IDICSO
Instituto de Investigación en Ciencias Sociales
Facultad de Ciencias Sociales
Universidad del Salvador

ÁREA EMPLEO Y POBLACIÓN

El análisis factorial

por **Horacio Chitarroni***

Buenos Aires, DIC/2002

* **CHITARRONI, Horacio.** Lic. en Sociología, Universidad Nacional de Buenos Aires (UBA). Docente, Facultad de Ciencias Sociales, Universidad del Salvador (USAL). Docente de la Maestría en Ciencias Sociales del Trabajo, Facultad de Ciencias Sociales, UBA. Investigador Principal, Área Empleo y Población, IDICSO, USAL. Consultor del Consejo Nacional de Coordinación de Políticas Sociales, SIEMPRO (Sistema de Evaluación, Seguimiento y Monitoreo de Programas Sociales).

BREVE HISTORIA DEL IDICSO. Los orígenes del IDICSO se remontan a 1970, cuando se crea el "Proyecto de Estudio sobre la Ciencia Latinoamericana (ECLA)" que, por una Resolución Rectoral (21/MAY/1973), adquiere rango de Instituto en 1973. Desde ese entonces y hasta 1981, se desarrolla una ininterrumpida labor de investigación, capacitación y asistencia técnica en la que se destacan: estudios acerca de la relación entre el sistema científico-tecnológico y el sector productivo, estudios acerca de la productividad de las organizaciones científicas y evaluación de proyectos, estudios sobre política y planificación científico tecnológica y estudios sobre innovación y cambio tecnológico en empresas. Las actividades de investigación en esta etapa se reflejan en la nómina de publicaciones de la "Serie ECLA" (SECLA). Este instituto pasa a depender orgánica y funcionalmente de la Facultad de Ciencias Sociales a partir del 19 de Noviembre de 1981, cambiando su denominación por la de Instituto de Investigación en Ciencias Sociales (IDICSO) el 28 de Junio de 1982.

Los fundamentos de la creación del IDICSO se encuentran en la necesidad de:

- Desarrollar la investigación pura y aplicada en Ciencias Sociales.
- Contribuir a través de la investigación científica al conocimiento y solución de los problemas de la sociedad contemporánea.
- Favorecer la labor interdisciplinaria en el campo de las Ciencias Sociales.
- Vincular efectivamente la actividad docente con la de investigación en el ámbito de la facultad, promoviendo la formación como investigadores, tanto de docentes como de alumnos.
- Realizar actividades de investigación aplicada y de asistencia técnica que permitan establecer lazos con la comunidad.

A partir de 1983 y hasta 1987 se desarrollan actividades de investigación y extensión en relación con la temática de la integración latinoamericana como consecuencia de la incorporación al IDICSO del Instituto de Hispanoamérica perteneciente a la Universidad del Salvador. Asimismo, en este período el IDICSO desarrolló una intensa labor en la docencia de post-grado, particularmente en los Doctorados en Ciencia Política y en Relaciones Internacionales que se dictan en la Facultad de Ciencias Sociales. Desde 1989 y hasta el año 2001, se suman investigaciones en otras áreas de la Sociología y la Ciencia Política que se reflejan en las series "Papeles" (SPI) e "Investigaciones" (SII) del IDICSO. Asimismo, se llevan a cabo actividades de asesoramiento y consultoría con organismos públicos y privados. Sumándose a partir del año 2003 la "Serie Documentos de Trabajo" (SDTI).

La investigación constituye un componente indispensable de la actividad universitaria. En la presente etapa, el IDICSO se propone no sólo continuar con las líneas de investigación existentes sino también incorporar otras con el propósito de dar cuenta de la diversidad disciplinaria, teórica y metodológica de la Facultad de Ciencias Sociales. En este sentido, las áreas de investigación del IDICSO constituyen ámbitos de articulación de la docencia y la investigación así como de realización de tesis de grado y post-grado. En su carácter de Instituto de Investigación de la Facultad de Ciencias Sociales de la Universidad del Salvador, el IDICSO atiende asimismo demandas institucionales de organismos públicos, privados y del tercer sector en proyectos de investigación y asistencia técnica.

