

# ANÁLISIS MULTIVARIADO

Un manual para investigadores



K. 6564551

BIBLIOTECA NUEVA UNIVERSIDAD  
MANUALES Y OBRAS DE REFERENCIA

SERIE PSICOLOGÍA  
Dirigida por Gualberto Buela-Casal

Andrés Catena  
Manuel M. Ramos  
Humberto M. Trujillo

## ANÁLISIS MULTIVARIADO

Un manual para investigadores

BIBLIOTECA NUEVA



Cubierta: A. Imbert

## ÍNDICE

© Los autores, 2003  
© Editorial Biblioteca Nueva, S. L., Madrid, 2003  
Almagro, 38  
28010 Madrid

ISBN: 84-9742-115-9  
Depósito Legal: M-231-2003

Impreso en Rógar, S. A.  
Impreso en España - *Printed in Spain*

Queda prohibida, salvo excepción prevista en la ley, cualquier forma de reproducción, distribución, comunicación pública y transformación de esta obra sin contar con la autorización de los titulares de propiedad intelectual. La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (arts. 270 y sigs., Código Penal). El Centro Español de Derechos Reprográficos ([www.cedro.org](http://www.cedro.org)) vela por el respeto de los citados derechos.

PRÓLOGO, por Agustín Escolano Benito .....	17
CAPÍTULO PRIMERO.—INTRODUCCIÓN AL ANÁLISIS MULTIVARIADO .....	
1. ¿Para qué sirve la estadística multivariada? .....	19
2. ¿Cuándo pueden aplicarse técnicas multivariadas? .....	22
2.1. Investigación en contextos no experimentales .....	23
2.2. Investigación en contextos experimentales .....	25
3. Clasificación de las técnicas multivariadas .....	26
3.1. Técnicas explicativas .....	26
3.1.1. Análisis de regresión múltiple .....	27
3.1.2. Análisis discriminante .....	27
3.1.3. Análisis multivariado de varianza y covarianza .....	28
3.1.4. Análisis de regresión logística .....	28
3.1.5. Ecuaciones lineales estructurales .....	29
3.2. Técnicas descriptivas .....	29
3.2.1. Análisis de datos categóricos .....	30
3.2.2. Análisis de correlación canónica .....	30
3.2.3. Análisis de <i>cluster</i> .....	31
3.2.4. Análisis factorial y de componentes principales .....	31
3.2.5. Análisis de supervivencia .....	32
3.3. ¿Cómo seleccionar la técnica de análisis adecuada? .....	32
4. Conceptos básicos .....	34
4.1. Definiciones útiles .....	34
4.1.1. Escalas de medida .....	34
4.1.2. Muestras y poblaciones .....	35
4.1.3. Vectores y matrices .....	35
4.1.4. Mezcla de variables .....	36
4.2. Contraste de hipótesis univariado y multivariado .....	37
4.3. Matrices .....	38
4.3.1. Operaciones elementales con matrices .....	39
4.3.2. Matrices de sumas de cuadrados y productos cruzados ..	40
4.3.3. Matrices de covarianzas y correlaciones .....	41
5. Organización de ficheros de datos en paquetes estadísticos .....	42
6. Lecturas recomendadas .....	43
7. Ejercicios .....	44

CAPÍTULO II.—ANÁLISIS UNIVARIADO DE VARIANZA .....	47
1. La inferencia estadística y el contraste de hipótesis .....	48
2. Diseños experimentales univariados .....	52
2.1. Diseños unifactoriales univariados .....	53
2.1.1. Diseño completamente aleatorizado .....	53
2.1.1.1. Contrastes planeados .....	56
2.1.1.2. Contrastes no planeados .....	59
2.1.1.3. Magnitud de los efectos .....	60
2.1.1.4. Potencia del contraste .....	61
2.1.1.5. Supuestos del análisis .....	62
2.1.1.5.1. Normalidad .....	62
2.1.1.5.2. Homogeneidad de las varianzas .....	62
2.1.1.5.3. Independencia de los errores .....	63
2.1.1.5.4. Aditividad .....	63
2.1.1.6. Análisis mediante SPSS .....	63
2.1.2. Diseños de medidas repetidas de un solo factor .....	66
2.2. Diseños factoriales univariados .....	74
2.2.1. Análisis global de varianza de los diseños de dos factores .....	77
2.2.2. Análisis de contrastes .....	83
2.3. Introducción a los diseños con covariantes .....	86
3. La secuencia de investigación orientada al análisis mediante ANOVA ..	88
4. Lecturas recomendadas .....	91
5. Ejercicios .....	92
 CAPÍTULO III.—ANÁLISIS DE DATOS CATEGÓRICOS .....	97
1. Conceptos fundamentales del análisis de datos categóricos .....	97
2. Las preguntas que el análisis de frecuencias ayuda a resolver .....	99
2.1. Parámetros .....	100
2.2. Importación de los niveles .....	100
2.3. Magnitud de la relación .....	100
2.4. Influencia de variables independientes sobre una variable dependiente .....	100
3. El análisis de tablas de frecuencia: Modelos logarítmico-lineales .....	101
3.1. Comprobación de efectos .....	102
3.2. Construcción del modelo de predicción .....	105
3.2.1. Tipos de modelos .....	106
3.2.1.1. Número de efectos .....	106
3.2.1.2. Estructura .....	106
4. Limitaciones y supuestos de los modelos log-lineales .....	116
4.1. Independencia de las observaciones .....	117
4.2. Número de variables de sujetos .....	117
5. La secuencia del análisis log-lineal .....	117
6. Lecturas recomendadas .....	119
7. Ejercicios .....	119

CAPÍTULO IV.—ANÁLISIS FACTORIAL Y DE COMPONENTES PRINCIPALES .....	121
1. Conceptos fundamentales del análisis factorial y de componentes principales .....	121
2. Los tipos de preguntas que resuelve el análisis factorial .....	128
2.1. Número y naturaleza de los factores .....	128
2.2. Importancia de los factores .....	128
2.3. Contraste de modelos .....	129
2.4. Estimación de las puntuaciones factoriales .....	129
3. El análisis factorial .....	129
3.1. Matriz de correlaciones .....	129
3.2. Extracción de factores .....	132
3.2.1. Análisis factorial por componentes principales .....	132
3.2.2. Número de componentes que se pueden extraer .....	137
3.2.3. Comunalidades .....	139
3.2.4. Rotación de factores .....	140
3.2.4.1. Rotación ortogonal .....	141
3.2.4.1.1. Rotación Varimax .....	141
3.2.4.1.2. Rotación Cuartimax .....	143
3.2.4.1.3. Rotación Ecuamax .....	143
3.2.4.2. Rotación oblicua .....	143
3.2.4.3. ¿Qué tipo de rotación aplicar? .....	145
3.2.5. El análisis factorial mediante SPSS .....	145
3.3. Puntuaciones factoriales .....	148
3.4. Otros métodos de extracción de factores .....	150
3.4.1. Método de los factores principales .....	150
3.4.2. Métodos de mínimos cuadrados .....	152
3.4.3. Método de máxima verosimilitud .....	152
3.4.4. Método alfa .....	153
3.4.5. Método de Imagen .....	153
4. Limitaciones y supuestos del análisis .....	153
4.1. Tamaño de la muestra y datos perdidos .....	154
4.2. Normalidad .....	154
4.3. Linealidad .....	154
4.4. Ausencia de puntos extremos .....	154
4.5. Multicolinealidad .....	155
4.6. Factorizabilidad de la matriz de correlaciones .....	155
5. La secuencia del análisis factorial .....	155
6. Lecturas recomendadas .....	158
7. Ejercicios .....	159
 CAPÍTULO V.—ANÁLISIS DE «CLUSTER» .....	163
1. Objetivos del análisis de <i>cluster</i> .....	163
2. Clases de cuestiones que el análisis de <i>cluster</i> ayuda a resolver .....	164
3. Medidas de semejanza .....	165
3.1. Medidas de distancia .....	166
3.2. Medidas de igualación .....	169
3.2.1. Medidas de igualación para variables dicotómicas .....	170
3.2.2. Medidas de igualación para frecuencias .....	171

	Índice
4. Métodos de agrupamiento .....	172
4.1. Métodos jerárquicos .....	172
4.1.1. Métodos jerárquicos de aglomeración .....	172
4.1.1.1. Método del centroide .....	173
4.1.1.2. Método de la vinculación promedio .....	174
4.1.1.3. Método del vecino más próximo .....	175
4.1.1.4. Método del vecino más lejano .....	176
4.1.1.5. Método de Ward .....	176
4.1.2. Agrupamiento jerárquico mediante SPSS .....	177
4.1.3. Evaluación de la solución de agrupamiento .....	179
4.1.3.1. Raíz cuadrática media de las desviaciones típicas (RMSSTD) .....	179
4.1.3.2. R cuadrado .....	180
4.1.3.3. R <sup>2</sup> semiparcial .....	181
4.1.3.4. Distancia entre <i>clusters</i> .....	182
4.1.4. Métodos de división .....	182
4.1.4.1. Distancia astilla-promedio .....	183
4.1.4.2. Detección automática de interacción .....	183
4.2. Métodos de partición .....	183
4.3. ¿Qué método de agrupamiento emplear? .....	186
4.3.1. Métodos jerárquicos .....	186
4.3.2. Métodos de partición .....	187
5. Limitaciones .....	188
6. La secuencia del análisis de <i>cluster</i> .....	189
7. Lecturas recomendadas .....	190
8. Ejercicios .....	191
<b>CAPÍTULO VI.—ANÁLISIS DE CORRELACIÓN CANÓNICA .....</b>	<b>193</b>
1. Conceptos fundamentales en el análisis de correlación canónica .....	193
2. Las cuestiones que el análisis de correlación canónica ayuda a resolver ..	195
2.1. Número de variados canónicos .....	196
2.2. Interpretación de la correlación canónica .....	196
2.3. Puntuaciones en los variados .....	197
3. El análisis de correlación canónica .....	197
4. Limitaciones y supuestos del análisis de correlación canónica .....	203
4.1. Linealidad .....	204
4.2. Multicolinealidad y singularidad .....	204
4.3. Normalidad multivariada .....	204
4.4. Ausencia de puntos extremos .....	204
5. La secuencia de investigación en el análisis de correlación canónica ..	205
6. Lecturas recomendadas .....	206
7. Ejercicios .....	207
<b>CAPÍTULO VII.—ANÁLISIS DE SUPERVIVENCIA .....</b>	<b>209</b>
1. Conceptos fundamentales del análisis de supervivencia .....	209
2. Análisis de supervivencia .....	210
2.1. Tablas de supervivencia .....	211
2.2. Kaplan-Meier .....	213

	Índice
2.3. Regresión de Cox .....	214
3. Limitaciones y supuestos .....	215
4. La secuencia de investigación orientada al análisis de supervivencias ..	215
5. Lecturas recomendadas .....	217
6. Ejercicios .....	217
<b>CAPÍTULO VIII.—ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE .....</b>	<b>223</b>
1. Aproximación conceptual al análisis de regresión múltiple .....	224
2. Regresión lineal simple .....	225
3. Regresión lineal múltiple .....	226
3.1. Contraste de hipótesis en la regresión múltiple .....	229
3.2. Variedades del análisis de regresión múltiple .....	231
3.2.1. Regresión simultánea .....	232
3.2.2. Regresión múltiple secuencial .....	232
3.2.3. Regresión múltiple por etapas ( <i>Stepwise</i> ) .....	234
3.2.3.1. Regresión por etapas hacia delante .....	235
3.2.3.2. Regresión por etapas hacia atrás .....	236
3.2.3.3. Regresión por etapas .....	237
3.2.3.4. ¿Qué método seleccionar? .....	238
3.3. Interpretación de los resultados del análisis de regresión .....	238
3.3.1. Método simultáneo .....	239
3.3.2. Métodos secuenciales .....	240
3.4. El análisis de varianza como análisis de regresión .....	241
4. Limitaciones y condiciones de aplicación .....	244
4.1. Linealidad .....	245
4.2. Homogeneidad del error de predicción .....	246
4.3. Independencia de los errores .....	248
4.4. Normalidad de las puntuaciones error .....	249
4.5. Puntos extremos .....	250
4.6. Multicolinealidad .....	251
4.7. Número de sujetos .....	251
5. La secuencia de investigación orientada al análisis mediante regresión múltiple .....	252
6. Lecturas recomendadas .....	253
7. Ejercicios .....	254
<b>CAPÍTULO IX. ANÁLISIS MULTIVARIADO DE VARIANZA (MANOVA) Y COVARIANZA (MANCOVA) .....</b>	<b>255</b>
1. Análisis univariado (ANOVA) frente a multivariado (MANOVA) ..	256
1.1. Bases conceptuales de MANOVA .....	256
1.2. El contraste de hipótesis en MANOVA/MANCOVA .....	261
2. El análisis de datos mediante MANOVA/MANCOVA .....	262
2.1. El análisis de diseños con un factor manipulado entre grupos ..	262
2.1.1. Partición de variabilidad .....	266
2.1.2. La lógica del contraste de hipótesis en MANOVA: variados, valores propios y coeficientes de los variados ..	267
2.1.3. Estadísticos de contraste .....	271
2.1.4. Potencia .....	273

	Índice
2.1.5. Magnitud del efecto .....	276
2.1.6. Qué estadístico seleccionar .....	276
2.1.7. Contrastes Multivariados .....	277
2.2. MANOVA con más de un factor entre grupos .....	280
2.3. MANOVA para diseños de medidas repetidas .....	284
2.3.1. Diseños de medidas repetidas con un solo factor .....	284
2.3.2. Diseños de medidas repetidas con dos factores .....	287
2.3.3. Diseños factoriales mixtos o parcialmente repetidos ..	289
2.4. Diseños que incluyen covariados .....	290
3. Limitaciones y supuestos del análisis multivariado de varianza y covarianza .....	294
3.1. Independencia .....	294
3.2. Homogeneidad de las matrices de varianzas-covarianzas .....	295
3.3. Normalidad multivariada .....	296
3.4. Linealidad .....	296
3.5. Multicolinealidad .....	297
3.6. Homogeneidad de los coeficientes de regresión .....	298
3.7. Observaciones extremas .....	298
4. La secuencia de investigación orientada al análisis mediante MANOVA/ MANCOVA .....	299
5. Lecturas recomendadas .....	301
6. Ejercicios .....	302
 CAPÍTULO X.—ANÁLISIS DISCRIMINANTE .....	 303
1. Conceptos fundamentales del análisis discriminante .....	303
2. Preguntas que el análisis discriminante puede ayudar a resolver .....	305
2.1. Cuestiones que ayuda a responder el análisis discriminante descriptivo .....	306
2.2. Cuestiones que ayuda a resolver el análisis discriminante predictivo .....	308
3. Análisis discriminante descriptivo .....	309
3.1. Determinación de las diferencias significativas entre los grupos ..	309
3.2. Funciones discriminantes lineales .....	310
3.3. La importancia de cada variable predictor: la matriz de estructura .....	315
3.4. Análisis discriminante por etapas .....	317
4. Análisis discriminante predictivo .....	319
4.1. Obtención de los coeficientes de clasificación lineal .....	319
4.2. Clasificación de individuos .....	322
5. Limitaciones y supuestos del análisis discriminante .....	325
5.1. Normalidad multivariada .....	326
5.2. Homogeneidad de varianzas-covarianzas .....	326
5.3. Linealidad y multicolinealidad .....	326
5.4. Muestras desiguales .....	326
5.5. Puntos extremos .....	327
6. La secuencia de investigación orientada al análisis discriminante .....	327
7. Lecturas recomendadas .....	329
8. Ejercicios .....	329

	Índice
CAPÍTULO XI.—ANÁLISIS DE REGRESIÓN LOGÍSTICA .....	331
1. Regresión logística: regresión con una variable dependiente no métrica .....	331
1.1. Las ecuaciones básicas de la regresión logística .....	332
1.2. Utilidad de la regresión logística .....	334
2. Fundamentos del análisis de regresión logística .....	336
3. Tipos de regresión logística .....	343
3.1. Regresión logística simultánea .....	343
3.2. Regresión logística secuencial .....	343
3.2.1. Regresión logística serial .....	343
3.2.2. Regresión logística secuencial .....	345
4. La clasificación de individuos .....	346
5. El análisis de regresión polítomico o multinomial .....	347
6. Limitaciones y supuestos .....	348
6.1. Linealidad de la función logit .....	348
6.2. Independencia de los errores .....	349
6.3. Multicolinealidad .....	349
6.4. Número de variables y número de sujetos .....	349
6.5. Puntos extremos .....	350
7. La secuencia de investigación orientada al análisis de regresión logística .....	350
8. Lecturas recomendadas .....	352
9. Ejercicios .....	352
 CAPÍTULO XII.—MODELOS DE ECUACIONES ESTRUCTURALES .....	 355
1. Conceptos básicos del análisis factorial confirmatorio y los modelos de ecuaciones estructurales .....	355
2. Las clases de cuestiones que SEM ayuda a resolver .....	356
2.1. Comprobación de teorías .....	357
2.2. Cantidad de varianza explicada .....	357
2.3. Mediación .....	357
2.4. Diferencias entre grupos e intrasujetos .....	358
3. Construcción de modelos y ecuaciones básicas .....	358
3.1. El modelo general de ecuaciones estructurales .....	361
3.2. Estimación de parámetros .....	366
3.3. Identificación del modelo .....	366
3.4. Los efectos directos, indirectos y totales .....	367
4. Ajuste del modelo .....	367
4.1. Valoración mediante chi-cuadrado .....	368
4.2. Valoración mediante medidas heurísticas .....	368
4.2.1. Índice de bondad de ajuste (GFI) .....	368
4.2.2. Índice ajustado de bondad de ajuste (AGFI) .....	369
4.2.3. Raíz de la media de cuadrados residual .....	370
4.2.4. Índices de no centralidad .....	370
4.2.5. ¿Por qué no ajusta un modelo? .....	371
4.2.6. Evaluación de los parámetros del modelo .....	372
4.3. Reespecificación del modelo .....	373
5. Aplicación .....	375

	Índice
6. Limitaciones y supuestos .....	379
6.1. Normalidad multivariada .....	379
6.2. Linealidad .....	379
6.3. Ausencia de multicolinealidad y singularidad .....	380
6.4. Indicadores múltiples .....	380
6.5. Modelos infraidentificados y justamente identificados .....	380
6.6. Tamaño de la muestra .....	381
7. La secuencia de investigación en CFA y SEM .....	381
8. Lecturas recomendadas .....	383
9. Ejercicios .....	384
 BIBLIOGRAFÍA .....	 387
 GLOSARIO .....	 395
 ÍNDICE DE MATERIAS .....	 409

## Prólogo

El lector tiene en sus manos un Manual de Análisis Multivariado elaborado desde la experiencia adquirida a lo largo de años de uso de las técnicas que se describen, pero también desde la experiencia que producen los años de enseñanza en cursos de Licenciatura y de Doctorado de esas técnicas. Como investigadores, la preocupación fundamental de los autores es, como la de tantos otros colegas, emplear las técnicas de una manera instrumental. En ese sentido, la idea más importante que late en todos los capítulos del Manual es la de que al investigador no le interesan tanto los entresijos matemáticos de las técnicas, a menudo más encubridores de su valía y posibilidades que clarificadores de su interés y uso potencial, cuanto en qué contextos de investigación pueden emplearse, qué clases de preguntas o problemas de investigación ayudan a resolver, cuáles son sus limitaciones de uso y cómo pueden interpretarse sus resultados. Sin embargo, en cuanto que docentes, el interés de los autores reside justamente en hacer comprender a los estudiantes los aspectos que para el investigador son menos relevantes, los relativos a las derivaciones matemáticas, los algoritmos de cálculo, las ecuaciones fundamentales. El presente Manual puede entenderse como un intento de integrar ambos puntos de vista. De ahí que se proporcionen las ecuaciones fundamentales para comprender los aspectos más importantes de las técnicas, pero también una guía de cuándo la técnica es aplicable y cuáles son los pasos fundamentales que es preciso seguir para que su uso e interpretación sean correctos.

