejes que están rotados y que explican la mayor cantidad de varianza. La varianza explicada con estos nuevos ejes se resta de la matriz, sin embargo todavía hay parte importante de la misma que no hemos logrado explicar, por lo que podríamos encontrar una nueva combinación lineal, es decir, unos nuevos factores, que minimicen la varianza que hasta el momento no hemos podido explicar.

Rotando constantemente los ejes iremos "dejando" cada vez menos varianza NO explicada hasta el punto en que la parte explicada sea lo suficientemente buena como para poder detener el proceso de rotación; es decir, estaremos haciendo un proceso iterativo.

Como ya mencioné antes, el procedimiento tradicional para la rotación de ejes es el *Análisis de Componentes Principales*, no obstante que existen otros procedimientos, para efectos de este documento nos centraremos en este para la segmentación. Para mayor claridad, a esta rotación también se le ha llamado *rotación ortogonal*, porque no permite la relación entre factores (tal y como vimos en la gráfica el F_1 y el F_2). Simplemente como información adicional, existe también la *rotación oblicua*, misma que SI permite que los factores se correlacionen entre sí, pero que no tocaremos en este documento.

V. Criterio para la Inclusión de Factores

Algo muy importante en el análisis factorial, es que en todo momento el investigador tendrá que estar tomando decisiones, a veces arbitrarias, basadas en su criterio. Un ejemplo de esto es en que momento parar la rotación o qué porcentaje de la varianza tenemos que explicar para tener un buen análisis.

Si quisiéramos tener el 100% de la varianza explicada entonces tendríamos que tomar todos los reactivos, por lo que el análisis factorial no tendría sentido alguno. Por el contrario, si nos quedamos con un solo factor estaremos asumiendo que todos los reactivos se comportan de la misma forma o son lo mismo, lo cual tampoco tiene sentido.

Una regla de dedo es que debemos tomar todos aquellos componentes o variables que tienen un autovalor mayor a 1 en la varianza explicada, aunque en algunas ocasiones (como en el ejemplo que verán a continuación), podemos modificar un poco el criterio bajando el autovalor a 0.9 o subiéndolo a 1.1.

En el siguiente cuadro muestro el mismo caso real del cual les había dicho que utilizamos un total de 15 reactivos para la evaluación de la categoría.

CUADRO 3

VARIANZA TOTAL EXPLICADA

	Autovalores iniciales			Sumas de saturación al cuadrado de la extracción			Suma de las saturaciones al cuadrado de la rotación		
Componente	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	7.724	51.493	51.493	7.724	51.493	51.493	4.176	27.837	27.837
2	1.283	8.553	60.046	1.283	8.553	60.046	3.217	21.445	49.282
3	.941	6.273	66.319	.941	6.273	66.319	2.556	17.037	66.319
4	.709	4.725	71.044						
5	.579	3.858	74.902						
6	.478	3.189	78.091						
7	.472	3.143	81.235						
8	.432	2.878	84.113						
9	.396	2.637	86.750						
10	.389	2.596	89.347						
11	.357	2.377	91.724						
12	.333	2.220	93.944						
13	.329	2.196	96.140						
14	.293	1.953	98.093						
15	.286	1.907	100.000						

Método de extracción: Análisis de Componentes Principales.

Explicaré lo más importante de este cuadro:

Columna componente: En esta columna se enlistan todos los reactivos que utilizamos en el cuestionario y equivaldrán a las combinaciones lineales que representan un factor. En el ejemplo anterior se puede ver que en este estudio tuvimos un total de 15 reactivos ya que ese sería el número máximo, más no óptimo, de factores que podemos tener.

Columna total: Esto es lo que conocemos como autovalor y aquí es en donde tomaremos la decisión de cuantas variables tomar. En este número es en el que nos tenemos que fijar que sea cercano a 1 y hasta ahí tomar los factores. Normalmente los paquetes estadísticos cuentan con este criterio pre-establecido y lo combinan con el porcentaje de la varianza total explicada para determinar cuantos factores tener .

% de la varianza: Esto es a lo que me refería cuando hablaba del porcentaje de la varianza que estamos explicando con el factor. Como pueden ver, el primer factor explica un 51.49% de la varianza total y el segundo tan solo un 8.55%.

% acumulado: Es simplemente el porcentaje de varianza total acumulada que explican nuestros factores. En el renglón dos pueden ver que es la suma de 51.493 y 8.553

Explicado lo anterior, en nuestro ejemplo vemos que de los 15 reactivos totales que teníamos en el estudio, hemos reducido a tan solo tres factores, que es en donde el autovalor llega a 0.941 y estamos explicando un poco más del 66% del total de la varianza con estos tres factores.