IDICSO

Departamento de Comunicación

Email: idicso@yahoo.com.ar

Web Site: http://www.salvador.edu.ar/csoc/idicso

El análisis factorial

Bajo la denominación de análisis factorial se agrupan un conjunto de técnicas diferentes, entre las cuales figuran el análisis factorial clásico y el análisis de componentes principales[1].

La idea básica de esta técnica consiste en la búsqueda de factores o dimensiones latentes – no inmediatamente aprehensibles – que se supone que subyacen a un conjunto mayor de variables. Un ejemplo – vinculado, por lo demás, al origen del análisis factorial – contribuirá a una mejor comprensión de esta idea.

La historia

A comienzos del siglo XX, estadísticos como Pearson y Spearman se interesaron en los resultados que arrojaba la aplicación de los test destinados a la medición de la inteligencia. Ocurría que ciertos test que apuntaban a medir ciertas destrezas específicas (por ejemplo, destrezas matemáticas, destrezas lógicas, destrezas relacionadas con el uso del lenguaje) aparecían positivamente correlacionados entre sí: quienes obtenían puntuaciones elevadas en unos, también tendían a obtener altas calificaciones en los otros. Esta comprobación aparecía como auspiciosa en relación con la medición de la inteligencia: una explicación de estas coincidencias era que realmente existiera un factor subyacente que pudiera denominarse como tal. Las personas "inteligentes" eran aquellas que obtenían altas puntuaciones en todas estas áreas de desempeño. O, mejor dicho, si obtenían altas puntuaciones ello se debía a que poseían esta condición latente: eran inteligentes.

Ambos modos de ver la cuestión remiten a diferentes concepciones acerca del factor latente:

Una postura *realista* sostendría que la inteligencia existía, aunque no pudiera verse ni medirse directamente. Y ella era la causa que explicaba el desempeño diferencial en diversos test que medían destrezas hipotéticamente independientes entre sí (se partiría del supuesto de que no habría razones para que alguien que se desempeñaba bien en matemáticas también se mostrara muy apto en relación con la comunicación verbal, salvo que fuera "inteligente"). Vale decir, la inteligencia explicaría la varianza en común de los respectivos test.

Una postura *nominalista*, en cambio, afirmaría que la inteligencia era una mera abstracción: llamaríamos así a esa coincidencia de aptitudes. Denominaríamos "inteligencia" a la covarianza entre los distintos test.

Pero lo cierto es que, en cualquiera de las dos posturas (se le adjudicara o no "existencia real"), el factor inteligencia, en caso de poder medirse, serviría para

[1] Este último es el de uso más corriente.

explicar la varianza en común en los distintos test, así como para predecir el desempeño de las personas en cada una de las áreas.

Descubriendo dimensiones latentes

¿Cómo medir este factor latente?. Si existiera un factor responsable de la varianza en común de todos los test, este factor debiera ser una variable tal que, calculada la correlación parcial entre todos los test o variables manifiestas y mensurables (matemáticas, lógica, comunicación verbal), manteniendo controlado el factor, tal correlación se volviera igual a cero:

$$r_{mlc.\,f} = 0$$

Es decir, si encontráramos unas puntuaciones capaces de volver cero la correlación parcial entre los diversos test una vez que se controla dicha variable, habríamos dado con la inteligencia: esa variable la mediría. Al menos, esto creería un *realista*, en tanto que un *nominalista* sólo diría que hemos inventado una variable útil para medir las covarianzas entre los diferentes test y para predecir sus puntuaciones.