Lograr ese objetivo fundamental es, pensamos, relativamente fácil con la ayuda de los paquetes estadísticos que tan disponibles están en el mercado actual. Sin duda uno de los paquetes que más amplia aceptación tiene es SPSS, un paquete cuyo interfaz de usuario ha mejorado considerablemente en las últimas versiones. Sus ventajas fundamentales son incuestionables: facilidad de manejo, marcos y ventanas cuya organización facilita la comprensión del análisis y su gran abanico de técnicas estadísticas. En la mayoría de los

temas la exposición viene acompañada por los comandos de SPSS que son necesarios para lograr los objetivos de análisis, acompañados por la salida que ofrece.

Cada capítulo tiene una organización bien definida. En primer lugar se presentan los objetivos de aprendizaje. Estos objetivos definen también el contenido básico del capítulo, de modo que pueden emplearse como núcleos en torno a los cuales agrupar todos los aspectos del mismo. Desde el punto de vista del uso instrumental, lo más importante es qué problemas son los que la técnica abordada en el capítulo resuelve. La presentación de las mismas se realiza en gran medida a través de ejemplos cuya intención es ayudar al lector a analogizar respecto de su propia investigación. A continuación se aborda la técnica en sí. En este caso se presentan las ecuaciones fundamentales de la técnica y su aplicación a un caso concreto, que se irá desarrollando a lo largo del resto del capítulo. La intención es que el lector pueda comprobar cómo se realiza el análisis completo de una investigación mediante la técnica analítica en cuestión. Los resultados de los análisis se presentan en cuadros de texto, lo que permite una lectura alternativa de cada capítulo: datos del ejemplo, comandos para su análisis en SPSS y resultados. Lógicamente la técnica de análisis no puede aplicarse siempre de manera automática. En general, las técnicas funcionan de manera adecuada cuando se cumplen los supuestos en los que se basan, y se han descartado las limitaciones que pueden hacer sus resultados no completamente adecuados, o incluso espurios. Finalmente, se presenta la secuencia de análisis característica de la técnica. La intención es proporcionar una guía rápida de cómo debe procederse en términos ideales para que el análisis se realice con garantías suficientes, para que la interpretación de los datos sea lo más inequívoca posible.

Finalmente, con respecto a las referencias, la pretensión no ha sido la exhaustividad, sino la utilidad. En opinión de los autores, era preferible presentar al lector textos adicionales que pudieran ser adecuados para completar la información que se presenta en cada capítulo. Por este motivo, las lecturas se han organizado en dos apartados, uno en el que incluimos lecturas básicas, en las cuales se presenta un contenido general y de fácil acceso para quienes sean relativamente novedosos en el tema, y otro de mayor nivel técnico, en el que se puede encontrar información extensa correspondiente a los diferentes epígrafes de cada tópico.

En suma, los autores han pretendido que el presente Manual de Análisis Multivariado pueda convertirse en una guía de investigadores y en un material de apoyo para quienes, como nosotros, intentan contribuir a la formación metodológica adecuada de los investigadores del futuro. En este sentido han seleccionado las técnicas que más frecuentemente están siendo utilizadas en la investigación psicológica en la actualidad (por ejemplo, MANOVA), o que tienen un mayor potencial de aplicación futura (por ejemplo, Modelos de Ecuaciones Estructurales).

## CAPÍTULO PRIMERO

### Introducción al análisis multivariado

#### OBJETIVOS DE APRENDIZAJE

- 1) Discriminar cuándo es aconsejable una aproximación multivariada en lugar de una univariada.
- 2) Conocer en qué contextos de investigación son relevantes las técnicas multivariadas.
- 3) Conocer los criterios de clasificación de las técnicas multivariadas.
- 4) Conocer en qué se diferencia un contraste multivariado de uno univariado.
- 5) Computar matrices de sumas de cuadrados, covarianzas y correlaciones.
- 6) Conocer los principios fundamentales de la construcción de ficheros para los ajustes estadísticos.

En este capítulo presentaremos los conceptos fundamentales de las técnicas multivariadas, su definición, sus clases y el proceso de decisión que conduce a la elección de la técnica de análisis apropiada para la investigación que se ha realizado. Adicionalmente presentaremos algunos conceptos básicos necesarios para entender los capítulos posteriores. La presentación es intuitiva, atendiendo más al objetivo de comprensión que a la técnica en sí.

#### 1. PARA QUÉ SIRVE LA ESTADÍSTICA MULTIVARIADA

La Psicología es una de las disciplinas que más ha contribuido al desarrollo y aplicación de la estadística moderna. A su vez, la estadística le ha proporcionado a la Psicología la posibilidad de ser una ciencia basada en algo más con-

sistente que la pura observación de fenómenos y la especulación sobre las causas de los mismos. Este fausto maridaje ha producido, creemos, consecuencias beneficiosas para ambos. Sin embargo, aunque actualmente está aumentando considerablemente el número de investigadores y profesionales que emplean técnicas de análisis más «complejas», es bien conocido que durante un buen número de años, y aun hoy todavía, en la gran mayoría de las publicaciones en revistas psicológicas hay un predominio de un conjunto de técnicas de análisis que suelen etiquetarse como «univariadas». Su nombre, que tiene resonancias de simplicidad, se deriva de un hecho elemental, a saber, el número de variables que se miden en una investigación.

Es bien conocido que, atendiendo a su función dentro de la investigación (su función metodológica), las variables pueden clasificarse simplemente en independientes y dependientes (por supuesto hay también variables relevantes o extrañas, pero éstas no hacen al caso por ahora en este contexto). Las independientes son las que se *manipulan* con la *intención* de producir un efecto. Por ejemplo, los tratamientos que se aplican para paliar un problema (pongamos terapia psicológica, frente a terapia psicológica mezclada con medicamentosa); o la edad (joven, mayor) cuando se está interesado en procesos de envejecimiento; o la distribución de la práctica (digamos masiva o espaciada) cuando nuestro interés son los procesos de aprendizaje. Estas variables son características de los sujetos (como la edad) o de la situación (la práctica, la terapia) que suponemos van a afectar a otras variables, las dependientes. Las variables dependientes son las que medimos porque consideramos que van a reflejar el efecto de nuestra manipulación. Podemos medir, por ejemplo, el grado de ansiedad preguntando a los sujetos cómo se sienten, pero también podemos medirla mediante registros electrofisiológicos. Así, pues, en este esquema tan simple, las variables independientes son causas potenciales de los cambios observados en las variables dependientes (los efectos). Naturalmente, cuando se conoce la causa de un fenómeno, se tiene una explicación del mismo. Esto ha hecho que se apliquen una gran variedad de nombres: estímulo-respuesta, predictor-criterio, entrada-salida, etc., respectivamente para independiente-dependiente. Sin embargo, no en todas las investigaciones el interés estriba en encontrar una relación causal entre variables. En muchas ocasiones se trata simplemente de encontrar si hay algún tipo de relación entre unas variables y otras, pero todas tienen el mismo estatuto. En este caso es preferible usar los términos de predictor o criterio.

La distinción entre técnicas univariadas y multivariadas es una consecuencia de la distinción anterior. Es obvio que en muchas investigaciones tenemos más de una variable independiente. De hecho podríamos decir que actualmente es lo más frecuente. Estos estudios son conocidos como factoriales o multifactoriales. Nótese que un factor es una variable independiente. Sin embargo, también debe ser obvio que en muchos estudios no se mide una sola variable dependiente. Esto es verdad incluso cuando la apariencia no está tan clara. Por ejemplo, los ítem de una encuesta pueden ser cada uno de ellos una

variable dependiente diferente. Podemos sumar los ítem y obtener una sola puntuación, pero ese variado (la suma aritmética de los ítem) que obtenemos es simplemente la mezcla lineal de las variables medidas. Pues bien, las técnicas se clasifican en univariadas, bivariadas o multivariadas en función del número de variables dependientes que consideren, cualquiera que sea el número de independientes que se hayan manipulado.

Las técnicas univariadas son las que se aplican para el análisis de investigaciones en las que sólo se tiene una variable dependiente. Por ejemplo, el rendimiento académico, medido mediante las calificaciones (la dependiente), puede ser una función del estatus socioeconómico y de la madurez psicológica (los factores). ¿Cómo podremos saber si esa dependencia es verosímil? Parece claro que tendríamos distintos grupos de personas definidos en función de los dos factores. Un análisis de varianza (ANOVA) para un diseño factorial entre grupos sería una técnica de elección en este caso. También lo sería si medimos el tiempo que las personas tardan en resolver un problema (digamos la Torre de Londres), en función de la dificultad del mismo (la independiente). Si las mismas personas trabajan con problemas de diferente dificultad, es evidente que el análisis de varianza que debe aplicarse sería el correspondiente a un diseño de medidas repetidas simple.

Las técnicas bivariadas no implican la distinción entre variables independientes y dependientes, porque están orientadas a resolver preguntas de naturaleza no causal, sino correlacional. Típicamente son las que deben emplearse cuando se está interesado en determinar si dos variables están o no relacionadas entre sí. Por ejemplo, ¿hay relación entre inteligencia y sueldo que se gana? En este caso basta un simple cálculo del coeficiente de correlación para poder responder a la pregunta. Nótese que es indiferente la cuestión de si la inteligencia causa el sueldo o al revés. Lo importante es si hay relación, no si una variable causa los cambios observados en la otra.

Las técnicas multivariadas, por el contrario, se aplican cuando se han medido varias variables dependientes simultáneamente y se desea analizarlas conjuntamente. Por ejemplo, si además de medir el rendimiento académico, medimos el éxito social en función del estatus socioeconómico y la madurez psicológica y pretendemos considerar conjuntamente ambas variables. *A priori* está claro que las técnicas univariadas como el análisis de varianza no nos sirven, puesto que sólo son capaces de tratar simultáneamente una dependiente. Pero nótese que una solución relativamente simple puede consistir en generalizar el ANOVA. Esto es, una solución es aplicar análisis multivariado de varianza (MANOVA).

Tanto MANOVA como otras técnicas multivariadas (discriminante, regresión logística, etc.) son, en principio, más complejas, menos transparentes, que las univariadas. Pero eso no debe ocultar un hecho importante, las técnicas multivariadas son comprensibles y accesibles para cualquier estudiante sin necesidad de adentrarse en lo intrincado de su matemática. Afortunadamente, el desarrollo de los ordenadores y de los paquetes de análisis estadístico ha con-

seguido algo que es bastante relevante desde un punto de vista práctico, a saber: es posible tener un manejo suficiente de las técnicas y de la interpretación de sus resultados aun desconociendo los detalles matemáticos de las técnicas. Esa es la perspectiva con la que se ha escrito este libro. En los capítulos que se presentan a continuación se exponen las técnicas con un objetivo fundamental, que se conozca la lógica intrínseca de las mismas (por qué se aplica, cuándo se puede aplicar, qué resultados nos ofrece, cómo pueden interpretarse) sin hacer especial hincapié en su matemática. Es por ese motivo por el que se ofrecen los procedimientos que deben emplearse en paquetes estadísticos como SPSS, y las salidas de estos paquetes. Si el lector, cuando acabe su lectura de este libro, puede decidir qué técnica elegirá para analizar su investigación, por qué la elegirá, y sabe también interpretar el resultado del paquete estadístico que seleccione, consideraremos nuestro objetivo cumplido.

## 2. ¿CUÁNDO PUEDEN APLICARSE LAS TÉCNICAS MULTIVARIADAS?

Las técnicas multivariadas son aplicables en contextos de investigación muy diferentes. Sin embargo, su uso e interpretación dependen precisamente del contexto en que se apliquen. Por ello, distinguiremos primero entre contextos de estudio y después examinaremos el uso de las técnicas multivariadas en cada caso. Recordemos, para empezar, que las investigaciones pueden diferenciarse en tipos en función del papel activo o pasivo que el investigador tenga en ellas.

Claramente, el investigador es un agente activo cuando realiza manipulaciones tendentes a que la situación de investigación tenga una estructura determinada. Esto ocurre en un gran número de estudios, por ejemplo, cuando manipulamos la edad y seleccionamos un grupo de adultos jóvenes y otro de personas mayores y medimos el tiempo que invierten en resolver un conjunto de tareas que miden diferentes funciones cognitivas (es decir, orientación de la atención, capacidad de memoria a corto plazo, capacidad de planificación y secuenciación de conductas y selección e inhibición atencional), hemos estructurado la situación de modo que tenemos una variable independiente y varias variables dependientes (los tiempos invertidos en cada una de las tareas). Observemos que la independiente está manipulada de forma indirecta; esto es, que los sujetos no reciben la edad que nosotros deseamos sino que son seleccionados en función de que tengan o no la edad que nos interesa. Este hecho, aunque aparentemente poco importante, implica, en realidad que nuestra investigación no es experimental en sentido estricto, puesto que no podemos garantizar que los grupos difieran única y exclusivamente en la variable independiente. La falta de equivalencia de los grupos nos permite etiquetar a esta investigación como *cuasi experimental*. Consideraremos ahora otro caso. Supongamos que queremos estudiar el efecto que la práctica produce en personas mayores sobre la velocidad con la que resuelven las distintas tareas. Selecciona-

mos un conjunto de personas mayores y las asignamos al azar a dos grupos, uno que recibirá poca práctica y otro mucha. Es evidente que en este caso el azar nos garantizará a largo plazo (si el número de personas por grupo es suficientemente grande) la equivalencia de los grupos. Nuestra investigación sería ahora propiamente experimental. Nótese que la diferencia entre ambos tipos es importante, porque en el primer caso no podemos estar seguros de si las diferencias se deben a la edad o a otros factores (por ejemplo: la salud general de las personas) que pueden diferir de un grupo a otro, mientras que en el segundo caso, prácticamente podemos garantizar que las diferencias en velocidad serán debidas a la práctica. Esto es, en un caso nuestra afirmación, aunque puede ser causal (por ejemplo: el envejecimiento produce deterioro en funciones cognitivas), es bastante más insegura que en el segundo caso (por ejemplo: la práctica mejora el rendimiento en tareas que miden funciones cognitivas).

En contraste con estas aproximaciones podemos considerar una tercera, cuya característica más relevante es que no introduce manipulación. Supongamos ahora que pasamos una encuesta electoral a una muestra representativa de la población general. La encuesta tendrá un buen número de preguntas (ítem), pero nuestra función es puramente pasiva, se trata de registrar únicamente las respuestas de los encuestados. Claramente se trata de una investigación no experimental en la que, en principio, no hay interés en establecer relaciones causales de ningún tipo.

Las técnicas multivariadas pueden aplicarse en los tres tipos de investigación. Pero también las técnicas univariadas y bivariadas. Sin embargo, las preguntas que pueden responderse son de naturaleza diferente, en general, en unos casos y en otros. La elección de la técnica de análisis que se ha de emplear no viene completamente determinada por el tipo de investigación, pero es conveniente tener presente que la interpretación de los resultados de la técnica sí.

### 2.1. INVESTIGACIÓN EN CONTEXTOS NO EXPERIMENTALES

¿Cómo podemos analizar los resultados de una encuesta electoral? Este es un caso típico de investigación no experimental o de encuesta. El número de variables que hemos medido es igual al número de preguntas que plantea la encuesta. Es obvio que no podemos aplicar técnicas de análisis de varianza (que nos sirven sólo cuando hay manipulación de variables), pero sí podemos aplicar técnicas bivariadas. Por ejemplo, si hemos preguntado la creencia religiosa y el partido al que se piensa votar, es razonable preguntarse por la relación entre ambas variables. Además podemos intentar predecir una (el voto) a partir de la otra (la creencia religiosa). Pero esta aproximación, aunque no inválida, no tiene en cuenta algo importante, a saber: con bastante probabilidad las variables de la encuesta están relacionadas entre sí. Esta relación no es considerada en absoluto por las técnicas bivariadas, pero sí por las multivariadas. En una aproximación bivariada podemos llegar a conclusiones erróneas, por ejemplo,

es posible que la creencia religiosa no permita predecir el voto, pero nosotros concluyamos que sí porque no hemos tenido en cuenta otros ítem de la encuesta que están relacionados con ambas variables.

Una segunda cuestión importante es que las técnicas bivariadas no nos permiten percatarnos de que varias o muchas variables de nuestro estudio están en realidad midiendo lo mismo, es decir, pueden agruparse en una, digamos, supervariable o variado. Por el contrario, algunas técnicas multivariadas, como el análisis de *cluster* de variables, el análisis factorial, o el análisis discriminante están específicamente diseñadas para agrupar variables en supervariables.

Supongamos que en la pregunta de la encuesta electoral, los encuestados responden que piensan votar a uno de entre, digamos por simplificar, cuatro partidos políticos. ¿Cómo podemos distinguir a los votantes de uno y otro partido? Una estrategia, poco razonable, podría consistir en considerar los votantes de cada partido como un grupo diferente de sujetos. Esto es una manipulación indirecta *ex post facto*, después de hecho el estudio. Podríamos ahora realizar un análisis de varianza por cada uno de los ítem para determinar en cuáles de ellos difieren los votantes de los diferentes partidos. Obviamente tras el análisis global tendríamos que aplicar pruebas *a posteriori*, como la LSD de Fisher, para concretar entre qué partidos concretos es capaz de discriminar el ítem en cuestión. Sin embargo, hay tres dificultades importantes en esta aproximación analítica. Primero, desde un punto de vista conceptual, el cuadro de resultados será tan complejo que difícilmente podremos extraer conclusiones de carácter general. ¿Qué podremos concluir si la pregunta de creencia religiosa distingue entre votantes del partido A y B por un lado y votantes de los partidos C y D por otro, mientras que la pregunta de frecuencia de asistencia a actos de su religión distingue entre votantes de A y D por un lado y votantes de B y C por otro? Aunque logremos concluir algo, quizás el esfuerzo necesario no lo haga suficientemente valioso. Segundo, desde un punto de vista metodológico es esperable que los ítem de la encuesta muestren un patrón de correlaciones complejo. Considerando cada ítem por separado estamos despreciando la correlación, pero esa correlación es valiosa, puede servir para descubrir la estructura interna de la encuesta, para descubrir supervariables que serán sin duda más informativas que las variables por separado. Tercero, si analizamos los ítem por separado, es obvio que el número de análisis y de decisiones sobre significación que habremos de tomar será elevado. La cuestión es sencilla, si en cada decisión corremos un riesgo de equivocarnos, es evidente que el riesgo global de error será también muy alto. En pocas palabras, es imposible mantener una adecuada tasa global de error tipo I (declarar diferencias cuando de hecho no las hay) sin adoptar restricciones sobre el error admisible en cada decisión sobre cada ítem. Cuarto, y también metodológico, si hay correlaciones entre variables, la aplicación del análisis de varianza, o de su análogo en el caso de tener sólo dos grupos, la *t* de Student es inadecuada. El estadístico *F* (o la *t*) son erróneos desde un punto de vista estadístico, precisamente porque no tienen en cuenta esas correlaciones.

Por el contrario las técnicas multivariadas nos permiten analizar estudios de este tipo de una manera natural. Por ejemplo, el análisis discriminante o la regresión logística nos permiten predecir el voto teniendo en cuenta simultáneamente todos los ítem de la encuesta. Esto implica, por supuesto, que se consideran las correlaciones entre los ítem. Además, es posible determinar cuáles de los ítem contribuyen a la predicción y qué otros pueden ser excluidos. Otras técnicas, como el análisis de *cluster* nos permiten formar grupos de individuos aunque *a priori* no tengamos ningún criterio de clasificación de los mismos. Los individuos pueden ser asignados a unos grupos u otros en función de su semejanza.