VI. Conjunción de Reactivos en un Factor

Hasta el momento ya tenemos todas las bases para encontrar el número de factores que estaremos utilizando en el estudio, por lo que el siguiente paso será determinar a qué reactivos agrupa cada uno de los factores.

El proceso es bastante sencillo ya que lo único que debemos hacer es observar las cargas que tiene cada uno de los reactivos por factor y así decidir como los agruparemos.

MATRIZ DE COMPONENTES ROTADOS

CUADRO 4

	Componente/factor 1	2	3	Comunalidades
Reduce la placa dentobacteriana	.807	.213	.181	.729
Elimina los gérmenes	.767	.199	.241	.686
Previene la caries	.763	.269	.184	.688
Previene la gingivitis	.743	.206	.219	.642
Ayuda a la higiene bucal	.666	.265	.358	.642
Reduce el mal aliento	.590	.316	.336	.561
Deja la boca fresca	.511	.467	.325	.585
No pica	.216	.784	.139	.681
Tiene diferentes sabores	.207	.774	.158	.666
Tiene buen sabor	.289	.771	.096	.687
Tiene aroma agradable	.211	.694	.373	.665
Es fácil de encontrarla	.184	.172	.847	.781
Es una marca de confianza	.375	.178	.776	.774
La recomiendan los dentistas	.406	.270	.562	.553
Tiene efecto duradero	.439	.452	.455	.604
Valor característico/propio	4.18	3.22	2.55	
Porcentaje de la varianza explicada	27.9%	21.4%	17.0%	
Porcentaje acumulado de la varianza explicada	27.9%	49.2%	66.3%	

Método de extracción: Análisis de componentes principales

Una vez más utilizaremos nuestro ejemplo para demostrar este fácil proceso. Después de haber encontrado que los 15 reactivos pueden ser agrupados en tan sólo tres factores y de haber hecho la rotación correspondiente, vemos que nuestra matriz de componentes rotados es el cuadro 4.

Antes de entrar en la explicación de cómo se agrupan los reactivos, considero importante definir el concepto de comunalidades:

La comunalidad es la proporción de la variabilidad del enunciado que está explicada por el conjunto de los factores extraídos del análisis, que en este caso son tres. Tradicionalmente a las comunalidades se les pone la nomenclatura H2 y matemáticamente es:

$$H^2 = (f_1)^2 + (f_2)^2 + (f_3)^2 + \ldots + (f_n)^2$$

En donde las f representan la carga específica de cada factor para el reactivo "X". En nuestro ejemplo, la comunalidad del primer reactivo se calcularía de la siguiente forma:

$$H^2 = (.807)^2 + (.213)^2 + (.181)^2 = 0.729$$

En nuestro ejemplo de enjuagues bucales vemos que la comunalidad más alta se encuentra en el reactivo "es fácil de encontrarla", lo que nos quiere decir que de todas las variables, ésta es la que mejor o más explicada se encuentra por nuestros tres factores.

Como análisis adicional que sugiero sea hecho siempre de primera instancia, en la tabla encontramos que todos los reactivos se explican en más del 50% por los tres factores utilizados, y 9 de los 15 están explicados en más del 65%, lo cual nos da mucha tranquilidad en el sentido que el proceso matemático aplicado, hasta el momento, "ajusta" de manera correcta a nuestros datos. Deberíamos sentirnos preocupados cuando las comunalidades tengan valores bajos en un varios de los reactivos, ya que nos estaría diciendo que muchos de ellos los estamos explicando pobremente.

Los valores característicos o propios, como algunas personas los llaman, es la suma del cuadrado de las cargas factoriales para cada factor; es decir, el valor característico del factor 1 es la suma de:

$$V^2 = (.807)^2 + (.767)^2 + (.763)^2 \;\ldots + (.439)^2 = 4.18$$

Este número nos sirve simplemente para poder calcular el porcentaje de varianza explicada por cada factor, ya que éste resulta de dividir el valor característico entre el número total de reactivos. En el caso del primer factor, el porcentaje de varianza explicada por el mismo se obtiene:

$$4.18 / 15 = 0.2787 = 27.9\%$$

Esto quiere decir que del total de la varianza, el primer factor solo explica el 27.9% de la misma.

Por último, y bastante obvio, el *porcentaje acumulado de la varianza explicada* simplemente resulta de sumar el porcentaje de la varianza explicada de cada factor, obteniendo en nuestro ejemplo un 66.3%. Si ustedes revisan el cuadro de varianza total explicada podrán comprobar que el programa se detuvo en el factor 3 en donde este número es exactamente el mismo que obtuvimos a través de la fórmula anterior. Así las cosas, ya sabemos qué es la varianza explicada, su importancia en el análisis factorial y como se obtiene paso por paso.