F

M L C

A poco, sin embargo, se descubrió que no existía necesariamente un solo y único factor que explicara toda la varianza conjunta de los test. En realidad, en relación con un conjunto *n* de variables (de test destinados a medir diferentes tipos de destrezas), habría un conjunto *m* (siendo $m < n$) de "factores". Algunos de estos factores estarían más vinculados a cierto grupo de test, mientras que otros se relacionarían más estrechamente con otro grupo. Por ejemplo, los test matemáticos (T_1, T_2 y T_3) estarían más relacionados con un factor F_1 que podríamos denominar "destreza matemática", que explicaría gran parte de la varianza de estos test. En tanto que un segundo factor F_2 aparecería más vinculado a los test de tipo lógico (T_4, T_5) y un tercer factor F_3 explicaría la varianza de los test orientados a las destrezas del lenguaje y la comunicación (T_6, T_7 y T_8). Esto no significa que el primer factor "matemático" F_1 no explicara, también, algo de la varianza de los test lógicos y aún de los relativos a la comunicación lingüística, pero seguramente lo haría en una proporción menor.

Esquemáticamente, sería así:

F_1 F_2 F_3

T_1 T_2 T_3 T_4 T_5 T_6 T_7 T_8

Dos objetivos del análisis factorial: parsimonia e interpretabilidad

Hacia mediados del siglo, Thurstone –que se embarcaba más bien en la postura *nominalista*– tendió a creer que estos factores no existían realmente, sino que eran abstracciones útiles, que permitían reducir un número elevado de variables manifiestas a un conjunto menor de factores latentes, cuyos puntajes, en última instancia, resultaban "inventados" mediante una serie de algoritmos. De hecho, los diversos procedimientos de "rotación" – cuyo concepto se examinará luego someramente – no serían sino maneras de "acomodar" estos factores, logrando de ellos el mejor rendimiento posible en términos de explicación de la varianza de las distintas variables, así como de nitidez en la asociación con cada subconjunto de ellas.

En última instancia, tal como lo sugiere el ejemplo hasta aquí utilizado de los test y los distintos tipos de destrezas, un análisis factorial debiera apuntar a dos objetivos:

- reducir un conjunto amplio de variables a un número menor (es un procedimiento de reducción dimensional, de simplificación). A esta simplificación se la denomina "parsimonia", que alude aquí a lograr expresar lo mismo con la mayor economía de medios, con cierta "elegancia": la misma diversidad de contenidos con un menor número de variables[2].

- lograr interpretabilidad o claridad: cada factor debería tener un significado identificable. Debiera expresar una dimensión separada, a la que pudiéramos designar con un nombre: por ejemplo, si F_1 se asociara muy claramente con T_1, T_2 y T_3 (y débilmente con los demás test) lo llamaríamos "Destreza matemática".

El análisis factorial clásico, partiendo de una postura más teórica o *realista*, apuntaría a buscar los "verdaderos" factores latentes, que explicaran la varianza en común que tienen las variables o test. En cambio, el análisis de componentes principales – de algún modo identificable con el *nominalismo* – procuraría encontrar unos factores capaces de reducir la cantidad de variables iniciales a un conjunto menor de factores que explicaran la mayor proporción posible de la varianza total de estas variables.

Una doble ecuación de regresión múltiple

Cada variable (en nuestro ejemplo, cada uno de los test: T_1, T_2, T_3...T_8) puede ser expresada como una combinación lineal de todos los factores (F_1, F_2 y F_3). Es decir, se puede escribir una ecuación de regresión múltiple donde el test T_1 sea la variable dependiente y cada uno de los factores (F_1, F_2 y F_3) sean las variables predictoras. Ocurrirá que cada uno de estos factores tendrá un cierto peso (o carga factorial o "saturación") en T_1. Y todos ellos explicarán una cierta parte de la varianza total de T_1, aunque seguramente no toda: quedará una parte de variabilidad única, individual o específica de la variable T_1, no asociada con ninguno de los tres factores, que se llama "singularidad":

[2] Así como uno diría que la teoría que más elegantemente da cuenta de uno o más fenómenos es la que lo hace con mayor simpleza. La más "económica".

$$T_1 = b_1*F_1 + b_2*F_2 + b_3*F_3 + e$$

Donde:

T_1: es la puntuación de un individuo cualquiera el test 1

F_1: es la puntuación del mismo individuo en el factor 1

b_1: es el coeficiente o carga factorial del factor 1 en el test 1

e: es el factor único o "singularidad" de la variable (en este caso, del test 1)

En el caso del análisis factorial, todas las variables (las puntuaciones de todos los test, en nuestro ejemplo) deben ser expresadas en una "medida común": esto es, en puntuaciones típicas Z[3]. De manera que los coeficientes b (las "cargas factoriales" o "factor loadings"[4]) nos dicen cuantos desvíos estándar aumenta (o disminuye, puesto que estos coeficientes pueden tener signo negativo) T_1 cada vez que el factor asociado (F_1 en el caso de b_1) aumenta en una unidad.