## 2.2. INVESTIGACIÓN EN CONTEXTOS EXPERIMENTALES

En un contexto experimental, como hemos señalado más arriba, manipulamos variables independientes con la intención de afectar a una o más variables dependientes. Las técnicas multivariadas son útiles cuando hemos medido más de una variable dependiente, particularmente si hay correlaciones entre las medidas. Como en el caso no experimental, las ventajas de las técnicas multivariadas residen en la posibilidad de mantener el error tipo I a un nivel aceptable, por ejemplo, igual o menor que 0,05, pero también en el hecho de que tienen en cuenta las correlaciones para construir supervariables o variados sobre los cuales se aplica el análisis (análisis discriminante, regresión logística, MANOVA). Finalmente, la aproximación a situaciones reales es también mayor desde una orientación multivariada. Consideremos, por ejemplificar, que estamos intentando comprobar la eficacia de un tratamiento de rehabilitación neuropsicológica y que consideramos que debe ejercer efecto sobre la atención y sobre la memoria operativa de los pacientes. Comparamos, por tanto, nuestro tratamiento frente a un tratamiento control, asignando los pacientes al azar a los grupos de rehabilitación. A cada paciente le realizamos una medida de atención (por ejemplo, mediante una prueba de interferencia Stroop) y su memoria operativa (por ejemplo, mediante una prueba de capacidad de memoria). Podríamos analizar las diferencias entre los dos grupos en cada medida por separado (mediante *t* de Student), pero, de nuevo, ese análisis no tiene en cuenta que habrá correlación entre las medidas. Esto es cierto también desde un punto de vista teórico, puesto que atención y memoria operativa comparten circuitos anatómicos y funcionales en el cerebro humano y además, comparten un proceso fundamental, el ejecutivo central. Así, no considerar la correlación es tanto un error metodológico como teórico. En efecto, el análisis por separado de las variables dependientes correlacionadas implica que comparten varianza y, por tanto, en cada análisis estamos considerando varianza que ya hemos considerado en análisis precedentes. En este ejemplo, un MANOVA sería una alternativa analítica apropiada. Por otro lado, el estudio es más realista, puesto que no parece haber dudas de que la rehabilitación debe

mejorar simultáneamente las capacidades de inhibir información irrelevante (atención) y la capacidad de memoria operativa.

En resumen, las técnicas multivariadas son aplicables tanto en contextos experimentales como no experimentales siempre que: (1) tengamos más de una variable medida (dependiente, criterio, etc.); (2) haya correlación entre las medidas; y (3) deseemos evitar cometer una tasa de error tipo I global elevada. Estas técnicas pueden considerarse una extensión de las univariadas (análisis de varianza, análisis de regresión simple, cuyos equivalentes multivariados serían el análisis multivariado de varianza y la regresión múltiple) y de las bivariadas (clasificación cruzada, correlación, que puede extenderse a correlación canónica), aunque hay técnicas que están diseñadas como multivariadas en su inicio, como el análisis factorial o el análisis de *cluster*.

### 3. CLASIFICACIÓN DE LAS TÉCNICAS MULTIVARIADAS

Las técnicas multivariadas son poco homogéneas y pueden clasificarse según diferentes criterios (Bisquerra, 1989). Uno de los más interesantes para nosotros es el que atiende a la función de la técnica. En este sentido se distinguen técnicas explicativas y técnicas descriptivas. Las técnicas explicativas se emplean en contextos de investigación experimentales mientras que las descriptivas se usan en contextos no experimentales. Las primeras nos permiten contrastar hipótesis relativas a la influencia que nuestras manipulaciones, las variables independientes, tienen sobre las variables dependientes. La etiqueta de explicativas obedece a que las variables independientes son consideradas las causas potenciales de los cambios observados en las variables dependientes. Por el contrario, las técnicas descriptivas se caracterizan por que nos permiten resumir el conjunto de datos de una forma adecuada a nuestros objetivos. Con estas técnicas no se pretende, ni se puede, contrastar hipótesis sobre efectos de factores, sino obtener una descripción ya de los diferentes patrones de respuestas, ya una información sobre la estructura interna (las supervariables) de nuestras medidas. A continuación presentaremos las técnicas que van a ser estudiadas a lo largo de este manual. En cada una de las que presentamos a continuación incluiremos una definición concisa de la técnica y de su objetivo fundamental. Explicaciones más detalladas se presentarán en los capítulos dedicados a cada una de ellas.

#### 3.1. TÉCNICAS EXPLICATIVAS

En este conjunto de técnicas incluiremos todos los análisis multivariados encaminados a explicar o predecir una variable o un conjunto de variables. Es necesario que primero realicemos una precisión. En este manual vamos a introducir no sólo técnicas multivariadas en sentido estricto, es decir, aquellas que

analizan simultáneamente más de una variable medida o dependiente, sino también técnicas multivariadas, es decir, que consideran simultáneamente un conjunto de variables predictoras o independientes (como la regresión múltiple).

##### 3.1.1. Análisis de regresión múltiple

En muchas investigaciones se mide una sola variable dependiente o criterio, y varias independientes o predictores, que no han sido manipulados, sino que, en general, son características que los individuos poseen en cierto grado. Cuando lo que se pretende es predecir el criterio desde el conjunto de predictores métricos (cuyos valores expresan magnitud de la variable y no simplemente orden o posesión de un atributo) medidos, la técnica de elección es el análisis de regresión múltiple. Este objetivo de predicción suele conseguirse combinando linealmente los predictores. La combinación lineal es simplemente una suma ponderada de los valores de los predictores. Por ejemplo, podemos sumar sin más las puntuaciones de los sujetos en un test de inteligencia como el WAIS. Pero esa suma es, en realidad ponderada, puesto que estamos asumiendo que el peso de cada uno de los ítem es el mismo. Una alternativa podría ser dar un peso diferente a cada ítem, lo que implica que su valor (asumiendo que la escala es la misma en todos) no es igual en cuanto a la medición de la inteligencia. Si pretendemos usar el test para predecir, digamos, el rendimiento académico, entonces cada ítem es un predictor que puede ser más o menos valioso que otros para hacer la predicción. Los pesos de cada ítem o variable se determinan habitualmente mediante un proceso de estimación conocido como mínimos cuadrados, cuyo objetivo es obtener pesos que hagan mínima la suma de los cuadrados de los errores de predicción. La regresión múltiple es utilizable siempre que las variables estén medidas en una escala de intervalo o de razón, es decir, siempre que sean escalares (números cuyo valor indique magnitud de la variable).

##### 3.1.2. Análisis discriminante

A menudo la variable que pretendemos predecir es dicotómica (puede adoptar únicamente dos valores) o politérmica (puede adoptar más de dos valores), pero no cuantitativa, no métrica, sino nominal u ordinal. Por ejemplo, si definimos el rendimiento académico en función de si los alumnos aprueban o no el curso completo, la variable adopta dos valores no métricos: apto y no apto. En estos casos, en que la muestra de individuos puede dividirse en grupos conocidos de antemano según un criterio, el análisis discriminante es una técnica útil, siempre que los predictores (las variables independientes) sean variables escalares. En el análisis se pretenden varios objetivos. En primer lugar, calcular la verosimilitud de que los individuos pertenezcan a uno u otro de los



grupos a partir del conjunto de predictores. En segundo lugar, determinar cuáles de las variables predictoras son realmente útiles para hacer la predicción. La idea en este caso es que algunos de los predictores medidos pueden ser irrelevantes para determinar la pertenencia a un grupo. Por ejemplo, la predicción del voto puede realizarse a partir del perfil de las respuestas a la encuesta electoral. La definición del perfil no se hará, probablemente, con todos los ítem, sino solamente con los que muestren un incremento significativo en la predicción cuando son introducidos en la ecuación de predicción.

### *3.1.3. Análisis multivariado de varianza y covarianza*

La extensión del análisis univariado de varianza (ANOVA) es el análisis multivariado de varianza (MANOVA). Como en aquél, el objetivo de MANOVA es determinar si una o más variables independientes producen efectos diferenciales sobre dos o más variables dependientes. Las dependientes tienen que ser variables métricas, pero las independientes pueden serlo o no. Es importante tener en cuenta que MANOVA es aplicable también cuando se ha medido la misma variable dependiente de forma reiterada en el tiempo a los mismos sujetos, esto es, en diseños de medidas repetidas. Cuando en el diseño no se ha podido realizar un control experimental adecuado es frecuente intentar un control estadístico midiendo variables (covariados) que estén relacionadas con las dependientes antes de que se apliquen las independientes a los sujetos. La estrategia pasa por asumir que lo que puede ser predicho por los covariados no es atribuible a las independientes. En este caso la técnica sería análisis multivariado de covarianza (MANCOVA). Tanto en MANOVA como en MANCOVA se trata de construir variados mezclando las variables dependientes de forma que discriminan lo mejor posible a los grupos.

### *3.1.4. Análisis de regresión logística*

Cuando las variables predictoras no son métricas, no es recomendable emplear análisis discriminante para tratar de predecir la verosimilitud de que un individuo pertenezca a un grupo u otro. En este caso la técnica de elección es la regresión logística. Su objetivo es similar al del análisis discriminante: predecir el grupo al que pertenece un individuo. El procedimiento también es similar, se trata de mezclar las variables predictoras con la ponderación adecuada, sin embargo, la relación entre la mezcla de variables y la variable predicha no se asume lineal como en el análisis discriminante. Además, la estimación de los pesos se realiza por un procedimiento iterativo conocido como máxima verosimilitud.

### *3.1.5. Ecuaciones lineales estructurales*

En ocasiones nuestro interés investigador no se centra en relaciones de predicción únicas, sino secuenciales. Por ejemplo, podemos pretender predecir la capacidad de memoria operativa a partir de un conjunto de tareas que miden cosas como la capacidad para coordinar la realización de tareas simultáneas, de alternar entre tareas o de actualizar el contenido de la información en memoria a medida que va apareciendo información nueva. A su vez, podemos intentar predecir la inteligencia general, medida, por ejemplo, mediante las Matrices Progresivas de Raven, a partir de la capacidad de memoria operativa. Nótese que en este caso una variable predicha se convierte en un predictor en el paso siguiente. La técnica adecuada para este tipo de análisis son las ecuaciones estructurales (LISREL es un nombre popular que se debe al programa más conocido para realizar los análisis). El análisis se caracteriza por la construcción de un modelo estructural en el que se especifica cómo se supone que están relacionados las variables predictoras y los criterios. La técnica en sí es neutral respecto de la estructura de las relaciones, por lo que el modelo suele basarse en presupuestos teóricos. Por ejemplo, las tareas mencionadas arriba se emplean para medir el ejecutivo central, un mecanismo encargado del control del procesamiento consciente y probablemente también no consciente (Catena y Valdés, 2002, para una revisión). A su vez, la memoria operativa se supone un compuesto de memoria a corto plazo y ejecutivo central. Por otra parte, las áreas cerebrales relacionadas con la memoria operativa y la inteligencia están solapadas (Duncan y cols., 2000). Este conjunto de conocimiento teórico nos indicará la estructura de relaciones más razonable entre las variables. En segundo lugar, el modelo de medida define qué variables se van a emplear para cada variable criterio o predictora. Por ejemplo, la capacidad de memoria operativa es un concepto que podemos evaluar mediante varias tareas, sean la tarea de amplitud lectora, la de amplitud de operaciones y la de dígitos hacia atrás de los test de Wechsler. El modelo de medida nos permite determinar cómo cada tarea contribuye a la medición de la capacidad de memoria operativa.

## **3.2. TÉCNICAS DESCRIPTIVAS**

No siempre el objetivo del análisis consiste en predecir una variable a partir de otras, en determinar si las manipulaciones que hemos realizado han producido efecto. A menudo, aunque sea como paso previo, estamos interesados en resumir nuestros resultados. Resumir significa reducir. Por ejemplo, cuando medimos la edad a un grupo de estudiantes y decimos que la media del grupo es 21, hemos resumido el conjunto de datos de edad a uno solo. El resumen hace más comprensible, y más tratable al grupo. Este dato representa al conjunto en mayor o menor medida, dependiendo de la variabilidad de la mues-

tra (otra forma de resumir). A continuación presentamos las técnicas multivariadas que persiguen el objetivo de resumir los datos.

### 3.2.1. Análisis de datos categóricos

En muchas investigaciones las variables que medimos no son escalares, sino de carácter categórico (nominal) u ordinal. Por ejemplo, el tipo de psicopatología que se sufre (esquizofrenia, depresión, etc.), el área geográfica donde se vive (norte, centro, sur), son variables que no pueden ser, en principio, tratadas numéricamente (aunque véase, Young, 1981, para una aproximación cuantitativa al análisis de datos cualitativos mediante escalamiento óptimo y mediante mínimos cuadrados). En muchos casos nuestro objetivo es determinar si hay relación entre las variables medidas, por ejemplo, entre área geográfica y psicopatología.

La forma más común de analizar las relaciones entre variables cualitativas es el análisis de frecuencias de múltiples vías, cuya extensión es más conocida como análisis log-lineal. Cuando sólo hay dos variables un modo simple de determinar si están relacionadas consiste en usar el estadístico  $\chi^2$ , que mide la asociación entre ambas. Cuando el número de variables es mayor de dos, la aproximación es el análisis de tablas de contingencia múltiples. Para ello se construyen tablas de asociación de dos, tres o más entradas y se construye un modelo lineal del logaritmo de las frecuencias esperadas de las casillas de las tablas. Los predictores son ahora los efectos de las variables y sus interacciones. El análisis progresará eliminando tantos predictores como es posible sin que se incremente significativamente la diferencia entre las frecuencias observadas y las predichas.

El análisis de datos categóricos también puede realizarse mediante una técnica conocida como análisis por correspondencias. Su objetivo es también estudiar las dependencias entre las variables. Y también procede computando tablas de contingencia. Sin embargo, realiza una transformación de los datos no métricos para convertirlos en datos métricos y obtiene las dimensiones subyacentes a los mismos y mapas que representan la cercanía de los valores de las variables medidas.

### 3.2.2. Análisis de correlación canónica

Supongamos que medimos a un grupo de sujetos dos conjuntos de variables, digamos, varias medidas de personalidad y varias medidas de funcionamiento cognitivo. En principio no pretendemos establecer una relación causal entre personalidad y cognición, pero sí es frecuente que estemos interesados en conocer si entre ambas existe un grado significativo de relación. El análisis de correlación canónica es idóneo para responder a ese tipo de cuestiones. Su

forma de proceder es semejante a la del análisis discriminante: lo que se correlaciona son combinaciones lineales de las variables de un conjunto con combinaciones lineales de las variables del otro conjunto. La semejanza es suficientemente grande como para que algunos paquetes como SPSS incorporen el análisis de correlación canónica dentro del programa de análisis discriminante.

### 3.2.3. Análisis de cluster

En ocasiones tenemos interés en clasificar individuos (o variables) semejantes entre sí en las variables que les hemos medido, pero no tenemos un criterio de clasificación *a priori*. El análisis de *cluster* es una técnica cuyo objetivo es precisamente agrupar a individuos (o variables) en función de su semejanza. Esto es lo mismo que decir que el análisis es idóneo cuando queremos saber si en nuestra muestra de sujetos podemos diferenciar grupos de individuos diferentes. Hay una gran variedad de técnicas para resolver este problema, pero, en general, todas coinciden en computar primero una medida de semejanza, distancia o asociación entre individuos y en agruparlos posteriormente mediante algún algoritmo de agrupación. Las técnicas difieren sobre todo en los algoritmos que emplean. Finalmente, los grupos se caracterizan en función de los individuos que los componen.

### 3.2.4. Análisis factorial y de componentes principales

Imaginemos que hemos aplicado un buen número de tareas para medir la memoria a corto plazo (por ejemplo, bloques de Corsi, amplitud de dígitos hacia delante, hacia atrás, amplitud lectora, etc.). Si calculamos las correlaciones entre ellas encontraremos que hay subconjuntos de variables que correlacionan alto entre sí, pero bajo con respecto a otros subconjuntos. Seguramente encontraremos que Corsi y dígitos hacia delante correlacionan entre sí, pero poco con dígitos hacia atrás y amplitud lectora, que a su vez correlacionarán entre sí. Agrupando las variables según la correlación podemos obtener factores, que se suponen indicadores del proceso que ha dado lugar a la correlación. Claramente en nuestro ejemplo, tendríamos dos factores, uno que podríamos etiquetar de memoria a corto plazo (en el que son importantes los dígitos hacia delante y Corsi) y otro de ejecutivo central (en el que son importantes dígitos hacia atrás y amplitud lectora). Tanto el análisis factorial como el de componentes principales tienen el objetivo de reducir los datos a partir de las correlaciones para encontrar un número reducido de factores que los expliquen. Como hemos podido comprobar las etapas son simples, medir las variables, computar las correlaciones y arreglarlas en una matriz, y extraer los factores, que son combinaciones lineales de las variables. A menudo, no siempre, es necesario rotar los factores con el fin de interpretar mejor los resultados.

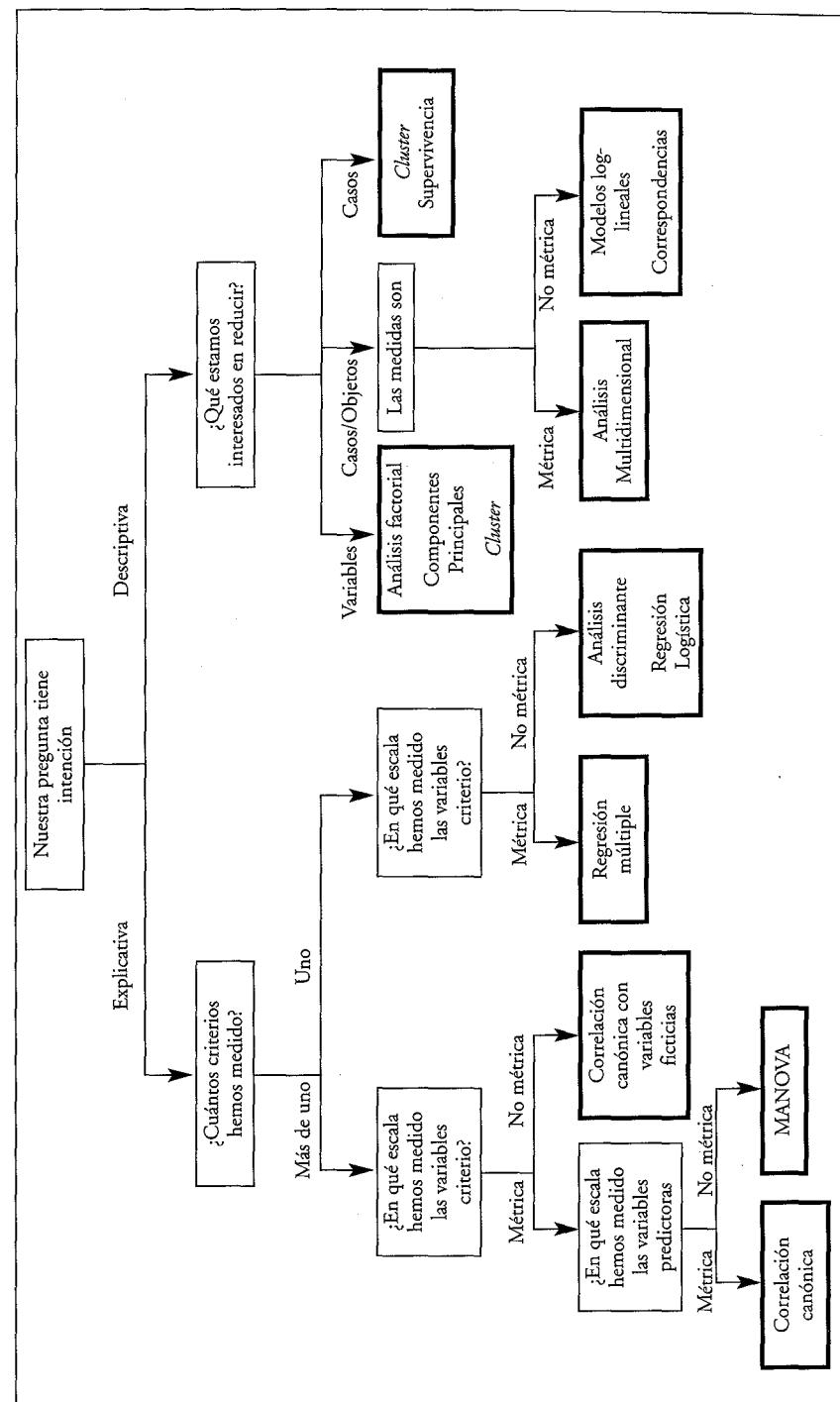
### 3.2.5. Análisis de supervivencia

En determinados contextos de investigación (por ejemplo, envejecimiento, rehabilitación, empleo, etc.) nuestro interés recae en determinar cuánto tiempo es necesario que transcurra para que suceda algo (muerte, cura, logro de empleo, etc.). En general, se trata de describir la proporción de casos que en diferentes momentos del tiempo permanecen en el estudio. ¿Cuántos pacientes siguen en rehabilitación después de transcurrido cierto tiempo? A menudo se emplea más de un grupo, por ejemplo, uno en rehabilitación y otro sin ella. El análisis presenta las curvas de supervivencia para cada grupo, pero también es posible realizar pruebas estadísticas para comparar los grupos.