Ya que hemos hecho todo lo anterior y revisado "el poder" de los factores encontrados a través de la varianza total explicada y las comunalidades, debemos decidir qué reactivos serán agrupados en el factor 1, cuáles en el 2 y cuáles en el 3. El proceso para esto es muy sencillo y consiste básicamente en destinar a cada factor los reactivos que muestren la mayor carga factorial. En el cuadro 5 podrán observar como quedarían agrupados los reactivos de nuestro ejemplo de enjuagues bucales.

Dice el dicho que si las cosas que valen la pena se hicieran fácilmente, cualquiera las haría, y en el caso del análisis factorial comúnmente debemos tomar decisiones apoyadas en nuestro criterio al momento de agrupar reactivos. Cuando la carga factorial es marcadamente alta el algún factor, el hecho de decidir poner el reactivo en ese factor es verdaderamente sencillo, tal es el caso de los reactivos 1, 2, 3, 4, 5 y 6 en el factor 1; pero existen algunas ocasiones en donde la carga factorial está muy dividida entre los factores encontrados, como en el caso de los reactivos 7 y 9, y es entonces cuando se nos presenta la disyuntiva de a qué factor lo cargamos.

La única solución a este problema es aplicar nuestro criterio como investigadores y decidir, de forma hasta cierto punto arbitraria, en qué factor debemos ponerlo. Para ello, la mejor herramienta al alcance es el sentido común, revisando con qué grupo de reactivos hace más lógica el unir aquellos que "están volando".

CUADRO 5

	Componente/factor 1	2	3	Comunalidades
Reduce la placa dentobacteriana	$\boxed{.807}$	.213	.181	.729
Elimina los gérmenes	$\boxed{.767}$	.199	.241	.686
Previene la caries	$\boxed{.763}$	.269	.184	.688
Previene la gingivitis	$\boxed{.743}$	.206	.219	.642
Ayuda a la higiene bucal	$\boxed{.666}$	.265	.358	.642
Reduce el mal aliento	$\boxed{.590}$	.316	.336	.561
Deja la boca fresca	$\boxed{.511}$	.467	.325	.585
No pica	.216	$\boxed{.784}$	.139	.681
Tiene diferentes sabores	.207	$\boxed{.774}$	.158	.666
Tiene buen sabor	.289	$\boxed{.771}$	.096	.687
Tiene aroma agradable	.211	$\boxed{.694}$	.373	.665
Es fácil de encontrarla	.184	.172	$\boxed{.847}$	.781
Es una marca de confianza	.375	.178	$\boxed{.776}$	.774
La recomiendan los dentistas	.406	.270	$\boxed{.562}$	.553
Tiene efecto duradero	.439	$\boxed{.452}$	.455	.604
Valor característico/propio	4.18	3.22	2.55	
Porcentaje de la varianza explicada	27.9%	21.4%	17.0%	
Porcentaje acumulado de la varianza explicada	27.9%	49.2%	66.3%	

En los ejemplos que he utilizado hasta este momento, todas las cargas factoriales han mostrado signo positivo, aunque, en muchas ocasiones, es perfectamente válido, y además común, tener valores negativos. Si este es el caso no hay nada de que preocuparse, ya que el criterio de inclusión que he descrito en este apartado se mantiene exactamente igual.

VII. Etiquetado de Factores

El último paso del análisis de factores es poner una etiqueta o nombre a cada factor. Resulta poco útil al momento de una presentación o análisis de resultados el estar hablando de factores simplemente a través de un número, ya que esto no nos dice nada, por lo que la solución es dar un nombre o etiqueta a cada factor, proceso también totalmente arbitrario que se encuentra a criterio del investigador. En el ejemplo que hemos estado siguiendo, podríamos poner las siguientes etiquetas:

Factor 1: Desempeño médico del producto
Factor 2: Desempeño cosmético del producto
Factor 3: Aspectos comerciales y soporte profesional.

VIII. Comentario Final

No me cansaré de repetir que lo hasta aquí expuesto es un análisis bastante somero y no derivativo del factorial, por lo que su claro y profundo entendimiento les ayudará a aplicarlo en ocasiones diversas a lo que en este documento he desarrollado.

Lo que sí nos debe dejar tranquilos es que una larga lista de reactivos puede ser disminuida significativamente, de tal forma que su análisis sea más ágil y objetivo.