Obviamente, una ecuación similar podría escribirse para expresar los valores de los otros test: T_2, T_3....T_8. Pero – al menos en nuestro ejemplo – en el caso de T_1, T_2 y T_3 serán grandes las cargas factoriales de F_1 y pequeñas las dos restantes: esos test están muy saturados por el factor 1. En cambio, F_1 mostrará poca carga en T_4 y T_5 (en ellos pesará F_2) y también en T_6, T_7 y T_8 (donde tendrá una carga grande F_3).

Inversamente, cada uno de los factores (F_1, F_2 y F_3) puede ser expresado como una combinación lineal de los diversos test, mediante otra ecuación de regresión múltiple (como si se tratara de una suerte de "juego de espejos"):

$$F_1 = a_1*T_1 + a_2*T_2 + ...+ a_8*T_8$$

Donde:

F_1: es la puntuación (o "score" factorial) de un individuo cualquiera en el factor 1

T_1: es la puntuación de un individuo cualquiera en el test 1

a_1: es la ponderación natural del test 1 en el factor 1

Los puntajes o "scores" factoriales pueden ser entendidos como una suerte de índice autoponderado de las variables (en este caso, de los test).

Aquí no hay, claro está, "singularidad" del factor, puesto que el factor es un "invento" generado a partir de las variables. No tiene, pues, "vida propia": en

[3] Tal como se recordará, esta conversión consiste en expresar los valores de una variable en distancias a la media, medidas en desvíos estándar: z = (x – X) / s.
[4] También se les suele llamar saturaciones.

otros términos, toda la varianza de un factor es explicada por el conjunto de variables.

Comunalidades y singularidades

El coeficiente de correlación múltiple de cada test con el conjunto de los factores (usados como variables independientes o "explicadoras"), elevado al cuadrado, se denomina "comunalidad" de esa variable. Indica el grado en que su varianza es explicada por todos los factores a la vez (en este caso, se trataría de la varianza de T_1 explicada por F_1, F_2 y F_3):

$$C^2 = R^2_{T1.\ F1F2F3}$$

Mientras que la "singularidad" (o varianza única) de cualquiera de las variables sería:

$$U^2 = 1 - R^2$$

Esta es la proporción de varianza que es propia de la variable y no resulta explicada por el conjunto de los factores.

El valor propio o "eigenvalue"

La suma de las cargas factoriales o "factor loadings" (elevadas al cuadrado) de un mismo factor con todas las variables, se denomina valor propio, autovalor o "eigenvalue" de dicho factor: debe interpretarse como la varianza que es capaz de explicar ese factor de todas las variables en conjunto. Por ejemplo, el factor 1 explicaría una proporción alta de los test 1, 2 y 3 y – probablemente – pequeñas proporciones de las varianzas de los test restantes (4 a 8). La suma de todas esas proporciones sería el "valor propio" de ese factor.

$$EV_{F1} = b^2_{1T1} + b^2_{1T2} + ... b^2_{1t8}$$

En esta ecuación:

EV_{F1} = valor propio del factor 1

b^2_{1T1} = carga factorial del factor 1 en el test 1

b^2_{1T2} = carga factorial del factor 1 en el test 2

Etc.