### 3.3. ¿CÓMO SELECCIONAR LA TÉCNICA DE ANÁLISIS ADECUADA?

La selección de la técnica que vamos a emplear para analizar nuestros datos es un proceso complejo, pero bien estructurado. Comencemos por establecer una cuestión fundamental. Toda investigación se realiza porque se pretende resolver un problema, porque se intenta responder a una pregunta. Eso significa que la investigación tiene un objetivo claro. Pues bien, el diseño de la investigación dependerá en gran medida de ese objetivo, pero también la técnica de análisis que empleemos. Por lo que hemos visto hasta aquí, parece claro que nuestra pregunta puede implicar una manipulación de variables o, al menos, una relación entre ellas. Está claro que si nuestro caso es el primero, nuestra elección tiene que realizarse entre las diferentes técnicas explicativas, pero si es el segundo, tendremos que elegir entre las descriptivas. En segundo lugar, nuestra pregunta puede requerir que midamos más de una variable dependiente (cuando hay variables predictoras). Por último, es importante considerar en qué escala hemos medido las variables, concretamente si son métricas o no. En función de estos tres criterios las decisiones pueden encadenarse de forma sistemática hasta llegar a una determinación completa de la técnica adecuada. El Cuadro 1.1. (página siguiente) presenta esta cadena de decisiones. Es importante tener presente que las decisiones no son rígidas, en el sentido de que una misma pregunta puede resolverse con diferentes técnicas.

Cuadro 1.1.  
La elección de la técnica de análisis según la intención del estudio, las variables predictoras y la escala de medida



## 4. CONCEPTOS BÁSICOS

A continuación presentaremos algunos conceptos fundamentales para la correcta comprensión del resto de los capítulos. En la medida de lo posible los epígrafes siguientes están orientados a una comprensión intuitiva, aunque en ocasiones la exposición será más técnica. No obstante, debe tenerse presente que el desarrollo actual del *software* estadístico se ocupa de forma completa de los cálculos, de manera que nuestra intención fundamental es facilitar que el detalle de las técnicas multivariadas pueda seguirse adecuadamente.

### 4.1. DEFINICIONES ÚTILES

Además de los términos que hemos ido introduciendo a lo largo de este capítulo es importante introducir algunas definiciones más que ayudarán a comprender tanto el análisis clásico como el multivariado.

#### 4.1.1. Escalas de medida

Medir es asignar nombres a las modalidades de una variable. Cuando medimos la edad, asignamos nombres que son valores numéricos a las diferentes modalidades, edades. También lo hacemos cuando medimos la creencia religiosa, pero ahora los nombres no representan magnitud alguna (católico, protestante, ortodoxo, musulmán, etc.). Según la escala de medida se distingue entre variables cualitativas y variables cuantitativas.

Las variables cualitativas se caracterizan porque se miden en una escala nominal. Esta escala asigna nombres a las modalidades sin ninguna referencia a magnitud (la creencia religiosa). No es posible hacer operaciones con estas variables, por ejemplo, no podemos decir cuánta diferencia hay entre católico y protestante.

Las variables cuantitativas se miden con escalas no nominales. Podemos distinguir dos tipos: ordinales y escalares. Las variables ordinales, también llamadas de rango, sí implican magnitud, aunque no podemos calcular tampoco diferencias entre unos valores y otros. Por ejemplo, si en una encuesta preguntamos «¿Asiste regularmente a actos religiosos?» y ofrecemos como alternativas de respuesta «poco, regular, mucho», está claro que no podemos saber qué diferencia hay entre poco y regular, o entre poco y mucho.

Las variables cuantitativas escalares se miden con escalas de intervalo o de razón. Las escalas de intervalo no tienen cero real. Por ejemplo, la escala CI de inteligencia carece de cero. Ello implica que es posible calcular diferencias entre, pongamos, dos individuos. Si uno tiene un CI de 100 y el otro de 120, la diferencia es de 20 puntos de CI, pero no podemos decir que el segundo

es 1,2 veces más inteligente que el primero. Por el contrario, en las escalas de razón, sí es posible decir cuántas veces un individuo es superior a otro en la variable. Si medimos el número de problemas resueltos en 20 minutos, y un individuo ha solucionado 10 y el otro 12, ahora sí podemos afirmar que el segundo ha resuelto 1,2 veces más problemas que el primero. Las variables escalares son variables métricas en sentido estricto.

La distinción es importante porque, como hemos visto, el análisis que es posible aplicar depende de manera fundamental del tipo de escala de medida con que hayamos medido la variable.

#### 4.1.2. Muestras y poblaciones

Investigamos prácticamente siempre con muestras, pero casi nunca nos interesa hacer afirmaciones sobre ellas, sino sobre poblaciones. Las muestras las extraemos de poblaciones, normalmente mediante métodos aleatorios, intentando garantizar que son representativas de las poblaciones de referencia. La inferencia estadística nos permite hacer afirmaciones sobre las poblaciones a partir de los datos que hemos registrado en las muestras. Cuando trabajamos con muestras usamos el término estadísticos para hacer referencia a sus características (media, varianza, etc.). En el caso de poblaciones utilizamos el término parámetros. La inferencia consiste, pues, en usar los estadísticos para hacer afirmaciones sobre los parámetros.

#### 4.1.3. Vectores y matrices

Si medimos a un individuo el tiempo que tarda en realizar una tarea, los aciertos que realiza, los errores de omisión y los errores de comisión, conocemos cuatro datos sobre él, que podemos ordenar. Un conjunto de números ordenados en una fila o en una columna es un vector (fila o columna, respectivamente). Por tanto, podemos referirnos al vector ( $\mathbf{x}$ ) del sujeto, para indicar las puntuaciones de ese sujeto. Las dimensiones del vector serán  $1 \times p$ , donde 1 se refiere a la fila y  $p$  a la columna, o  $p \times 1$ . Si en lugar de tener un solo sujeto tenemos un conjunto  $n$ , y ordenamos los datos de manera que cada sujeto ocupe una fila y cada variable medida ocupe una columna, tendremos una matriz de puntuaciones ( $\mathbf{X}$ ). Las matrices son, pues, conjuntos de números ordenados en filas y columnas, que podemos descomponer en vectores. Su dimensión es  $n \times p$ , donde  $n$  es el número de filas y  $p$  el de columnas. Nótese que la letra minúscula se usa para el vector y la mayúscula para la matriz. La negrita indica que no se trata de un único escalar, sino de un conjunto. Para referirnos a los elementos de la matriz emplearemos la letra minúscula sin negrita y con dos subíndices, el primero para la fila y el segundo para la columna. Así,  $x_{21}$ , indica el valor obtenido por el sujeto 2 en la variable 1.

Cuando tenemos un grupo de sujetos medidos en varias variables podemos calcular la media del grupo en cada variable. De esta forma podemos obtener un vector de medias que representa al grupo. Este vector de medias suele llamarse centroide o centro del grupo. Por ejemplo, si el grupo tiene dos sujetos con puntuaciones:

$$\begin{array}{l} S_1: \quad 4 \quad 6 \quad 3 \quad 2 \\ S_2: \quad 6 \quad 4 \quad 3 \quad 4 \end{array}$$

Entonces el centroide del grupo será:  $c = (5 \quad 5 \quad 3 \quad 3)$ .

Dos casos importantes de matrices nos interesarán especialmente, las de sumas de cuadrados y productos cruzados y las de correlaciones. La primera contiene las sumas de los productos de unas variables por otras y la segunda las correlaciones entre variables. Estas matrices desempeñan un papel crucial en una gran cantidad de análisis multivariados. Volveremos sobre este punto cuando presentemos las operaciones que pueden realizarse con matrices.

#### 4.1.4. Mezcla de variables

Muchas técnicas multivariadas mezclan variables. Por ejemplo, el análisis discriminante mezcla los predictores para predecir la verosimilitud de pertenencia a un grupo. La mezcla de variables suele ser lineal, exceptuando algunos casos como la regresión logística o los modelos log-lineales. La combinación lineal de variables es conceptualmente muy simple. Se trata de sumarlas pero teniendo en cuenta que cada una de ellas puede tener un peso diferente. Por ejemplo, si hemos medido las cuatro variables mencionadas en el epígrafe anterior (tiempo,  $x_1$ , aciertos,  $x_2$ , errores de comisión,  $x_3$ , y errores de omisión,  $x_4$ ), podemos mezclarlas dándole un peso ( $b$ ) propio a cada una, de este modo tendremos una nueva puntuación ( $y'$ ) resultante de la mezcla, lo que podemos llamar una supervariable o un variado. El variado no ha sido medido directamente, pero teóricamente puede ser muy relevante porque nos está proporcionando información sobre el significado subyacente del conjunto de variables. Esto es:

$$y' = b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4$$

Es interesante destacar que los pesos de las variables proporcionan información sobre el valor de la variable para hacer la predicción. Puede haber variables, y de hecho frecuentemente las hay, que no aportan nada significativo y simplemente es recomendable suprimirlas. Evidentemente una parte fundamental del análisis pasa por el cálculo de los pesos de las variables. La estimación de pesos no se realiza normalmente explorando el espacio de parámetros, aunque pueda hacerse, sino por procedimientos algorítmicos que

suelen basarse en las matrices de correlaciones o de sumas de cuadrados y productos cruzados.

#### 4.2. CONTRASTE DE HIPÓTESIS UNIVARIADO Y MULTIVARIADO

Una hipótesis es un establecimiento de hechos, o mejor, una predicción sobre hechos. Las hipótesis se formulan con referencia siempre a un problema de investigación. Así, si nos planteamos como problema si el envejecimiento normal produce un déficit cognitivo (de memoria, atención, etc.) de carácter subclínico, una hipótesis es simplemente una solución posible del problema. Naturalmente siempre es posible establecer dos soluciones genéricas. En primer lugar, podemos hipotetizar que el envejecimiento no produce déficit cognitivo. Por supuesto, la hipótesis complementaria es que sí lo produce. Éstas son hipótesis de investigación. Puesto que son complementarias tomar una decisión sobre una, la primera, implica tomarla también sobre la otra. Pero las hipótesis de investigación no pueden ponerse a prueba sin más. Necesitamos traducirlas, dotarlas de contenido estadístico. La concreción no es posible sin saber qué vamos a medir. Puesto que nuestro objetivo es el déficit cognitivo durante el envejecimiento, podríamos emplear un grupo de adultos jóvenes y otro de personas mayores, digamos de 60 años en adelante. El déficit cognitivo lo podríamos medir con una tarea de atención, una tarea de memoria a corto plazo y otra de memoria a largo plazo. Supongamos que en la tarea de atención medimos el tiempo de reacción, que en la de memoria a corto plazo medimos la amplitud lectora y en la de memoria a largo plazo el recuerdo correcto de pares asociados. Así, tenemos un diseño con una variable independiente manipulada por selección (la edad) y tres variables dependientes diferentes.

Como hemos señalado antes nuestra estrategia de análisis podría consistir en analizar cada variable dependiente por separado. Por tanto, tendríamos tres hipótesis de investigación posibles:

- 1) Los tiempos de reacción en la tarea de atención no son afectados por la edad.
- 2) La amplitud lectora es la misma en las diferentes edades.
- 3) El recuerdo correcto es idéntico en las diferentes edades.

Nótese que las tres indican que no hay diferencias entre los grupos. Ahora bien, la manera en que resumimos la ejecución es crucial. Parece claro que resumir es encontrar una puntuación que represente al conjunto. La media es una buena alternativa. Caigamos ahora en la cuenta de que no necesitamos hacer ningún contraste si nuestra intención fuese hablar solamente de los grupos que hemos seleccionado. Nos bastaría con mirar sus medias y concluir en función de si son o no diferentes. Esto es, lo que nos interesa no son los grupos, sino las poblaciones de las cuales los hemos extraído. Lo importante es

poder afirmar si la población de personas mayores difiere de la de adultos jóvenes. Por tanto, las medias que nos interesan, y sobre las que de hecho hacemos las afirmaciones, son poblacionales. Así pues, tendremos tres hipótesis estadísticas, que afirman que las medias poblacionales (llamémoslas  $\mu_i$ ) son iguales:

- 1) Tiempo de reacción:  $H_{0(1)}: \mu_{11} = \mu_{21} \Rightarrow \mu_{11} - \mu_{21} = 0$
- 2) Amplitud lectora:  $H_{0(2)}: \mu_{12} = \mu_{22} \Rightarrow \mu_{12} - \mu_{22} = 0$
- 3) Recuerdo correcto:  $H_{0(3)}: \mu_{13} = \mu_{23} \Rightarrow \mu_{13} - \mu_{23} = 0$

donde el primer subíndice hace referencia al grupo y el segundo a la variable dependiente. Obsérvese que expresadas en forma diferencial, como aparecen a la derecha, las hipótesis nos ponen sobre la pista de algo fundamental, a saber: para tomar una decisión, hacer el contraste, partimos de la diferencia entre las medias. En este caso simple, la diferencia elevada al cuadrado es el numerador de la varianza. Esto es, el contraste puede consistir en realizar tres análisis de varianza, uno por cada criterio.

Como ya hemos señalado uno de los problemas fundamentales de esta aproximación es que no considera la correlación entre las variables medidas, lo que puede llevar a errores en la toma de decisiones. Por ejemplo, es posible que en cada medida por separado no haya diferencias en función de la edad, que sí aparezcan cuando consideramos las medidas conjuntamente. El contraste multivariado se caracteriza en este caso justamente por tener en cuenta todas las medidas de forma simultánea. Esto es, el contraste se hace sobre los centroides, de grupo en nuestro ejemplo:

$$H_0: \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \end{pmatrix} = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{pmatrix}$$

Más adelante veremos cómo se pueden contrastar centroides. Por ahora, quedémonos con la idea de que el contraste se realiza de forma simultánea sobre el conjunto de variables medidas.

#### 4.3. MATRICES

Las matrices son conjuntos de números ordenados en filas y columnas. Su ventaja es que podemos realizar operaciones con ellas como si fueran un solo escalar. A continuación presentaremos algunos de los cálculos más importantes para comprender las técnicas estadísticas.

##### 4.3.1. Operaciones elementales con matrices

Con matrices se pueden realizar sumas y multiplicaciones. Como veremos la división tiene que realizarse por un procedimiento especial. Para exemplificar consideremos la matriz (X) simple de más arriba, que reproducimos inmediatamente debajo:

$$\begin{pmatrix} 4 & 6 & 3 & 2 \\ 6 & 4 & 3 & 4 \end{pmatrix}$$

y que contiene los datos de dos sujetos medidos en cuatro variables. Supongamos ahora otra matriz (M)

$$\begin{pmatrix} 5 & 6 & 2 & 3 \\ 6 & 2 & 4 & 5 \end{pmatrix}$$

Sumar ambas matrices consiste en sumar los elementos correspondientes, por tanto, la matriz suma (S) será:

$$\begin{pmatrix} 4+5 & 6+6 & 3+3 & 2+3 \\ 6+6 & 4+2 & 3+4 & 4+5 \end{pmatrix}$$

Restar ambas matrices consiste, de manera similar en sustraer una de la otra. Por ejemplo, la matriz X-M, será:

$$\begin{pmatrix} 4-5 & 6-6 & 3-3 & 2-3 \\ 6-6 & 4-2 & 3-4 & 4-5 \end{pmatrix}$$

La multiplicación de matrices requiere que el número de columnas de la primera sea igual al número de filas de la segunda. Así, las matrices X y M no pueden multiplicarse. Un caso importante para nuestros objetivos es el de una multiplicación de una matriz por sí misma, o mejor dicho, por su traspuesta. Trasponer una matriz es intercambiar filas por columnas. La traspuesta de X se escribe como  $X'$ , y será:

$$\begin{pmatrix} 4 & 6 \\ 6 & 4 \\ 3 & 3 \\ 2 & 4 \end{pmatrix}$$

Veamos el producto de matrices en el apartado siguiente.

### 4.3.2. Matrices de sumas de cuadrados y productos cruzados

Una matriz traspuesta multiplicada por la matriz original produce una matriz de sumas de cuadrados y sumas de productos. Estas matrices son fundamentales en muchos cálculos en análisis multivariado. El producto  $X'X$  será:

$$\begin{pmatrix} 4 & 6 \\ 6 & 4 \\ 3 & 3 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 4 & 6 & 3 & 2 \\ 6 & 4 & 3 & 4 \end{pmatrix} = \begin{pmatrix} 4x^4+6x^6 & 4x^6+6x^4 & 4x^3+6x^3 & 4x^2+6x^4 \\ 6x^4+4x^6 & 6x^6+4x^4 & 6x^3+4x^3 & 6x^2+4x^4 \\ 3x^4+3x^6 & 3x^6+3x^4 & 3x^3+3x^3 & 3x^3+3x^4 \\ 2x^4+4x^6 & 2x^6+4x^4 & 2x^3+4x^3 & 2x^2+4x^4 \end{pmatrix}$$

Esto es, el producto se obtiene multiplicando los elementos de una fila de la primera matriz por los elementos correspondientes de cada columna de la otra y sumando los productos. Nótese que en la matriz resultado, cada elemento es una suma de productos. Observemos también que en la diagonal principal (de izquierda a derecha) tenemos una variable multiplicada por sí misma, esto es, las sumas de los cuadrados de las variables.

En estadística multivariada hay otra operación con matrices que tiene que realizarse con bastante frecuencia. Aunque afortunadamente los paquetes estadísticos realizan el cálculo sin que tengamos que preocuparnos por él. El Cuadro 1.2., que aparece en la página siguiente, presenta los pasos necesarios para invertir una matriz para el lector interesado.

Cuadro 1.2. Cálculo de la inversa de una matriz

En el cálculo matricial no existe la división de matrices. Si tenemos dos matrices, pongamos  $W$  y  $B$ , para dividir  $B$  entre  $W$  es necesario invertir primero  $W$ . La inversa de la matriz  $W$  es una nueva matriz, denominada como  $W^{-1}$ , que premultiplicada por la matriz original da como resultado la matriz unidad (con unos en la diagonal y ceros fuera de ella), esto es:  $I = W^{-1}W$ .

La matriz inversa se obtiene dividiendo la matriz de cofactores de la transpuesta de la original por el determinante de la original. Esto es:

$$W^{-1} = (W')^*/|W|$$

Por ejemplo, sea la matriz

$$W = \begin{pmatrix} 1 & 2 & 4 \\ 3 & 2 & 5 \\ 4 & 1 & 2 \end{pmatrix}$$

para hallar su inversa obtenemos primero su traspuesta, cambiando filas por columnas:

$$W' = \begin{pmatrix} 1 & 3 & 4 \\ 2 & 2 & 1 \\ 4 & 5 & 2 \end{pmatrix}$$

Después obtenemos la matriz de cofactores, teniendo en cuenta que el cofactor de un elemento es el menor complementario del mismo multiplicado por  $-1^n$  de fila + n de columna y que el menor complementario es el determinante de la submatriz que queda cuando se eliminan los elementos de la fila y columna correspondientes al dado). En el ejemplo, tendremos que:

$$(W')^* = \begin{pmatrix} -1 & 0 & 2 \\ 14 & -14 & 7 \\ 5 & 7 & -4 \end{pmatrix}$$

Y, puesto que el determinante de la matriz original es 7 [=  $(1*2*2)+(3*1*4)+(2*5*4)-(4*2*4)-(3*2*2)-(1*1*5)$ ], la inversa será:

$$W^{-1} = \begin{pmatrix} -1/7 & 0 & 2/7 \\ 2 & -2 & 1 \\ -5/7 & 1 & -4/7 \end{pmatrix}$$

y ahora sólo nos queda comprobar que, como ocurre de hecho,  $W^{-1}W = I$ . La inversa de una matriz puede obtenerse fácilmente en excel mediante la instrucción MINVERSA(w).

### 4.3.3. Matrices de covarianzas y de correlaciones

Si los elementos de la matriz resultado los dividimos por los grados de libertad (el número de productos sumados para obtener cada elemento menos 1), obtendremos una matriz de varianzas-covarianzas. Dividiendo por los productos de desviaciones típicas adecuados (los de las variables cuya correlación estamos calculando) podemos derivar la matriz de correlaciones. Siguiendo con nuestro ejemplo, ya sabemos que la matriz de sumas de cuadrados y productos cruzados es:

$$\begin{pmatrix} 52 & 56 & 30 & 32 \\ 49 & 64 & 30 & 28 \\ 30 & 36 & 18 & 18 \\ 32 & 32 & 18 & 20 \end{pmatrix}$$

Por tanto, puesto que hemos sumado dos productos para obtener esa matriz, la de covarianzas la obtendremos dividiendo por (2-1) cada elemento. Lógicamente eso implica que con dos sujetos la matriz de sumas de cuadrados y productos cruzados coincide con la de covarianzas. Cuando dividimos los elementos fuera de la diagonal principal (la que va desde la esquina superior izquierda hasta la inferior derecha) por la raíz cuadrada de los elementos situados en la diagonal correspondientes en fila y columna con el primero, obtenemos una matriz de correlaciones. Esto se debe a que los elementos fuera de la diagonal son covarianzas y a los situados en la diagonal son las varianzas correspondientes a cada variable. Por ejemplo, para convertir la covarianza entre la variable de la primera fila y la variable de la última columna (su valor es 32), tendremos que dividir por la raíz del producto de 52 (el elemento diagonal correspondiente a la fila) por 20 (el elemento diagonal correspondiente a la columna).

## 5. ORGANIZACIÓN DE LOS FICHEROS DE DATOS EN PAQUETES ESTADÍSTICOS

La estructura de los ficheros que contienen los datos de la investigación que serán analizados mediante algún paquete estadístico es extraordinariamente sencilla. Supongamos que hemos realizado un estudio para evaluar el efecto de la edad sobre el estado cognitivo de las personas. Se emplean dos grupos, uno de adultos jóvenes y otro de mayores y se les evalúa su rendimiento en una tarea de atención y otra de memoria. Supongamos además que hemos empleado cinco personas por grupo, y que los datos obtenidos han sido los siguientes:

SUJETOS	JÓVENES		MAYORES		
	A	M	SUJETOS	A	M
1	12	4	1	10	3
2	14	6	2	9	4
3	13	5	3	8	4
4	11	7	4	12	3
5	9	4	5	8	5

¿Cómo introduciríamos esos datos en un fichero para poder realizar los análisis adecuados? La respuesta a esta cuestión es muy simple. Los paquetes estadísticos utilizan una plantilla de hoja de cálculo para introducir los datos. Esas plantillas tienen una organización de tabla, en filas y columnas. La cues-

tión se reduce a tener presente que las filas están reservadas a los sujetos: cada fila es ocupada por un sujeto diferente. Las columnas están reservadas a las variables que hemos medido a los sujetos: una columna por variable. En nuestro ejemplo, debemos caer en la cuenta de que tenemos un total de 10 sujetos, 5 por grupo, por tanto el fichero tendrá 10 filas. Además, de cada sujeto conocemos tres cosas, el grupo al que pertenece (si es joven o mayor), su puntuación de atención (A) y su puntuación de memoria (M), por tanto, el fichero tendrá tres columnas. Su apariencia será, pues, la siguiente:

EDAD	A	M
1	4	12
1	6	14
1	5	13
1	7	11
1	4	9
2	3	10
2	4	9
2	4	8
2	3	12
2	5	8

Donde en la columna de edad 1 designa a jóvenes y 2 a mayores. El orden de las columnas es irrelevante, pero sí es absolutamente imprescindible que en cada columna se introduzca el tipo de datos adecuado. Para introducir estos datos en SPSS simplemente se tendría que crear una hoja de datos nueva, definir las variables y teclearlos en sus posiciones adecuadas. Es importante tener presente que prácticamente en todos los análisis estadísticos que se presentan en este manual, la organización de los ficheros será la descrita. El lector puede, por tanto, crear sus propios ficheros con los ejemplos que se presentan y seguir el curso del manual mientras realiza simultáneamente los análisis.

## 6. LECTURAS RECOMENDADAS

- AMÓN, J., *Introducción al análisis multivariante (Cálculo matricial)*, Barcelona, PPU, 1991.
- BISQUERRA, R., *Introducción conceptual al análisis multivariante*, Barcelona, PPU, 1989.
- CAMACHO, J., *Estadística con SPSS para Windows*, Madrid, Ra-Ma, 2000.
- HAIR, J. F. y cols., *Análisis multivariante*, Madrid, Prentice-Hall, 1998/1999.
- SÁNCHEZ-CARRIÓN, J. J., *Introducción a las técnicas de análisis multivariante aplicadas a las ciencias sociales*, Madrid, CIS, 1984.
- TABACHICK, B. G. y FIDELL, L. S., *Using multivariate statistics*, Boston, Allyn & Bacon, 2001.

## LECTURAS ADICIONALES

- ARNAU, J., *Diseños experimentales multivariados*, Madrid, Alianza, 1990.
- *Métodos y técnicas avanzadas de análisis de datos en ciencias del comportamiento*, Barcelona, UB, 1996.
- GRIMM, L. G. y YARNOLD, P. R., *Reading and Understanding Multivariate Statistics*, Washington, APA, 1995.
- JOHNSON, R. A. y WICHERN, D. W., *Applied Multivariate Statistical Analysis*, Prentice Hall, 1998.
- KACHIGAN, S., *Multivariate Statistical Analysis: A Conceptual Introduction*, Nueva York, Radius Press, 1991.
- NESSELROADE, J. R. y CATTELL, R. B., *Handbook of multivariate experimental Psychology*, Nueva York, Plenum Press, 1988.
- SHARMA, S., *Applied multivariate techniques*, Nueva York, Wiley, 1996.
- STEVENS, J., *Applied multivariate statistics for the social sciences*, Hillsdale, LEA, 1986.
- YOUNG, F. W., «Quantitative analysis of qualitative data», *Psicometrika*, 46, 1981, págs. 357-388.

## 7. EJERCICIOS

Seleccione la alternativa adecuada a cada una de las cuestiones siguientes. Medite su respuesta y trate de encontrar una justificación de la misma.

En un laboratorio de Psicología se intenta medir la función inhibitoria de la atención en niños. Los investigadores deciden emplear la tarea de interferencia Stroop en la que los niños tienen que indicar el color de la tinta en la que están escritas palabras que son nombre de color. Deciden introducir una condición congruente (por ejemplo, la palabra rojo escrita en tinta roja) y otra incongruente (por ejemplo, rojo escrito en tinta verde). Miden el tiempo que tardan en decir la tinta en la hoja de cada condición y el número de errores que cometen.

- 1) El contexto de investigación es:
  - a) experimental.
  - b) no experimental.
- 2) Una de las variables dependientes es:
  - a) la congruencia tinta-palabra.
  - b) el tiempo que se invierte en decir la tinta.
- 3) La variable dependiente errores está medida es una escala:
  - a) ordinal.
  - b) nominal.
  - c) de intervalo.
  - d) de razón.

- 4) La técnica más aconsejable para analizar los datos sería:
  - a) análisis univariado, dato que las variables dependientes no es esperable que correlacionen.
  - b) análisis multivariado, porque las variables dependientes pueden correlacionar.
- 5) Si la técnica que empleamos para hacer el análisis es MANOVA, podemos decir que:
  - a) hacemos un contraste simultáneo de las variables dependientes.
  - b) hacemos un contraste variable a variable.
- 6) ¿Cree que MANOVA sería apropiado dadas las características del estudio?
  - a) Sí, porque, tenemos una variable independiente, medida en escala nominal, y dos dependientes métricas.
  - b) No, porque MANOVA no es aplicable cuando se manipulan variables nominales.
- 7) Multiplique esta matriz por su traspuesta, obtenga también la matriz de sumas de cuadrados y productos cruzados, la de varianzas-covarianzas, y la de correlaciones:

$$\begin{pmatrix} 4 & 5 \\ 3 & 4 \\ 6 & 3 \end{pmatrix}$$

## RESPUESTAS A LOS EJERCICIOS

- 1) a)
- 2) b)
- 3) d)
- 4) b)
- 5) a)
- 6) a)
- 7)

Sumas de Cuadrados y Productos	Covarianzas	Correlaciones
41 32 32 25	20,5 16 16 12,5	1 0,9995 0,9995 1

## CAPÍTULO II

### Análisis univariado de varianza

#### OBJETIVOS DE APRENDIZAJE

- 1) Conocer los contextos de investigación en que son aplicables los análisis de varianza y de covarianza (ANOVA y ANCOVA).
- 2) Conocer las preguntas de investigación que estos análisis ayudan a responder.
- 3) Conocer las limitaciones y supuestos de ANOVA y ANCOVA.
- 4) Conocer los principales diseños de investigación univariados.
- 5) Comprender la lógica subyacente a los análisis de los diferentes diseños.
- 6) Conocer los aspectos fundamentales del cálculo de los efectos producidos por las variables independientes y sus interacciones.
- 7) Saber analizar las interacciones en términos de efectos principales y de interacciones parciales (Contrastes).
- 8) Saber interpretar los resultados de los análisis en los diferentes diseños.

No corremos mucho riesgo de equivocarnos si comenzamos afirmando que una parte sustancial de los resultados de las investigaciones actuales es analizada mediante análisis univariados, aunque de hecho estén planificadas inicialmente como investigaciones multivariadas. La importancia del análisis univariado de varianza en muchos campos de la Psicología es tan extraordinaria que podríamos considerarlo como la técnica estándar de la investigación. Una gran mayoría de las investigaciones actuales incluyen una manipulación de una o más variables independientes, y miden su efecto sobre una o más variables dependientes. Por tanto, el interés de estas investigaciones reside, fundamentalmente, en determinar si la manipulación produce o no efecto. Esto es, en contrastar

tar hipótesis. Es justamente en este marco de contraste de hipótesis donde el análisis univariado de varianza (ANOVA) es aplicable, pero eso no implica necesariamente que sea la técnica de elección, la más apropiada. Hay una buena cantidad de técnicas que están orientadas justamente en esa misma dirección, que son apropiadas para contrastar hipótesis sobre los efectos de las manipulaciones. Afortunadamente cada día es mayor el número de investigadores que son conscientes de este hecho, pero aún queda bastante camino por recorrer. Ésta es una de las razones fundamentales para que incluyamos un capítulo de análisis univariado en un manual dedicado al análisis multivariado. Sin embargo, no es la única. Consideramos que este capítulo puede ayudar a situar al lector en el marco conceptual mínimo necesario para comprender el resto de las técnicas que aquí se presentan. En las páginas que siguen realizaremos primero un repaso de conceptos básicos de inferencia estadística, para adentrarnos después en los principales diseños de investigación, su análisis global y pormenorizado, para concluir haciendo referencia a las limitaciones de los análisis y la secuencia de actividades característica de estas investigaciones.

## 1. LA INFERENCIA ESTADÍSTICA Y EL CONTRASTE DE HIPÓTESIS

No hay investigación científica en la que no se trate de resolver un problema, en la que no se aborden fenómenos cuya explicación no está clara, en los que se necesite adquirir conocimiento acerca de cuáles son sus posibles causas. Por lo general ese problema está formulado de manera explícita en forma de una pregunta en la que se establece una conexión, sea causal o no, entre variables. Las preguntas que los investigadores realizamos pueden ser más o menos interesantes, dependiendo del conocimiento que tengamos del dominio teórico en el que nuestra investigación se enmarca. Sin embargo, todas admiten una respuesta. Por ejemplo, podemos preguntarnos si una terapia que combine componentes psicológicos y farmacológicos (A1) puede ser diferente de una terapia meramente farmacológica (A2) para tratar problemas de depresión. Esta pregunta es, claramente, de naturaleza causal, puesto que estamos sugiriendo que un tipo de terapia produce mejorías más importantes que la otra. Además, tiene dos respuestas posibles: no es cierto, las dos terapias son iguales o sí es cierto, las dos son diferentes. Ahora lo importante es que esas dos respuestas son suposiciones previas al experimento que podemos caracterizar como hipótesis. La primera es una hipótesis sobre igualdad, mientras que la segunda es una hipótesis sobre diferencias. Estas hipótesis no son contrastables sin más. Para decidir entre ellas necesitamos poder traducirlas a términos estadísticos. La traducción se realiza de manera automática apenas decidimos cómo puede evaluarse cada terapia. La forma más frecuente pasa, sin duda, por suponer que las medias de tratamiento reflejarán el posible efecto de la manipulación. Por otra parte, nuestro interés no reside en si los sujetos que han recibido

do las dos terapias difieren, sino en si esa diferencia es suficiente para pensar que en la población las cosas serían de la misma manera. Es decir, lo que nos interesa no es establecer hipótesis sobre muestras o grupos de personas, sino sobre poblaciones de personas. Por tanto, las hipótesis estadísticas hacen alusión a medias poblacionales, no a medias de muestras. De hecho, es casi seguro que si un grupo de pacientes recibe el tratamiento A1 y otro el A2, las medias de grupo no serán iguales. Lo que nos interesa es poder decidir hasta qué punto si en lugar de usar esos sujetos empleásemos otros diferentes obtendríamos una diferencia como la observada. El contraste de hipótesis sirve precisamente para eso, para poder establecer si la diferencia observada se debe a diferencias entre poblaciones (y entonces podría observarse con sujetos diferentes) o es producto del azar. En definitiva, tendremos una hipótesis, la nula, que afirma que las medias poblacionales son iguales, y otra, la alternativa, que establece que no lo son.

¿Cómo se realiza el contraste de hipótesis? En situaciones de investigación como la descrita más arriba el contraste se realiza calculando un estadístico, un valor, cuyas características de distribución sean conocidas. El estadístico más empleado cuando, como en el ejemplo, tenemos dos grupos que reciben tratamientos diferentes, es la *t* de Student. Este estadístico se computa a partir de los datos del experimento de una forma muy simple. Basta con aplicar la siguiente ecuación:

$$t = \frac{\bar{X}_{A1} - \bar{X}_{A2}}{\sqrt{\frac{s_{A1}^2}{n_{A1}} + \frac{s_{A2}^2}{n_{A2}}}} \quad (\text{Ec. 2.1})$$

donde en el numerador tenemos las medias de tratamiento y en el denominador los errores de cada uno de ellos. Por ejemplificar, supongamos que hemos aplicado cada tratamiento a un grupo de 5 sujetos y que hemos usado el inventario de Beck (BDI). Supongamos que los resultados obtenidos tras la aplicación del tratamiento fueron los que aparecen en la tabla siguiente, en la que también aparecen las medias y varianzas ( $s^2$ ) de grupo:

Tabla 2.1

A1	B1
14	17
16	23
12	19
14	16
9	20
Media	13
Varianza ( $s^2$ )	7
	7,5

Ahora  $t$  es computable aplicando directamente la ecuación 2.1:

$$t = \frac{13 - 19}{\sqrt{\frac{7}{5} + \frac{7,5}{5}}} = -3,52$$

Una vez obtenido el estadístico de contraste el paso siguiente consiste en tomar una decisión sobre si su valor es indicativo de que hay diferencias poblacionales o no. Para ello, necesitamos conocer qué valores pueden aparecer cuando no hay diferencias poblacionales. Esto es, su distribución cuando la hipótesis nula es verdadera. Esta distribución depende de los grados de libertad (gl) de la  $t$  de Student, que son el número total de observaciones, menos el número total de restricciones. Las restricciones son dos, las medias de tratamiento. Por tanto, en general tendremos  $2n-2$  grados de libertad. En nuestro ejemplo  $2(5)-2 = 8$  gl. Puesto que nuestra hipótesis no establecía claramente si las diferencias entre A1 y A2 serían positivas o negativas, debemos realizar un test de dos colas. En la tabla de  $t$  podemos encontrar que la probabilidad de obtener un valor de  $t$  mayor o igual (en valores absolutos) que el observado en el estudio es menor de 0,025, por lo que rechazamos la hipótesis nula. Esto nos lleva a creer verosímil que los tratamientos son diferentes entre sí. La cuestión de cuál de ellos es el más adecuado es soluble simplemente examinando el signo de la  $t$ , o las medias y comprobando que la de A1 es menor que la de A2.

¿Por qué funciona este planteamiento tan sencillo de toma de decisiones sobre hipótesis? Esta pregunta podemos replantearla de otro modo: ¿qué reflejan las diferencias entre las medias y qué reflejan las varianzas de tratamiento? Comencemos por las medias. Parece razonable pensar que la puntuación de cada sujeto del estudio sea debida a la contribución de dos componentes. Es evidente que uno de ellos debe ser el tratamiento que el sujeto recibe. Y también debe ser evidente que el resto de cosas son irrelevantes para nosotros, son error, entre ellas está claro que deben incluirse las características del propio sujeto, personales, sociales, etc. Si asumimos que el efecto que produce el tratamiento es el mismo en todos los sujetos que lo reciben, y que su contribución se suma a la del error, tendremos que:

$$\text{Puntuación}_{ij} = \text{Tratamiento } j + \text{Error}_{ij}$$

De forma que la media del grupo, que obtenemos sumando todas las puntuaciones del grupo y dividiendo por el total de sumandos, contendrá efecto del tratamiento y error. Por tanto, la diferencia de medias se debe a las diferencias entre los tratamientos y las diferencias en los errores de grupo. Esto es,

$\text{Diferencia de Medias} = \text{Diferencia entre los efectos de los Tratamientos} + \text{Diferencia Errores}$
---

Por otra parte, la varianza intragrupo se calcula también a partir de diferencias, aunque en este caso entre las puntuaciones de los sujetos y la media del grupo correspondiente. Por lo que llevamos dicho, estará claro que en cada una de estas diferencias no puede haber influencia del tratamiento, puesto que todos los sujetos reciben el mismo y su efecto es independiente de los sujetos, y, por tanto, sólo pueden ser debidas a los sujetos, esto es, sólo contienen error. Es decir:

$\text{Puntuación}_{ij} - \text{Media}_j = (\text{Ef. Del Tratamiento } j + \text{Error}_{ij}) - (\text{Ef. Del Tratamiento } j + \text{Error}_j)$
--

donde  $\text{Error}_j$  es el error promedio del grupo. La  $t$  de Student es, pues, un cociente entre dos tipos de diferencias, las del numerador, que se deben a los efectos y al error, y las del denominador, que contienen sólo error. Obviamente, cuanto más se aleje, en valores absolutos, la  $t$  de 1, más evidencia aporta a favor de la existencia de diferencias entre los tratamientos que se comparan. Es importante caer en la cuenta, sin embargo, de que puede producirse cualquier valor de  $t$  aun cuando los tratamientos sean iguales (las poblaciones lo sean). Justamente esto es lo que hace necesario considerar la probabilidad de que el valor de  $t$  obtenido en un estudio sea debido a la hipótesis nula. Habitualmente, si la probabilidad de obtener una  $t$  como la observada o mayor es 0,05 o menor para una cola (si la hipótesis incluye el signo de la diferencia; 0,025 si es de dos colas, la hipótesis sobre la diferencia no incluye signo), se rechaza la hipótesis nula. Antes de finalizar este punto es importante realizar una consideración adicional. Algunos investigadores tienden a creer que los valores de probabilidad extraordinariamente bajos son indicadores de que tienen un gran tamaño de efecto, de que su variable produce grandes efectos. Eso no es cierto, si se pretende conocer el tamaño del efecto es preferible utilizar estadísticos adecuados, del tipo de los que presentaremos más adelante.

La  $t$  de Student que acabamos de presentar es utilizable cuando los tratamientos se aplican a grupos diferentes de sujetos, a muestras independientes. ¿Qué podemos hacer para contrastar hipótesis cuando los tratamientos se aplican a un mismo grupo de sujetos en momentos diferentes del tiempo, esto es, cuando las muestras están correlacionadas? Podemos seguir usando la  $t$  de Student, sin embargo, su ecuación de cálculo debe adaptarse para tener en cuenta que las puntuaciones correspondientes a uno y otro tratamiento pueden estar correlacionadas. Para hacer los cálculos nos basta con calcular las diferencias sujeto a sujeto. La media de las diferencias será el numerador, como en el caso anterior, y el error típico de la diferencia será el denominador. Esto es:

$$t = \frac{\bar{X}_{A1} - \bar{X}_{A2}}{\sqrt{\frac{s_D^2}{n}}} \quad (\text{Ec. 2.2})$$

La t de Student que acabamos de presentar es equivalente a otro estadístico muy empleado en el contexto de análisis de varianza, la F de Snedecor. La F del experimento puede calcularse, cuando hay dos grupos, simplemente elevando la t al cuadrado. Eso implica que el numerador de la F ya no contiene una diferencia, sino una varianza debida a los efectos, mientras que el denominador contiene varianza error.