Es evidente que cada factor sólo podría ser responsable, como máximo, del 100% de la varianza de cualquiera de las variables. En cuyo caso, $b^2 = 1$[5] (por ejemplo, si el factor 1 explicara toda la varianza del test 1, $b^2_{1T1} = 1$). Por lo tanto, el valor máximo posible del "eigenvalue" de un factor cualquiera es igual al

[5] En una ecuación de regresión con las variables expresadas en puntuaciones típicas Z, donde la función pasa por el origen, el coeficiente b es equivalente a r. De manera que b^2 equivale a la proporción de la varianza de una variable que resulta explicada por la otra.

número total de variables: si el factor 1 explicara el 100% de la varianza de cada uno de los ocho test, entonces sus cargas factoriales elevadas al cuadrado (b^2) serían iguales a uno con todos los test. Y su sumatoria sería ocho.

Y el promedio de estas cargas factoriales (es decir, el "eigenvalue" dividido por la cantidad de variables) no podría exceder de uno. Si calculamos este promedio (igual o menor a 1) y lo multiplicamos por 100, obtenemos la proporción de varianza de todas las variables que resulta explicada por el factor.

Así, si el factor 1 explicara el 70% de la varianza del test 1, el 60% de la varianza del test 2, el 80% de la varianza del test 3, el 10% de la varianza del test 4, el 15% de la varianza del test 5, el 5% de la varianza del test 6, el 3% de la varianza del test 7 y el 1% de la varianza del test 8, la suma de sus cargas factoriales con cada test, elevadas al cuadrado (es decir, su "eigenvalue"), sería:

$EV_{F1} = 0,70 + 0,60 + 0,80 + 0,10 + 0,15 + 0,05 + 0.03 + 0.01 = 2,44$

Y el promedio sería 2,44 / 8 = 0,305. Esto significa que el factor 1 explica el 30.5% de la varianza del conjunto de los test.

Obviamente, puesto que cada variable "se explica" a sí misma, puede decirse que cada una de ellas es responsable, en promedio, de 1/n de la varianza total conjunta. Vale decir, si tenemos ocho test, cada uno de ellos explicará (siempre en promedio) 1/8 = 0,125 de la varianza conjunta de todos ellos (es decir, 12,5%). Para que un factor sea útil (para que realmente reduzca ventajosamente el número de variables) tendría que aventajar ese porcentaje. De lo contrario, el análisis factorial no ofrecería ganancia alguna.

La tabla de "eigenvalue"

La tabla de "eigenvalues" nos permite seleccionar los factores que resultan adecuados. En el ejemplo de los test, con que venimos trabajando, supongamos que dicha tabla nos mostrara lo siguiente:

Total Variance Explained			
	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	2,4400	30,5000	30,5
2	2,3715	29,6438	60,1
3	2,3400	29,2500	89,4
4	0,4200	5,2500	94,6
5	0,2592	3,2400	97,9

6	0,0959	1,1983	99,1
7	0,0503	0,6288	99,7
8	0,0231	0,2891	100,0
Extraction Method: Principal Component Analysis.			

El método aquí aplicado es el de componentes principales: se trata de uno de los más usuales y consiste en la búsqueda, no de un único factor (la inteligencia) sino de las principales dimensiones subyacentes, que permiten reducir el número inicial de variables: los factores o componentes principales.

En este caso, vemos que los tres primeros componentes o factores son los que permiten "mejorar" el desempeño de las variables o test individuales. El "eigenvalue" de estos factores es superior a 1 y el promedio de estos valores propios, multiplicado por 100 nos dice cuánta varianza explica cada uno. Los tres primeros logran explicar casi 90% de la varianza de los test o variables individuales. Obviamente, el resto son desdeñables (aunque el sistema los va extrayendo a todos, hasta completar el 100% de la varianza explicada).

La matriz factorial

Ahora, para el caso de los tres factores o componentes principales que son de utilidad, veremos la matriz factorial. Esta matriz nos muestra el cruce de los factores (sólo los que resultan "buenos") con cada una de las variables o test individuales. Las celdas de la matriz muestran las "cargas factoriales" o "factor loadings" (aquellos coeficientes b de la ecuación inicial) de estos factores con cada test.