## 2. DISEÑOS EXPERIMENTALES UNIVARIADOS

Los diseños experimentales clásicos están orientados al contraste de hipótesis sobre los efectos de una o más variables independientes. Las diferencias entre ellos pueden establecerse en función de las características que esas variables presenten en cada estudio en particular. Especialmente relevantes son la forma de manipular la variable y el número de variables. Las variables independientes directas pueden manipularse de dos maneras bien diferenciadas, que suelen denominarse como entre grupos (también entre o intersujetos) e intra (dentro de) sujetos. En la manipulación entre grupos cada sujeto recibe sólo un nivel de la variable, mientras que en la intrasujetos cada uno recibe todos los niveles. Esta diferencia, como hemos podido comprobar más arriba, resulta fundamental, puesto que la correlación de las muestras debe ser tenida en cuenta en el cálculo del error en el segundo, pero no en el primer caso.

El segundo criterio importante es el número de variables independientes. Cuando en el diseño se incluye sólo una variable, los diseños pueden ser univariados entre grupos o intrasujetos. Cuando el diseño incluye más de una variable independiente pueden ocurrir varias cosas. En primer lugar que se obtengan todas las combinaciones posibles de los distintos niveles de las variables. Por ejemplo, si tenemos dos variables, manipuladas ambas a tres niveles, las combinaciones posibles serán nueve, puesto que cada nivel de la primera puede combinarse con cada nivel de la segunda. Si etiquetamos como A y B a las variables, a los niveles de A como  $a_1, a_2$  y  $a_3$ , y a los de B como  $b_1, b_2$  y  $b_3$ , entonces el cruce factorial sería el que aparece en la siguiente tabla:

	b1	b2	b3
a1			
a2			
a3			

donde las casillas representan las combinaciones posibles, que llamaremos condiciones experimentales. Una segunda posibilidad, casi siempre indeseable, es que las variables sean mezcladas por confusión. Por ejemplo, si aplicamos  $a_1b_1$  a unos sujetos,  $a_2b_2$  a otros y  $a_3b_3$  a otros, el problema más importante que se plantea es la imposibilidad de discernir si las diferencias entre unos sujetos y otros se deben a una variable, a la otra o a ambas a la vez. La tercera posibilidad es que una variable no se repita dentro de la otra, sino que sus niveles cambien cuando cambiamos de nivel en la primera. Supongamos que A tiene dos niveles y B cuatro, entonces podríamos mezclar  $a_1b_1, a_1b_2, a_2b_3$  y  $a_2b_4$ . Este tipo de diseños se denominan anidados y suelen aparecer en investigaciones realizadas en ambientes aplicados, donde las variables no se pueden manipular de forma arbitraria. Por ejemplo, si A son colegios y B son profesores, es obvio que los profesores no pueden ser trasladados de colegio para impartir docencia. Dadas las dificultades de analizar e interpretar los diseños anidados y de confusión, lo más frecuente es encontrar investigaciones en las que las variables se han cruzado factorialmente, esto es, que emplean diseños factoriales. En las páginas siguientes presentaremos el análisis de los diseños de uno o más factores de forma detenida.

### 2.1. DISEÑOS UNIFACTORIALES UNIVARIADOS

Cuando se manipula una sola variable independiente a varios niveles pueden emplearse dos tipos de diseños diferentes, según se manipule entre grupos o intrasujetos. En el primer caso el diseño más característico es el conocido como completamente aleatorizado. En el segundo, el de medidas repetidas. A continuación presentaremos brevemente el análisis global de cada uno de estos dos diseños, además de su análisis detallado a través de contrastes.

#### 2.1.1. Diseño completamente aleatorizado

La característica más definitoria de este diseño es que los sujetos son asignados al azar a los grupos de tratamiento. Por tanto, cada sujeto recibe uno y sólo uno de los niveles de la variable independiente. Para cada sujeto tendremos una sola medición en la variable dependiente, de modo que en total habrá tantas observaciones como sujetos. Por ejemplo, consideremos que además de emplear dos terapias de la depresión hubiésemos utilizado un grupo de control, que no recibe tratamiento alguno (A3). Tendremos, pues, una variable independiente, el tratamiento, con tres niveles, A1, A2 y A3. Si disponemos de un total de 15 sujetos, cinco serían asignados al azar al tratamiento A1, otros 5 al A2 y otros 5 al A3. A cada sujeto le hemos medido la variable dependiente, usando BDI, después del período de terapia. En diseños de este tipo, la mejor pregunta de partida es: ¿cómo podemos explicar la puntuación obtenida por

cada sujeto en la variable dependiente? Como hemos visto antes, la puntuación se obtiene sumando dos componentes: efecto del tratamiento y error. Si expresamos la puntuación en términos de diferencias, tendremos:

$$\text{Puntuación}_{ij} - \text{Media total} = (\text{Media}_j - \text{Media total}) + (\text{Puntuación}_{ij} - \text{Media}_j)$$

Obsérvese que una forma de conceptuar el error consiste en considerar que contiene todo lo que no es efecto del tratamiento. Estas diferencias pueden transformarse en variabilidades de una forma muy sencilla, elevando al cuadrado y sumando para todos los sujetos de todos los grupos (puesto que la correlación tratamiento-error es nula). Es decir:

$$\sum_i \sum_j (\text{Puntuación}_{ij} - \text{Media tot})^2 = \sum_j n_j (\text{Media}_j - \text{Media tot})^2 + \sum_i \sum_j (\text{Puntuación}_{ij} - \text{Media}_j)^2$$

Que es la ecuación fundamental del diseño, expresada normalmente como:

$$SC_{\text{TOTAL}} = SC_{\text{EG}} + SC_{\text{IG}}$$

La suma de cuadrados total (variabilidad total) es obtenida sumando la suma de cuadrados entre grupos (variabilidad debida al tratamiento) y la suma de cuadrados intragrupo (la variabilidad error). Recordemos que el objetivo es tomar una decisión sobre el efecto del tratamiento, y que para eso se necesita un estadístico de contraste. El estadístico por defecto en este caso es la F de Snedecor, que, como hemos visto, es un cociente entre varianza de tratamiento y varianza error. Por tanto, necesitamos transformar las sumas de cuadrados en varianzas (que son denominadas medias de cuadrados y denotadas como MC). Las medias de cuadrados se obtienen dividiendo las sumas de cuadrados por sus grados de libertad correspondientes. Puesto que los grados de libertad son número de sumandos menos número de restricciones, nos basta mirar las ecuaciones anteriores para darse cuenta de que la  $SC_{\text{EG}}$  tiene tantos sumandos como grupos tenemos ( $a$  grupos) y que la única restricción es la media total. Por tanto, sus grados de libertad serán  $a-1$ . Es decir:

$$MC_{\text{EG}} = \frac{SC_{\text{EG}}}{a-1} \quad (\text{Ec. 2.3})$$

Por otra parte, la  $SC_{\text{IG}}$  está compuesta por tantos sumandos como sujetos hay en el experimento (esto es  $a \cdot n$ , puesto que tenemos  $a$  grupos y  $n$  sujetos por grupo). El número de restricciones es obvio, puesto que necesitamos conocer cada una de las medias de grupo para poder computar las diferencias. Por tanto, tendremos tantas restricciones como grupos. En resumen, los grados de libertad del error serán:  $a \cdot n - a$ , o mejor aún  $a(n-1)$ . Luego, la MC de error será:

$$MC_{\text{IG}} = \frac{SC_{\text{IG}}}{a(n-1)} \quad (\text{Ec. 2.4})$$

El cociente entre ambas sumas de cuadrados es la F del experimento:

$$F = \frac{MC_{\text{EG}}}{MC_{\text{IG}}} \quad (\text{Ec. 2.5})$$

La distribución de F cuando la hipótesis nula es verdadera depende de los grados de libertad del numerador y del denominador.

Consideremos los datos que aparecen en la Tabla 2.2 que contiene puntuaciones ficticias de 15 sujetos, 5 por grupo de tratamiento.

Tabla 2.2

	A1	A2	A3
14	17	22	
16	23	23	
12	19	25	
14	16	24	
9	20	26	
Media	13	19	24
Varianza ( $s^2$ )	7	7,5	2,5

El cálculo de las sumas de cuadrados es muy sencillo. Para empezar, el cálculo de la  $SC_{\text{EG}}$  requiere primero computar la media total, que será  $(13+19+24)/3 = 18,67$ , es decir, la media de las medias de grupo. Ahora tendremos que:

$$SC_{\text{EG}} = 5 * [(13-18,67)^2 + (19-18,67)^2 + (24-18,67)^2] = 303,33$$

Por su parte, la  $SC_{\text{IG}}$ , se obtiene como:

$$SC_{\text{IG}} = [(14-13)^2 + (16-13)^2 + \dots + (20-19)^2 + \dots + (26-24)^2] = 68,$$

Luego  $SC_T = 303,33 + 68 = 371,33$ . Teniendo en cuenta que los grados de libertad entre grupos son  $(3-1)$  y los del intragrupo  $3(5-1)$ , las medias de cuadrados serán:

$$MC_{\text{EG}} = 303,33/2 = 151,67,$$

y

$$MC_{\text{IG}} = 68/12 = 5,67.$$

Nótese que la misma  $MC_{IG}$  se obtendría si sumamos las varianzas de grupo y dividimos por el número de grupos. Esto es:

$$MC_{IG} = \sum s_j^2 / a = (7 + 7,5 + 2,5) / 3 = 5,67$$

Lo que indica claramente que la  $MC_{IG}$  es simplemente la media de los errores de grupo. Esta cuestión es relevante para comprender uno de los supuestos más importantes del análisis de varianza en estos diseños, la homogeneidad de las varianzas error.

Finalmente, la F del experimento será el cociente entre las varianzas entre grupos e intragrupo:  $F = MC_{EG}/MC_{IG} = 151,67 / 5,67 = 26,76$ . Consultando la tabla de F de Snedecor comprobaremos que la probabilidad de obtener un valor de F igual o mayor que el obtenido, con 2 y 12 gl, es menor de 0,05, por lo que rechazamos la hipótesis nula. Podemos concluir que los tratamientos parecen producir resultados diferenciales. Ésta es una afirmación concreta cuando tenemos dos tratamiento (obviamente uno de ellos ha de producir mejores resultados que el otro), pero es ambigua cuando tenemos, como en este ejemplo, más de dos tratamientos.

#### 2.1.1.1. «Contrastes planeados»

La pregunta clave ahora es: ¿qué tratamientos difieren de qué otros? La respuesta puede obtenerse de varias maneras diferentes. Si antes de realizar el experimento se tenían hipótesis claras sobre las diferencias esperables, entonces pueden emplear comparaciones planeadas para contrastarlas. Por ejemplo, si pensábamos que recibir algún tratamiento (A1 y/o A2) sería mejor que no recibir ninguno (A3), y creímos también que recibir un tratamiento combinado (A1) sería mejor que recibir sólo un tratamiento farmacológico (A2), las comparaciones serían las siguientes:

*Primera comparación:* Es obvio que A1 y A2 se agrupan juntos frente a A3, por tanto, se deberían combinar sus datos y compararlos con los de A3. En términos de hipótesis estadísticas esto sería:

$$(\mu_{A1} + \mu_{A2})/2 - \mu_{A3} = 0; \text{ o lo que es lo mismo}$$

$$(\mu_{A1} + \mu_{A2}) - 2\cdot\mu_{A3} = 0, \text{ lo que puede escribirse como}$$

$1\cdot\mu_{A1} + 1\cdot\mu_{A2} - 2\cdot\mu_{A3} = 0$ , y, por tanto, la comparación queda expresada cuando conocemos los coeficientes de cada una de las medias de grupo: {1,1,-2}.

*Segunda comparación:* En este caso, A3 se descarta, por consiguiente:

$$\mu_{A1} - \mu_{A2} = 0; \text{ y esto es lo mismo que}$$

$$1\cdot\mu_{A1} - 1\cdot\mu_{A2} + 0\cdot\mu_{A3} = 0, \text{ luego la comparación es } \{1, -1, 0\}$$

Estas comparaciones se llaman planeadas o *a priori* por que están planificadas antes de realizar el experimento. La única condición que deben cumplir los coeficientes para que pueda definirse una comparación es que su suma sea cero (teniendo en cuenta los sujetos de cada grupo), esto es:

$$\sum n_j c_j = 0 \quad (\text{Ec. 2.7})$$

Su análisis es muy sencillo, puesto que puede realizarse siguiendo el mismo esquema del análisis global de varianza. Esto es, se trataría de obtener una varianza de los efectos (de la comparación) en el numerador y emplearíamos la varianza error del análisis global en el denominador. Por tanto, el problema se reduce a calcular la suma de cuadrados de la comparación,  $SC_j$ , que obtendremos aplicando la siguiente expresión:

$$SC_j = \frac{\left( \sum_{j=1}^p n_j c_j \bar{X}_j \right)^2}{\sum_{j=1}^p n_j c_j^2}, \quad (\text{Ec. 2.8})$$

lo que en el caso de la primera comparación quedaría como:

$$SC_{\varphi_1} = \frac{(1\cdot13 + 1\cdot19 - 2\cdot24)^2}{5\cdot1^2 + 5\cdot1^2 + 5\cdot(-2)^2} = 213,33$$

La suma de cuadrados de la segunda sería:  $SC_{\varphi_2} = 90$ . Obsérvese que las dos sumas de cuadrados de comparación sumadas equivalen a la  $SC_{EG}$ , lo que ocurrirá siempre que las comparaciones sean ortogonales (no correlacionen entre sí). La ortogonalidad se define siempre sobre un par de comparaciones. Se produce cuando la suma de los productos de los coeficientes de ambas comparaciones sea nula. Esto es:

$$\sum n_j c_j n_j c_j = 0 \quad (\text{Ec. 2.9})$$

lo que ocurre en las dos comparaciones que estamos considerando:

Comparaciones	A1	A2	A3	Sumas
Primera	5·1	5·1	5·(-2)	0
Segunda	5·1	5·(-1)	5·0	0
Productos: 1. <sup>a</sup> , 2. <sup>a</sup>	25	-25	0	0

La ortogonalidad no es una condición necesaria de las comparaciones planeadas, pero tiene la ventaja de que reduce el número de comparaciones que es preciso emplear para explicar los efectos observados.

El resto del análisis es muy simple, puesto que estas comparaciones tienen sólo un grado de libertad, y, por tanto, las medias de cuadrados de la comparación coinciden con las sumas de cuadrados. Finalmente, la F de la comparación es el cociente entre la media de cuadrados de la comparación y la media de cuadrados del error calculado en el análisis global. En nuestro ejemplo, tendremos que:

$$F_{\varphi 1} = \frac{MC_{\varphi 1}}{MC_{IG}} = \frac{213,33}{5,67} = 37,64$$

Consultando en la tabla de F de Snedecor, con 1 y 12 gl en numerador y denominador respectivamente, comprobaremos que la probabilidad de obtener un valor de F mayor o igual que el observado es menor que 0,05, por lo que rechazamos la hipótesis nula, y, por tanto, consideraremos que recibir algún tratamiento es diferente a no recibir ninguno. El lector puede realizar la segunda comparación para comprobar que su F es 15,88, y que no produce el mismo efecto el tratamiento combinado que el tratamiento simple.

Cuando la variable independiente no es categórica (como lo es en el ejemplo), sino métrica, es frecuente que el investigador tenga interés en realizar comparaciones de un tipo especial (en cuanto a significado), denominadas tendencias o polinomiales. Imaginemos, por exemplificar, que nuestro interés hubiese sido comprobar cómo mejoran los pacientes a medida que incrementa el número de días de terapia. En ese caso podríamos manipular la variable independiente también entre grupos a varios niveles, pongamos a 5, 15 y 25 días de terapia. Una pregunta importante sería cuál es la relación funcional entre variable independiente y variable dependiente. Por ejemplo, ¿a medida que el número de días de terapia incrementa, se produce una mejoría de los pacientes? ¿Se produce una mejora hasta cierto número de días y a partir de ahí ya no se producen cambios en los pacientes? En el primer caso estaríamos suponiendo que la relación es lineal (más días, más mejora), y en el segundo que la relación, además de lineal es cuadrática (hay un punto de cambio en la relación). La respuesta a preguntas de este tipo puede obtenerse realizando análisis de tendencias. Este análisis se realiza exactamente igual que el de comparaciones, exceptuando que los coeficientes deben ser computados (o seleccionados de una tabla) según la relación que pretenda contrastarse. En este sentido es importante tener en cuenta que las tablas de coeficientes son útiles cuando la distancia entre dos niveles consecutivos cualesquiera de la variable independiente es constante. Esto es, la distancia entre el primer nivel (5 días) y el segundo (15 días) es la misma que entre el segundo y el tercero (25 días), y que el número de sujetos por grupo es el mismo. Los paquetes estadísticos como SPSS tie-

nen esto en cuenta, siempre y cuando en la definición de la variable de agrupamiento se indiquen los valores reales de los niveles.

### 2.1.1.2. «Contrastes no planeados»

En bastantes ocasiones el investigador no puede planificar las comparaciones debido a que no tiene información suficiente sobre cuál puede ser el efecto de su variable independiente. En estos casos, la ausencia de hipótesis previas obliga a que se realicen todas las comparaciones posibles entre los niveles de la variable independiente. Este hecho, puede parecer irrelevante, pero, sin embargo, resulta fundamental respecto de la ejecución de los contrastes. Caigamos en la cuenta, para empezar, de que el número de comparaciones posibles puede ser muy elevado. Por ejemplo, con tres niveles, las comparaciones entre pares de tratamiento son  $3*(3-1)/2 = 3$  (tratamiento 1 con tratamiento 2, tratamiento 1 con 3, tratamiento 2 con 3). Con cuatro niveles, el número sube a  $4*(4-1)/2 = 6$ , y así sucesivamente. En general el número de comparaciones entre pares de tratamientos es:

$$c = \frac{a(a - 1)}{2}, \quad (\text{Ec. 2.10})$$

donde a es el número de niveles de la variable independiente y c el número de comparaciones. Si asumimos que en cada comparación tomamos una decisión con un riesgo de error  $\alpha$ , entonces la probabilidad de acertar en cada una será  $1-\alpha$ , y la probabilidad de que todas las decisiones sean correctas será  $(1-\alpha)^c$ , por tanto, la probabilidad de que alguna será errónea será  $1-(1-\alpha)^c$ . Esta tasa de error tipo I a lo largo del experimento ( $\alpha_E$ ), será obviamente mayor cuanto mayor sea el número de comparaciones que se realicen. Así, si realizamos 3 comparaciones, con un riesgo de 0,05, tendremos que  $\alpha_E=1-(1-\alpha)^3 = 0,1426$ , lo que está claramente por encima del límite de error admitido de forma estándar del 0,05. Cuando c es pequeño, la expresión anterior es prácticamente equivalente a la siguiente desigualdad:  $\alpha_E \leq c\alpha$ , conocida como desigualdad de Bonferroni. Por tanto, una estrategia posible para realizar las comparaciones será ejecutarlas como si fuesen planeadas, pero utilizar un  $\alpha=c/\alpha_E$  (derivando a partir de Bonferroni), o bien  $\alpha=1-(1-\alpha_E)^{1/c}$  (corrección de Sidak, si pretendemos mayor exactitud) donde  $\alpha_E$  se ha fijado en 0,05 o menos.

Lo que pone de manifiesto Bonferroni es que no es posible (o al menos no se debe) emplear un procedimiento de análisis que no tenga en cuenta el hecho de que el número de comparaciones incrementa la tasa de error tipo I en el conjunto del experimento. El número de procedimientos disponibles que teóricamente permiten controlar el crecimiento de esta tasa de error es bastante grande. SPSS incluye 14 pruebas que asumen igualdad de varianzas, y 4 que

asumen desigualdad. En principio es posible usar cualquiera de ellas, sin embargo, debe tenerse en cuenta que no todas tienen la misma potencia ni protegen igual de bien frente al crecimiento del error tipo I. Por ejemplo, la prueba DMS (diferencia mínima significativa de Fisher), basada en la t de Student, no introduce ningún control del error alfa, que puede crecer libremente hasta límites intolerablemente altos. Algo similar ocurre con la prueba S-N-K (Student-Newman-Keuls), basada en el estadístico q de rango estudiantizado, cuya protección es menor de lo que sería recomendable. Las pruebas, REGWF y REGWQ (Ryan-Einot-Gabriel-Walsh, basadas en la F o en la q) y la de Duncan son similares a la prueba SNK, aunque protegen levemente mejor que ésta. La prueba de Tukey, conocida también como HSD (diferencia honestamente significativa), asegura que la tasa global de error no excederá la fijada por el investigador (0,05, por ejemplo). Resultados similares a los de Tukey proporcionan pruebas como la GT2 de Hochberg o la prueba de Gabriel, aunque ésta última corrige menos de lo necesario cuando el número de sujetos es muy desigual de un grupo a otro. Las pruebas de Bonferroni y Sidak usan las correcciones que hemos mencionado más arriba. Respecto de las que asumen varianzas desiguales, la más recomendable, debido a su potencia y flexibilidad es la de Games-Howell, el resto deben emplearse solamente cuando el control de la tasa de error sea absolutamente prioritario (véase Glosario para una descripción más detallada).

#### 2.1.1.3. «Magnitud de los efectos»

Los contrastes anteriores nos sirven para decidir si la variable independiente produce efectos diferenciales o no, pero no son útiles para evaluar la importancia que la variable independiente tiene para explicar la variabilidad que se observa en la variable dependiente. Erróneamente algunos investigadores tienden a utilizar el nivel de significación como un indicio de la magnitud del efecto que la independiente produce sobre la dependiente. Sin embargo, esta estrategia es inadecuada porque, entre otras cosas, el nivel de significación depende del número de sujetos por grupo y del número de niveles del factor, de manera que un mismo tamaño de efecto tiene asociada una significación mayor cuando se incrementa el número de sujetos o de variables. Por tanto, es preferible utilizar índices que sean independientes de estos factores. Se han propuesto varios índices, entre los que destacan  $R^2$  (o  $\eta^2$ ),  $R^2_p$  (coeficiente de determinación parcial),  $\omega^2$  (omega cuadrado) y el coeficiente de correlación intraclass.

El coeficiente  $R^2$  es un cociente entre la variabilidad del efecto y la variabilidad total, por tanto, indica la proporción de la variabilidad total de los datos que se debe a la variable independiente.  $R^2$  se computa como  $SC_{EG}/SC_{TOTAL}$ . Uno de los problemas más importantes que tiene este índice es que tiende a disminuir a medida que incrementamos el número de variables independientes, razón por la que se introdujo el índice  $R^2_p = SC_{Efecto}/(SC_{Efecto} + SC_{Error})$ .

Naturalmente, cuando sólo hay una variable independiente los dos coinciden exactamente. En nuestro ejemplo,  $R^2 = 303,33/371,33 = 0,82$ , lo que indica que el 82 por 100 de la variabilidad total se debe a los tratamientos.

El coeficiente  $\omega^2$  puede interpretarse de la misma forma que los anteriores, pero difiere de ellos en que tiene en cuenta el tamaño del error. Su ecuación es:

$$\omega^2 = \frac{SC_{EG} - (a - 1)MC_{IG}}{SC_T + MC_{IG}} = \frac{303,33 - (3-1)5,67}{371,33 + 5,67} = 0,83 \quad (\text{Ec. 2.11})$$

Finalmente, el coeficiente de correlación intraclass se computa como:

$$\rho^2 = \frac{MC_{EG} - MC_{IG}}{MC_{EG} + (a - 1)MC_{IG}} = \frac{151,67 - 5,67}{5,67 + (3-1)5,67} = 0,90 \quad (\text{Ec. 2.12})$$

lo que implica que el 90 por 100 de la variabilidad total se debe a los tratamientos. Como hemos podido comprobar, el coeficiente más conservador es  $R^2$ , que tiende a infraestimar, en relación con los demás, la magnitud de la asociación entre variable independiente y variable dependiente.

#### 2.1.1.4. «Potencia del contraste»

La potencia del contraste es la probabilidad de rechazar una hipótesis nula falsa. Su cálculo requiere conocer las características de la distribución de F cuando la hipótesis alternativa es verdadera (hay efecto de los tratamientos), lo que, en general, no es posible saber *a priori*, puesto que esta distribución depende del parámetro de no centralidad. La potencia depende de la magnitud del efecto, de la tasa de error tipo I, y del tamaño y variabilidad de la muestra. La relación de los tres primeros con la potencia es directa, esto es, cuanto mayor es la magnitud del efecto (o el tamaño de la muestra o la tasa de error tipo I), mayor es la potencia. Por el contrario, a mayor variabilidad de la muestra, menor potencia. SPSS y otros paquetes estadísticos proporcionan la potencia *a posteriori*. En general se considera que potencias por encima de 0,80 (Cohen, 1988) son aceptables. Dado que por lo común no es posible manipular el tamaño del efecto, la potencia deseada puede alcanzarse principalmente por dos vías: incrementando el tamaño de la muestra y/o cambiando la tasa de error tipo I. Algunos paquetes estadísticos, como STATISTICA 6.0, proporcionan programas que permiten estimar *a priori* la potencia, y/o el número de sujetos necesario para alcanzar una potencia deseada.

### 2.1.1.5. «Supuestos del análisis»

El análisis de varianza funciona correctamente cuando se cumplen una serie de supuestos concernientes tanto al modelo estructural como al estadístico de contraste, la F de Snedecor.

#### 2.1.1.5.1. Normalidad

Los errores de las puntuaciones de los individuos de cada población de tratamiento se distribuyen siguiendo una normal de media 0 y varianza  $s^2_e$ . Las formas más típicas de no cumplimiento del supuesto son dos: *a)* las distribuciones son no normales, pero tienen igual forma, y *b)* las distribuciones son no normales y tienen diferente forma. Las primeras investigaciones Montecarlo indicaban que el análisis de varianza es afectado poco por el no cumplimiento del supuesto. Sin embargo, estudios más recientes (Rasmussen, 1986) han indicado que la violación de la normalidad sí afecta a la distribución de F cuando la hipótesis nula es falsa, por tanto, sí afecta a la potencia del contraste (la probabilidad de rechazar la hipótesis nula falsa). El problema de la baja potencia es que pueden escaparse efectos que son realmente significativos. Por ello es importante comprobar el supuesto (por ejemplo, mediante las pruebas de Shapiro-Wilk o Kolmogorov-Smirnov), y en caso de incumplimiento puede ser conveniente realizar transformaciones para normalizar los datos.

#### 2.1.1.5.2. Homogeneidad de las varianzas error

Este supuesto implica que las varianzas de las distintas poblaciones de tratamiento son iguales. La cuestión esencial es si su no cumplimiento afecta de manera importante a la tasa de error tipo I real y a la potencia del test. Wilcox (1987) indica que en ambos casos la respuesta es un claro sí. Con un número igual de sujetos por grupo, el error alfa real es siempre mayor que el error alfa que hemos fijado para tomar la decisión (Tomarken y Serlin, 1986), siendo más grave la diferencia cuanto mayor es el número de sujetos. Cuando el número de sujetos por grupo es desigual, el error alfa real es mayor que el que hemos fijado para tomar la decisión siempre que las varianzas mayores correspondan a los tratamientos con menor número de sujetos. Por el contrario, cuando las varianzas mayores corresponden a los grupos de mayor número de sujetos, el error alfa real es menor que el nominal (Bishop, 1976). Algo semejante puede decirse respecto de la potencia, que disminuye si hay correlación negativa entre número de sujetos y varianzas (Boehnke, 1984). Por tanto, parece claro que es necesario comprobar el cumplimiento de este supuesto. En la literatura se han descrito una gran cantidad de pruebas (alrededor de 60) para

contrastarlo, pero sin duda, la que ha ganado mayor aceptación es la prueba de Levene, debido a que no es afectada por el hecho de que se incumpla el supuesto de normalidad, lo que no ocurre con otras, como la de Hartley o la de Cochran. La prueba de Levene consiste en realizar un análisis de varianza sobre puntuaciones  $Z_{ij}$ , que son el valor absoluto de la diferencia entre la puntuación de cada sujeto y la media de su grupo. Si la prueba de Levene es significativa, las varianzas no son iguales, entonces es conveniente realizar transformaciones para tratar de conseguir la homogeneidad.

#### 2.1.1.5.3. Independencia de los errores

Los errores cometidos en el experimento no deben estar correlacionados entre sí. Habitualmente se consigue el cumplimiento del supuesto controlando las variables relevantes, particularmente evitando que los sucesivos sujetos conozcan las respuestas que hayan podido dar sus predecesores, y asignando al azar los sujetos a los tratamientos. Este supuesto es tanto un requisito estadístico como del diseño, sin cuyo cumplimiento no pueden extraerse conclusiones experimentales válidas. En efecto, la dependencia de las observaciones altera la forma de la distribución de F, de manera que el error alfa real es mayor que el que hemos fijado para tomar la decisión si hay correlación positiva en la serie de datos, y es menor si la correlación es negativa. Puede emplearse bien una prueba de rachas, bien el coeficiente de autocorrelación serial para evaluar si el supuesto se cumple o no.

#### 2.1.1.5.4. Aditividad

Como hemos visto más arriba, el análisis funciona correctamente siempre que la puntuación del cualquier sujeto del experimento pueda obtenerse sumando dos componentes, el efecto del tratamiento y el error. Si ambas aportaciones no son independientes, el análisis no es válido, pero pueden intentarse transformaciones para conseguir la aditividad. Por ejemplo, la forma más común de violación es que la variable dependiente se obtenga como el producto de tratamiento y error (Cox, 1958), por ello, una transformación logarítmica haría que el supuesto se cumpliese, puesto que, como es bien conocido, el logaritmo de un producto es la suma de los logaritmos.

#### 2.1.1.6. «Análisis mediante SPSS»

SPSS permite analizar los diseños univariados unifactoriales entre grupos mediante dos programas diferentes, a través de las opciones incluidas en el grupo de programas Modelo Lineal General, y a través del grupo de programas comparar medias. El más recomendable, por las opciones que ofrece respecto

de los contrastes es el segundo, concretamente mediante el subprograma ANOVA de un factor. En la ilustración llamaremos Terapia a la variable independiente y BDI a la dependiente. Los comandos necesarios para realizar el análisis global de varianza, así como los dos contrastes que hemos exemplificado más arriba y la prueba *a posteriori* de Tukey son los siguientes:

Analizar → Comparar Medias → ANOVA de un factor → Dependientes: BDI; Independiente: Terapia → Contrastos → Coeficientes: 1 → Añadir → 1 → Añadir → -2 → Añadir → Siguiente → 1 → Añadir → -1 → Añadir → 0 → Añadir → Continuar → *Post hoc* → Tukey → Continuar → Opciones → Estadísticos: Homogeneidad de varianzas → Continuar → Aceptar (análisis)

Los resultados ofrecidos por SPSS aparecen en el Cuadro 2.1. (véase la página siguiente). Primero se presenta la prueba de homogeneidad de varianzas de Levene, que indica que las varianzas son homogéneas, lo que permite asumir que los resultados del análisis son, en general, fiables (asumiendo que la asignación de los sujetos a las terapias ha sido aleatoria y que los datos son normales). A continuación aparece la tabla del análisis global de varianza, que, como ya hemos visto, indica que el efecto de la variable terapia es significativo. Finalmente, los coeficientes de las comparaciones planeadas y sus correspondientes análisis ponen de manifiesto que recibir terapia es mejor que no recibirla y que la terapia 1 es preferible a la 2.

El análisis *a posteriori* mediante la prueba de Tukey indica que esas conclusiones son válidas, puesto que todos los tratamientos difieren entre sí. Las estimaciones de la magnitud del efecto y de la potencia *a posteriori* no están incluidas en el programa ANOVA de SPSS, por lo que debe emplearse el grupo de programas Modelo Lineal General, concretamente el programa Univariante. La secuencia de comandos que habría que emplear sería:

Modelo Lineal General → Univariante → Dependiente: BDI → Factores fijos: Terapia → Opciones → Estimaciones del Tamaño del Efecto → Potencia observada → Continuar → Aceptar (análisis)

Cuadro 2.1. Análisis de varianza para un diseño unifactorial entre grupos

Prueba de homogeneidad de varianzas								
BDI								
	Estadístico de Levene	gl1	gl2	Sig.				
,615		2	12	,557				
ANOVA								
BDI	Suma de cuadrados	gl	Media cuadrática	F	Sig.			
	Intergrupos 303,333	2	151,667	28,785	,000			
	Intragrupos 68,000	12	5,667					
	Total 371,333	14						
Coeficientes de contraste								
Contraste	TERAPIA							
	1,00	2,00	3,00					
	1 1	1	-2					
2	1	-1	0					
	Pruebas de contraste							
	Contraste	Valor de contraste	Error típico	t	gl	Sig. (bilateral)		
BDI	Suponer igualdad de varianzas	1 -16,0000	2,6077	-6,136	12	,000		
		2 -6,0000	1,5055	-3,985	12	,002		
	No asume igualdad de varianzas	1 -16,0000	2,2136	-7,228	11,698	,000		
		2 -6,0000	1,7029	-3,523	7,990	,008		
Comparaciones múltiples								
Variable dependiente: BDI								
HSD de Tukey								
(I) Terapia	(J) Terapia	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%			
					Límite inferior	Límite superior		
1,00	2,00	-6,0000*	1,5055	,005	-10,0166	-1,9834		
	3,00	-11,0000*	1,5055	,000	-15,0166	-6,9834		
2,00	1,00	6,0000*	1,5055	,005	1,9834	10,0166		
	3,00	-5,0000*	1,5055	,016	-9,0166	-,9834		
3,000	1,00	11,0000*	1,5055	,000	6,9834	15,0166		
	2,00	5,0000*	1,5055	,016	,9834	9,1066		

\* La diferencia entre las medidas es significativa al nivel ,05

### 2.1.2. Diseños de medidas repetidas de un solo factor

En Psicología una de las formas más frecuentes de manipular las variables independientes consiste en aplicar todos sus niveles al mismo grupo de sujetos, esto es una manipulación intrasujeto. De esta forma, la misma variable dependiente es medida a los mismos sujetos tantas veces como niveles tiene la variable independiente. En principio, cualquier variable independiente directa puede ser manipulada tanto entre grupos como intrasujetos. Sin embargo, según la investigación en ocasiones es preferible una forma de manipulación a la otra. Las terapias del ejemplo anterior podrían ser manipuladas intrasujetos, de manera que se les aplicaría primero una de ellas, luego otra y después la tercera. Pero el lector habrá percibido ya que ésta sería una investigación de dudoso alcance, puesto que la última terapia tendería a ser siempre mejor que la penúltima y, por supuesto, bastante mejor que la primera que se aplicó. Ese efecto no se debería tanto a la propia terapia, cuanto a la secuencia en que se aplicaron. Sin duda esta variable debe manipularse entre grupos. Por el contrario, si nuestro interés residiese en comprobar el curso de la mejora tras la aplicación de una de las terapias, sea la combinada, entonces la variable días de terapia podría ser manipulada preferentemente intrasujeto, puesto que obtendríamos ventajas de economía de tiempo y sujetos, de control de variables relevantes de sujeto, y sería la forma idónea de estudiar los efectos de la secuencia (los sucesivos días de terapia).

¿Cómo puede decidirse si la variable independiente produce efectos diferenciales o no? El análisis de medidas repetidas puede comprenderse rápidamente si se comprende la pregunta fundamental relativa a cómo puede explicarse la puntuación de cualquier sujeto en cualquier nivel de la variable independiente. La respuesta es fácil. Habrá dos tipos de diferencias posibles que sumadas darán la diferencia total. En primer lugar, los sujetos pueden diferir unos de otros (al fin y al cabo son personas diferentes). En segundo lugar, los sujetos pueden diferir de sí mismos de una medida a otra (a la poste son mediciones realizadas con diferente tratamiento y en momentos temporales diferentes). Por tanto:

$$\text{Diferencia Total}_{ij} = \text{Diferencia entre sujetos}_i + \text{Diferencia intrasujetos}_{ij}$$

Ahora bien, la diferencia entre sujetos no es descomponible, pero sí la diferencia de un sujeto con respecto a sí mismo. Esta diferencia puede deberse a dos causas fundamentales, primero, el tratamiento cambia de una medida a otra, por ejemplo, una medida de la variable dependiente se realiza tras 1 día de terapia y otra tras 15 días. Segundo, el error experimental, cuya fuente fundamental será la interacción de los sujetos con la variable independiente, pero que también incluye cualquier otro factor diferente de la propia variable independiente. Por tanto:

$$\text{Diferencia intrasujetos}_{ij} = \text{Efecto del tratamiento}_j + \text{Error}_{ij}$$

Como ya sabemos, las diferencias se traducen en sumas de cuadrados elevándolas al cuadrado y sumando. La ecuación fundamental del diseño será, pues:

$$\text{SC}_{\text{TOTAL}} = \text{SC}_{\text{ENTRE S}} + \text{SC}_{\text{INTRA S}} \quad (\text{Ec. 2.13})$$

$$\text{y} \quad \text{SC}_{\text{INTRA S}} = \text{SC}_{\text{EFECTO}} + \text{SC}_{\text{ERROR IS}} \quad (\text{Ec. 2.14})$$

$$\text{luego} \quad \boxed{\text{SC}_{\text{TOTAL}} = \text{SC}_{\text{ENTRE S}} + \text{SC}_{\text{EFECTO}} + \text{SC}_{\text{ERROR IS}}} \quad (\text{Ec. 2.15})$$

Las sumas de cuadrados se transforman en varianzas, medias de cuadrados, dividiéndolas por sus grados de libertad correspondientes. Los grados de libertad asociados a cada una de ellas, son el número de sumandos que contribuyen menos el número de restricciones (el número de medias que necesitamos conocer para computarlas), por tanto, tendremos:

$$gl_{\text{TOTAL}} = an-1; \quad (\text{Ec. 2.16})$$

$$gl_{\text{ES}} = n-1; \quad (\text{Ec. 2.17})$$

$$gl_{\text{EFECTO}} = a-1, \quad (\text{Ec. 2.18})$$

$$\text{y} \quad gl_{\text{ERROR}} = (a-1)(n-1) \quad (\text{Ec. 2.19})$$

donde a es el número de niveles de la variable independiente y n es el número de sujetos. Finalmente, la F del experimento se obtendrá dividiendo la  $MC_{\text{EFECTO}}$  por la  $MC_{\text{ERROR}}$ .