Component Matrix			
	Component		
Variables	1	2	3
T1	**0,837**	-0,254	-0,152
T2	**0,775**	0,352	0,215
T3	**0,894**	-0,297	-0,421
T4	-0,316	**0,854**	0,213
T5	0,387	**0,985**	0,112
T6	-0,224	0,352	**0,981**
T7	0,173	-0,419	**0,753**
T8	-0,100	0,186	**0,799**

Extraction Method: Principal Component Analysis.	
a	3 components extracted.

En este caso, puesto que se trata de un ejemplo imaginario, todo funciona en forma óptima. Hemos logrado extraer tres factores sobre ocho variables originales, con lo que logramos una apreciable reducción ("parsimonia"). Y estos factores se asocian, cada uno de ellos, inequívocamente con algunas de las variables originales y no con las demás: el factor o componente 1 se asocia con los test 1 a 3 (los de destreza matemática), mientras que el factor 2 lo hace con los test 4 y 5 (los de habilidades lógicas) y el tercer factor lo hace con los test 6, 7 y 8 (que miden habilidades lingüístico-comunicativas). Así, los factores son perfectamente "interpretables": podríamos, sin dificultad, llamar al primero "Inteligencia matemática", al segundo "Inteligencia Lógica" y al tercero "Inteligencia comunicativa".

La idea de la rotación

En la práctica, no suelen salir tan bien las cosas. Muchas veces, la matriz factorial muestra ambigüedades y no es fácil interpretar los factores. Supongamos que, en un ejemplo de cinco variables originales se obtuvieran dos factores o componentes principales y que la matriz resultante fuera la siguiente[6]:

Component Matrix		
	Component	
Variables	1	2
A	3	-4
B	2	-3
C	4	3
D	3	4
Extraction Method: Principal Component Analysis.		
A	2 components extracted.	

Aquí, la interpretación no es fácil: independientemente del signo de las cargas factoriales, el factor 1 aparece asociado a las variables A, C y D, mientras que el factor 2 lo está a las variables A, B, C y D. No hay, pues, una distinción nítida entre ambos.

[6] En este caso se adjudica a los loadings números enteros, para simplificar. En realidad, no pueden serlo, porque sus cuadrados no pueden superar la unidad: ningún factor puede explicar más del 100% de la varianza de una variable.

En estos casos, a veces, algún método de rotación mejora la situación. Se trata de un conjunto de algoritmos sumamente complejos, pero el esquema que sigue puede aclarar la idea central: en este esquema, los factores originales (F1 y F2) han sido representados sobre unos ejes. Y cada una de las variables (A, B, C, D) fue ubicada en el espacio determinado por estas coordenadas según los puntajes de las cargas o "loadings" con cada uno de los factores: por ejemplo, la variable A queda en la intersección del valor 3 del factor 1 y –4 del factor 2.

En un segundo momento, los ejes se rotan, como si se tratase de una tómbola que gira sobre el punto de intersección de dichos ejes: resultan de ello los factores rotados F1' y F2'. La posición de los puntos respecto de estos nuevos factores es diferente (vale decir, son diferentes los "loadings"): ahora la variable A queda situada en 1 del factor 1 y en –5 del factor 2.

Factores rotados:

La nueva matriz, con los nuevos componentes 1' y 2', resultantes de la rotación, tendría otras características:

Component Matrix		
	Component	
Variables	1'	2'
A	1	**-5**
B	0	**-4**
C	**5**	1
D	**5**	2
Extraction Method: Principal Component Analysis.		
A	2 components extracted.	

Ahora resultaría muy claro que el primer factor o componente satura fuertemente las variables C y D, mientras que el segundo tiene altas cargas en las variables A y B.

Al hacer la rotación, se recalculan las puntuaciones factoriales de cada sujeto para los nuevos factores rotados. Estas puntuaciones expresarían las dimensiones latentes. Por supuesto que, en la práctica, pocas veces una rotación resulta en un resultado tan exitoso en términos de "interpretabilidad"...

Métodos de rotación

Estas rotaciones de los factores pueden realizarse de más de un modo. Básicamente, puede emplearse una rotación ortogonal[7], en la que los ejes se rotan manteniendo un ángulo de 90° entre ellos, o bien una rotación oblicua[8], en cuyo caso los ejes de los factores rotados forman un ángulo menor. Las rotaciones oblicuas son menos usadas y algunos de los métodos existentes son el *Oblimin* y el *Promax*.

Entre las rotaciones ortogonales – como sería el caso del ejemplo expuesto en el esquema – existen dos métodos predominantes: el *Varimax* y el *Cuartimax*. El primero de ellos procura "simplificar" las columnas de la matriz factorial, logrando factores que tiendan a tener cargas máximas con algunas de las variables y muy bajas con las otras, en tanto que el segundo tiende a simplificar las filas de la matriz, de modo que cada una de las variables aparezca muy saturada por uno de los factores y presente cargas bajas en los otros. El procedimiento *Varimax* es uno de los más usados[9] y suele considerarse más robusto.

Component Matrix				
	Component			
Variables	1	2	3	
A	1	0	0	Cuartimax
B	0			
C	0			
D	0			
	Varimax			

¿Cuántos factores extraer?

Esta pregunta no tiene una respuesta única. Pero hay algunos criterios básicos: uno de ellos – ya visto[10] - consiste en extraer sólo aquellos factores cuyo autovalor supere 1, para que obtenga ventaja sobre las variables originales. Un criterio alternativo es que la varianza acumulada del conjunto de las variables explicada por los factores alcance a cierta proporción: por ejemplo, no menos de 60% del total.

[7] Ortogonal significa que los factores que se extraen no están correlacionados entre sí.
[8] La rotación oblicua genera factores correlacionados entre sí.
[9] Suele estar disponible en casi todos los soft informáticos de aplicación estadística.
[10] Y que es el adoptado por defecto en algunos soft estadísticos, como el SPSS.

Interpretación de los factores

¿Qué significa cada uno de los factores?. Lo fundamental, para interpretar los factores, es ver con cuáles de las variables se asocia, es decir, a cuáles satura. Esta información nos es proporcionada por las cargas factoriales o "loadings". El cuadrado de estas cargas factoriales debe interpretarse como la proporción de la varianza de la variable que resulta explicada por el factor (así, una carga factorial de 0,80 del factor 1 en la variable o test 1, implicaría un 64% de la varianza de esa variable explicada por el factor.

Esto sugiere, asimismo, que si una variable muestra una alta carga en todos los factores, oscurecerá la interpretación de los mismos. Es probable que convenga eliminarla del análisis.

Requisitos de la matriz de datos

¿Qué variables conviene incluir en un análisis factorial?. Puesto que los factores son, fundamentalmente, combinaciones lineales de las variables, es fácil comprender que las variables que se incluyen en el análisis deben ser las que muestran ciertas correlaciones con algunas de las otras. Un examen de la matriz de correlaciones bivariadas (r de Pearson) entre todas las variables involucradas nos permitirá una buena aproximación: aquellas variables que arrojen correlaciones muy débiles con todas las otras resultarán poco productivas, puesto que tienen escasa "comunalidad" y mucha "singularidad".

Lo anterior supone que si el examen de la matriz no revela un buen número de correlaciones relativamente altas entre las variables[11], seguramente no convendría emplear el análisis factorial. Asimismo, es deseable que exista multicolinealidad, vale decir cierta varianza común entre algunas de las variables, para que sea posible generar factores significativos. Una prueba para verificar la existencia de estas multicolinealidades es el cálculo de la correlación parcial para cada par de variables, manteniendo controladas todas las demás: si estas correlaciones parciales son similares a las originales, el análisis factorial no sería adecuado, por falta de multicolinealidad[12] .

Asimismo, el tamaño de la matriz debe ajustarse a ciertas exigencias mínimas. Es poco aconsejable calcular un análisis factorial con un número de casos inferior a 100 o, al menos, 50 casos. Y el cociente entre la cantidad de casos y la cantidad de variables no debiera ser menor a 10 (lo que demandaría un tamaño muestral de 100 casos para incluir 10 variables en el análisis factorial). A título ilustrativo, debe considerarse que si empleáramos 30 variables, el número de correlaciones bivariadas entre ellas sería:

30 * (30 – 1) / 2 = 435

[11] Como mínimo r > 0,30.
[12] Puesto que querría decir que las demás variables no intervienen en la relación entre el par original.

Con un nivel de significación de 0,05, aproximadamente 20 de ellas podrían ser casuales. Ello sugiere un tamaño muestral suficiente para poder descartar esa probabilidad con una significación más elevada. Según lo indican algunos cálculos, con un n = 100 podrían tomarse como estadísticamente significativas cargas factoriales de 0,55 y superiores, en tanto que con n = 50, estas cargas debieran no ser menores a 0,75[13].

La validación del análisis factorial

Acerca de la validez y fiabilidad de los factores extraidos, puede recurrirse a una prueba relativamente sencilla. Siempre que se cuente con un número suficiente de observaciones o casos, la muestra podría ser subdividida al azar en dos grupos, para practicar el análisis factorial, separadamente, en cada uno de ellos. Si los resultados en términos de las cargas de los factores extraídos son similares, ello aumentará considerablemente la confianza en la robustez de los resultados.

¿De qué sirven los factores?

¿De qué nos sirven los factores?. Como se ha dicho, los puntajes factoriales son una suerte de índices ponderados de las variables saturadas por ellos. Por una cuestión de economía y comodidad, suele ser mejor emplear, en lugar de una gran cantidad de indicadores empíricos, un número menor de dimensiones subyacentes a ellos.

Pero además, si realmente los factores resultan interpretables, el análisis factorial puede ser un elemento auxiliar valioso para el desarrollo teórico: al descubrirnos una dimensión subyacente a un conjunto de variables mensurables, nos forzaría a hacerla explícita.

Supongamos que nuestros casos son países y que contamos con información referida a un conjunto de variables sociales, económicas, etc. Si un análisis factorial nos permitiera extraer dos componentes principales y comprobáramos que los puntajes del primero se asocian fuertemente al ingreso per cápita, a las tasas de inversión y al crecimiento económico, mientras que las puntuaciones del segundo se correlacionan con las tasas de alfabetización, la esperanza de vida, la mortalidad infantil, etc., diríamos que hemos "descubierto" dos modalidades de *desarrollo*: *desarrollo económico* y *desarrollo humano*, respectivamente...

En cambio, supongamos que la distribución de un fondo de desarrollo educativo requiere de un criterio de asignación por provincias, basado en el grado de necesidad de cada una de ellas respecto de la educación. Estas carencias estarán expresadas por indicadores tan diversos como la cantidad de chicos en edad escolar, la proporción de población pobre, las tasas de repitencia y sobreedad, los puntajes promedio en las pruebas anuales que evalúan el rendimiento educativo, la cantidad de alumnos por docente, la

[13] Hair, Joseph y otros, *Análisis Multivariante*, Prentice Hall, Madrid, 1999.

cantidad de vacantes disponibles, etc. Aquí, un análisis factorial pondría el énfasis en hallar un solo factor latente que explica la máxima proporción posible de la varianza de todas estas variables. Este factor sería una suerte de índice ponderado de carencia educativa, que permitiría ordenar a las jurisdicciones para la asignación de los subsidios.

En este mismo ejemplo, podría suceder que prefiriéramos construir un índice o escala aditiva con las variables mencionadas. En este caso, necesitaríamos saber cuáles de ellas debieran incluirse en ese indicador compuesto, puesto que presentan cierta unidimensionalidad (miden la dimensión latente "carencia educativa"). El análisis factorial nos permitiría seleccionar adecuadamente estas variables.

En un terreno más empírico, los puntajes factoriales pueden ser empleados como variables dependientes o independientes en una gran diversidad de análisis estadísticos.

Buenos Aires, DIC/2002

por **Horacio Chitarroni**
Investigador Principal, Área Empleo y Población, IDICSO, USAL.
Email: hchitarroni@siempro.gov.ar