Ejemplificaremos el análisis suponiendo que hemos manipulado el número de días de la terapia combinada a tres niveles (1, 15 y 25 días) y que hemos medido la depresión, mediante el test BDI, después de cada bloque de días. Los datos ficticios de los cinco sujetos del experimento aparecen en la tabla siguiente:

Sujeto	Días de Terapia			Medias de Sujeto
	1	15	25	
1	25	14	9	16
2	26	9	7	14
3	24	12	9	15
4	24	16	8	16
5	21	14	7	14
Medias de tratamiento	24	13	8	15

El análisis con SPSS debe realizarse mediante el Modelo Lineal General, concretamente a través de Medidas Repetidas. La secuencia de comandos sería la siguiente, suponiendo que en el fichero hemos denominado T1, T15 y T25 a los niveles del factor:

Análisis → Modelo Lineal General → Medidas Repetidas → Nombre del factor intrasujetos: Días; Número de niveles: 3 → Añadir → Definir → Variables intrasujetos: T1, T15, T25 ➤ Aceptar (análisis)

La interpretación de la salida de SPSS requiere que realicemos algunas consideraciones previas sobre los supuestos del análisis. Para empezar, debemos caer en la cuenta de que las puntuaciones de los niveles de la variable independiente pueden correlacionar entre sí. El análisis exige que se cumpla el supuesto de esfericidad. Para comprender su significado es preciso primero definir una matriz de contrastes ortonormales. Esta matriz contiene un conjunto de comparaciones ortogonales cuyos coeficientes han sido normados a varianza unidad. La normalización se consigue dividiendo los coeficientes de cada comparación por la raíz cuadrada de la suma de los cuadrados de sus coeficientes. Por ejemplo, en nuestro caso, el conjunto ortogonal de contrastes tendrá dos comparaciones, sean sus coeficientes los siguientes:

	T1	T15	T25	Sumas	SC
$\varphi_1$	1	-1	0	0	2
$\varphi_2$	1	1	-2	0	6
$\varphi_1 * \varphi_2$	1	-1	0	0	

Puede verse (suma de la fila  $\varphi_1 * \varphi_2$ ) que las comparaciones son ortogonales. A la derecha (columna SC) aparecen las sumas de los cuadrados de los coeficientes. Ahora dividimos por la raíz cuadrada correspondiente y tendremos la matriz ortonormal (que llamaremos C, por ser una matriz de coeficientes):

	T1	T15	T25
$\varphi_1$	0,707	-0,707	0
$\varphi_2$	0,408	0,408	-0,82

Además, necesitamos obtener la matriz de varianzas-covarianzas error. Para computarla definiremos el error de cada puntuación como lo que resta de esa puntuación cuando se han excluido tratamiento y sujetos. Esto es:

$$\text{Error}_{ij} = \text{Puntuación}_{ij} - \text{Media Tratamiento}_j - \text{Media Sujeto}_i + \text{Media total}$$

Por tanto, para el sujeto 1 en el tratamiento 1, tendremos:

$$\text{Error}_{11} = 25 - 24 - 16 + 15 = 0$$

Y así sucesivamente con el resto de los sujetos, hasta obtener la siguiente tabla de errores (nótese que las sumas son cero en las dos direcciones de ésta):

Sujetos	Días de Terapia		
	T1	T15	T25
1	0	0	0
2	3	-3	0
3	0	-1	1
4	-1	2	-1
5	-2	2	0

A continuación podemos calcular las varianzas y covarianzas de los errores de la tabla, de modo que obtendríamos (llamaremos S a la matriz —tabla— resultante):

Días de Terapia	T1	T15	T25
	T1	3,5	-3,8
T15	-3,8	4,5	-0,8
T25	0,25	-0,8	0,5

Ahora sabemos que la covarianza entre 1 día y 15 días es -3,8, o que la covarianza entre 1 y 25 días es 0,25, o que la varianza de los errores en el segundo día es 3,5. Es decir, las varianzas de los errores aparecen en la diagonal principal (desde la esquina superior izquierda hasta la esquina inferior derecha) y las covarianzas fuera de ella. Definiremos a continuación la matriz producto CSC', que denominaremos  $\Sigma_e$ . Esto es:

T1	T15	T25	$\Sigma_e$
0,707	-0,71	0	7,75
0,408	0,408	-0,82	-0,87
			=
0,71	0,41		0,75
-0,7	0,41		
0	-0,8		

Por tanto,  $\Sigma_e$  contiene las varianzas-covarianzas error de los contrastes ortonormalizados. Ahora podemos formular el supuesto de esfericidad, que afirma que la matriz  $\Sigma_e$  es proporcional a una matriz identidad, esto es, que los elementos de la diagonal principal son todos iguales y los elementos fuera de la diagonal son cero (o en cualquier caso no diferentes entre sí). El contraste de

este supuesto puede realizarse mediante la prueba de Mauchly, cuyo estadístico W, se computa de la siguiente manera:

$$W = \frac{|\Sigma_e|}{\left[ \frac{\text{Traza} \Sigma_e}{a-1} \right]^{a-1}} \quad (\text{Ec. 2.20})$$

donde en el numerador tenemos el determinante de la matriz, en el denominador la Traza de la matriz es la suma de los elementos de su diagonal principal, y  $a-1$  son los grados de libertad del efecto (número de niveles de la variable intrasujeto menos 1). Por tanto, en nuestro ejemplo tendremos que:

$$W = \frac{5,06}{\left[ \frac{7,75+0,75}{3-1} \right]^{3-1}} = 0,28$$

La siguiente transformación de W se distribuye como una  $\chi^2$  con v grados de libertad:

$$\chi^2_{(v)} = - \left[ (n-1) - \frac{2a^2 - 3a + 3}{6(a-1)} \right] \ln(W) \quad (\text{Ec. 2.21})$$

$$\text{siendo } v = \frac{a(a-1)}{2} - 1$$

Luego, en nuestro ejemplo, obtendremos que  $v=[3(3-1)/2]-1=2$ , y:

$$\chi^2_{(2)} = - \left[ (5-1) - \frac{2 \cdot 3^2 - 3 \cdot 3 + 3}{6(3-1)} \right] \ln(0,28) = 3,816$$

Consultando en la tabla de chi-cuadrado comprobaremos que la probabilidad de obtener un valor igual o mayor que el observado es 0,148, por lo que no podemos rechazar la hipótesis nula (que la matriz  $\Sigma_e$  es proporcional a la matriz identidad), lo que nos llevaría a pensar que no hay violación del supuesto de esfericidad y que el análisis de varianza es adecuado.

Sin embargo, algunos autores (véase, por ejemplo, Vallejo, 1991) han señalado que incluso violaciones no significativas del supuesto pueden alterar la forma de la distribución F, lo que implicaría que la prueba de Mauchly es demasiado insensible para comprobarlo, y que su uso puede producir conse-

cuencias indeseadas en la tasa de error tipo I real. Una forma alternativa de comprobar qué modificación puede producirse en la forma de F consiste en utilizar un indicador de la alteración, que será, como el lector habrá adivinado ya, un indicador del cambio que se produce en los grados de libertad de numerador y denominador de F como consecuencia de cualquier desviación, por mínima que sea, del supuesto. Dos son los indicadores más frecuentemente empleados, la épsilon de Greenhouse-Geisser ( $\epsilon_{GG}$ ) y la épsilon de Huynh-Feldt ( $\epsilon_{HF}$ ). La  $\epsilon_{GG}$  se define a partir de la matriz S, del siguiente modo:

$$\epsilon_{GG} = \frac{a^2 (s_{ii} - s)^2}{(a-1) \left( \sum_{i=1}^a \sum_{j=1}^a s_{ij}^2 - 2a \sum_{i=1}^a \bar{s}_i^2 + a^2 \bar{s}^2 \right)} \quad (\text{Ec. 2.22})$$

donde  $s_{ij}$  es cualquier elemento de S,  $s_{ii}$  son los elementos de la diagonal principal,  $\bar{s}_i$  son las medias de las filas de S y  $\bar{s}$  es la media de todos los elementos de S. Aplicando a la matriz S del ejemplo que venimos analizando los cálculos son simples, puesto que las sumas de filas y la suma total son todas cero, por tanto, sólo nos queda obtener la suma de los cuadrados de los elementos (=62,125), y la suma de los elementos de la diagonal (=2,83), por tanto:

$$\epsilon_{GG} = \frac{3^2 (2,83 - 0)^2}{(3-1)(62,125 - 2 \cdot 3 \cdot 0 + 3^2 \cdot 0)} = 0,581$$

Así pues, los grados de libertad del numerador de F serán  $(a-1)\epsilon_{GG} = (3-1)*0,581 = 1,163$ , y los del denominador  $(a-1)(n-1)$   $\epsilon_{GG} = (3-1)*(5-1) 0,581 = 4,652$ . Obsérvese que el cambio en la forma de F es notable a tenor del gran cambio en los grados de libertad, a pesar de que la prueba de Mauchly indica que el supuesto de esfericidad no se ha violado. Para evaluar la magnitud del cambio es interesante caer en la cuenta de que el valor mínimo que puede adoptar  $\epsilon_{GG}$  es  $1/(a-1)$ , esto es,  $1/2$  en nuestro ejemplo.

La corrección que introduce la épsilon de Greenhouse-Geisser es demasiado grande, hasta el punto de que algunos autores (por ejemplo, Howell, 1989) recomiendan que se use la corrección menos exigente de Huynh-Feldt, cuando hay razones para creer que épsilon se situará por encima de 0,75. El indicador de Huynh-Feldt, basado en el de Greenhouse-Geisser, se computa como:

$$\epsilon_{HF} = \frac{n(a-1)\epsilon_{GG} - 2}{(a-1)[(n-1) - (a-1)\epsilon_{GG}]} \quad (\text{Ec. 2.23})$$

lo que en nuestro ejemplo se traduce en:

$$\epsilon_{HF} = \frac{5(3 - 1)(0,581 - 2)}{(3 - 1)[(5 - 1) - (3 - 1)(0,581)]} = 0,672$$

Según  $\epsilon_{HF}$ , los grados de libertad de numerador y denominador de F serán ahora, respectivamente, 1,345 y 5,379. La corrección con el índice de Huynh-Feldt es prácticamente siempre menos severa que con el de Greenhouse-Geisser.

¿Qué podemos hacer cuando el supuesto de esfericidad no se cumple? Como acabamos de comprobar hay varias alternativas disponibles. En primer lugar, podemos intentar continuar dentro del marco de ANOVA, para lo cual nos basta con utilizar el análisis estándar pero ajustando los grados de libertad según alguno de los dos valores de épsilon, o bien alguna medida derivada de ambos, como su media (Stevens, 1992). En segundo lugar, todavía dentro del marco de ANOVA, es posible tomar una decisión adecuada realizando un conjunto de comparaciones ortogonales (los paquetes estadísticos suelen realizar un análisis de tendencias, que es siempre ortogonal, aunque, obviamente no debe interpretarse como tal si los niveles de la variable no están igualmente espaciados). La ventaja que tiene este análisis es que no cabe que se viole el supuesto, puesto que las comparaciones se establecen siempre entre dos condiciones (o dos grupos de condiciones, que es lo mismo). En otras palabras, si sólo hay un grado de libertad, sólo hay una covarianza y no cabe hablar de falta de esfericidad. Si alguno de los contrastes es significativo, entonces al menos hay una diferencia significativa y por tanto, podemos rechazar la hipótesis nula global. La tercera alternativa consiste en salir del marco de ANOVA y emplear análisis multivariado de varianza (véase Capítulo IX). Una cuarta alternativa, útil en realidad sólo en diseños de una sola variable independiente, consiste en transformar la escala de la variable dependiente, realizando un análisis no paramétrico, por ejemplo, mediante la prueba de Friedman (si tenemos más de dos niveles) o la prueba de Wilcoxon (para comparar dos niveles entre sí).

Ahora ya podemos interpretar la salida proporcionada por SPSS (o cualquier otro paquete como STATISTICA) de manera más adecuada. El Cuadro 2.2. (véase página siguiente) presenta los aspectos esenciales del análisis del ejemplo realizado con SPSS. En primer lugar SPSS presenta los resultados de las pruebas multivariadas (véase Capítulo IX), que permiten rechazar la hipótesis nula. En segundo lugar, la prueba de esfericidad de Mauchly que hemos presentado más arriba, que parece indicar que no hay violación del supuesto de esfericidad. Sin embargo, en esa misma tabla podemos comprobar que los valores de épsilon se alejan de 1 (el límite superior), que implica ningún cambio en los grados de libertad (ninguna violación del supuesto), y se acercan al límite inferior, 0,5, en nuestro ejemplo, e implican que la violación, aun no siendo significativa según Mauchly, es suficientemente grande como para ser tenida en cuenta. En la tabla de ANOVA podemos comprobar que la hipótesis nula puede rechazarse asumiendo la esfericidad e incluso en el peor de los casos de violación, cuando épsilon adopta su límite inferior. El análisis de tendencias

que SPSS realiza por defecto abunda en la misma dirección, puesto que el componente lineal (una comparación entre el primer y el último nivel, T1 y T25) es significativo. Nótese, además que en cualquiera de las pruebas la magnitud del efecto (eta cuadrado o R cuadrado) se sitúa por encima de 0,95, y que la potencia observada es 1 (la tasa de error tipo II es cero). En consecuencia, la variable días de tratamiento produce efectos diferenciales.

Cuadro 2.2. Análisis de varianza para un diseño de un solo factor repetido con tres niveles

Contrastes multivariados								
Efecto	Valor	F	Gl de la hipótesis	Gl del error gl	Sig.	Eta cuadrado	Parámetro de no centralidad	Potencia observada
DÍAS	Traza de Pillai ,994	270,556	2,000	3,000	,000	,994	541,111	1,000
	Lambda de Wilks ,006	270,556	2,000	3,000	,000	,994	541,111	1,000
	Traza de Hoteling 180,370	270,556	2,000	3,000	,000	,994	541,111	1,000
	Raíz mayor de Roy 180,370	270,556	2,000	3,000	,000	,994	541,111	1,000
Prueba de esfericidad de Mauchly								
Medida: MEASURE_1								
Efecto intra-sujetos	W. de Mauchly	Chi-cuadrado		gl	Sig.	Épsilon		
		aprox				Greenhouse-Geisser	Huynh-Feldt	Límite-inferior
DÍAS	,280	3,816	2	,148	,581	,672	,500	
Pruebas de efectos intrasujetos								
Medida: MEASURE_1								
Fuente		Suma de cuadrados tipo III		gl	Media cuadrática	F	Sig.	Eta cuadrado
DÍAS	Esfericidad asumida	670,000	2	335,000	78,824	,000	,952	157,647
	Greenhouse-Geisser	670,000	1,163	576,107	78,824	,000	,952	91,670
	Huynh-Feldt	670,000	1,345	498,259	78,824	,000	,952	105,992
	Límite-inferior	670,000	1,000	670,000	78,824	,001	,952	78,824
Error (DÍAS)	Esfericidad asumida	34,000	8	4,250				
	Greenhouse-Geisser	34,000	4,652	7,309				
	Huynh-Feldt	34,000	5,379	6,321				
	Límite-inferior	34,000	4,000	8,500				
<sup>a</sup> Calculado con alfa = ,05								
Pruebas de contrastes intrasujetos								
Medida: MEASURE_1								
Fuente	DÍAS	Suma de cuadrados tipo III		gl	Media cuadrática	F	Sig.	Eta cuadrado
DÍAS	Lineal	640,000	1	640,000	365,714	,000	,989	365,714
	Cuadrático	30,000	1	30,000	4,444	,103	,526	4,444
Error (DÍAS)	Lineal	7,000	4	1,750				
	Cuadrático	27,000	4	6,750				
<sup>a</sup> Calculado con alfa = ,05								

Desafortunadamente SPSS es muy pobre por lo que hace referencia a las comparaciones entre niveles de las variables, tanto planeadas como no planeadas. Las comparaciones planeadas disponibles son realmente inservibles para un investigador que pretende comprobar hipótesis específicas. Por ejemplo, el contraste simple permite comparar cada nivel con un nivel de referencia, lo que

evidentemente, a no ser que uno de los niveles de la variable independiente sea una condición de control, carece de utilidad. Por tanto, la única opción disponible en este paquete es utilizar la t de Student para muestras correlacionadas para hacer las comparaciones que se hayan planificado. Por otra parte, las comparaciones no planeadas no pueden realizarse en SPSS más que sobre factores manipulados entre grupos. Si el investigador está interesado en realizar pruebas *a posteriori*, bien debe utilizar un paquete alternativo, como STATISTICA, bien debe realizarlas en SPSS como si fuesen planeadas, mediante la t de Student y emplea la desigualdad de Bonferroni para ajustar el error alfa al que va a tomar cada decisión.

## 2.2. DISEÑOS FACTORIALES UNIVARIADOS

Sin duda los diseños de investigación más frecuentemente empleados en la actualidad son los que incluyen dos o más variables independientes, cruzadas factorialmente. Como hemos señalado más arriba, el cruce factorial implica que las condiciones experimentales se obtienen realizando todas las combinaciones posibles de los niveles de todas las variables manipuladas. El número de condiciones resultantes es siempre el producto de los niveles de los factores que se mezclan. Estos diseños tienen una ventaja fundamental respecto de los unifactoriales, su acercamiento a una realidad compleja en la que, por lo general, más de un factor está influyendo en el comportamiento de las personas. Por ejemplo, no es frecuente que se intente estudiar la eficacia de varios tratamientos sin tener en cuenta el número de días que se aplican. Un estudio mucho más realista consistiría, pues, en cruzar factorialmente ambas variables. Si mantenemos la manipulación de niveles que hemos realizado en los dos ejemplos precedentes tendríamos un total de nueve condiciones experimentales: cada tratamiento es estudiado a lo largo de tres niveles de la variable días de terapia.

Supongamos que los datos de BDI obtenidos hubiesen sido los siguientes:

A1			A2			A3		
T1	T15	T25	T1	T15	T25	T1	T15	T25
25	15	5	24	16	14	24	21	22
24	14	7	21	14	15	25	23	20
18	11	5	26	16	13	26	23	21
26	13	4	25	14	16	23	20	19
22	12	4	24	15	17	22	23	18

Es evidente que podemos obtener tres clases de efectos, con independencia de si las variables están manipuladas entre grupos o intrasujeto. Por una parte, podemos calcular el efecto diferencial de la terapia, sin tener en cuenta la variable días. Llamaremos a éste el efecto principal del factor terapia. Para computarlo será necesario obtener los promedios adecuados, sumando todas las pun-

taciones correspondientes a cada terapia, y dividiendo por el número de sumandos, tendremos que:

A1	A2	A3
13,67	18	22

Por otro lado, sumando las puntuaciones correspondientes a cada uno de los niveles de la variable días y dividiendo por el número de sumandos obtendremos los promedios que nos permitirán conocer los efectos principales del factor días:

T1	T15	T25
23,67	16,67	13,33

Finalmente, sumando las puntuaciones correspondientes a cada condición experimental y dividiendo por el número de sumandos, obtendremos los promedios de las condiciones experimentales:

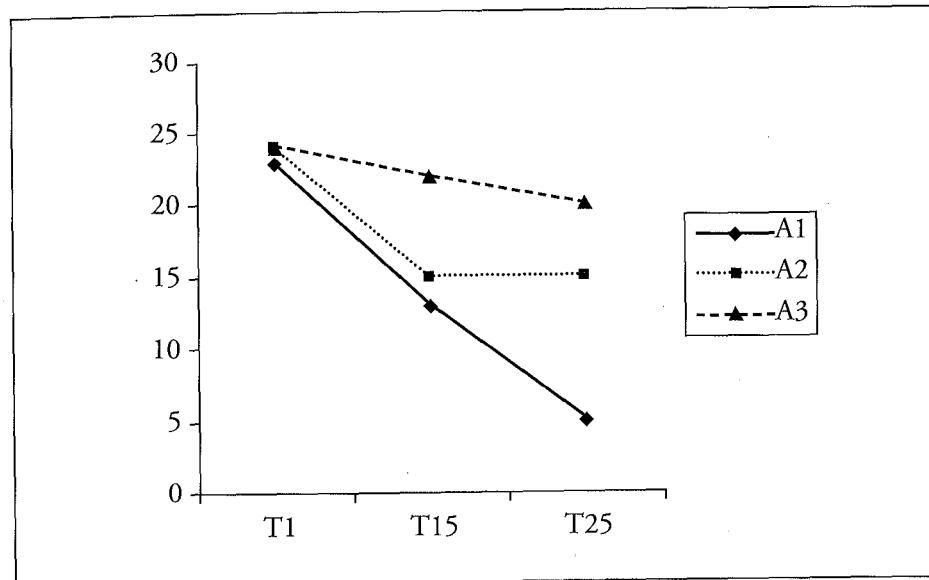
	T1	T15	T25
A1	23	13	5
A2	24	15	15
A3	24	22	20

Estos promedios de condición experimental son fundamentales para computar el efecto que supone una aportación radical de los diseños factoriales, la interacción de factores. El concepto de interacción puede entenderse fácilmente si representamos gráficamente los datos. La Figura 2.1. contiene el gráfico de la interacción. Obsérvese que a medida que transcurren los días de terapia la puntuación BDI va disminuyendo de forma clara cuando consideramos la terapia combinada (A1), pero no ocurre igual con la terapia farmacológica (A2) y menos aún con el control de no terapia (A3). Esto mismo puede comprobarse examinando las medias, o todavía más examinando los efectos de interacción, computables como lo queda de las medias de casilla cuando eliminamos los efectos principales de las dos variables, esto es:

$$\text{Efecto Interacción}_{jk} = \text{Casilla}_{jk} - \text{Media}_j - \text{Media}_k + \text{Media Total}$$

	T1	T15	T25
A1	3,556	0,556	-4,11
A2	0,222	-1,78	1,556
A3	-3,78	1,222	2,556

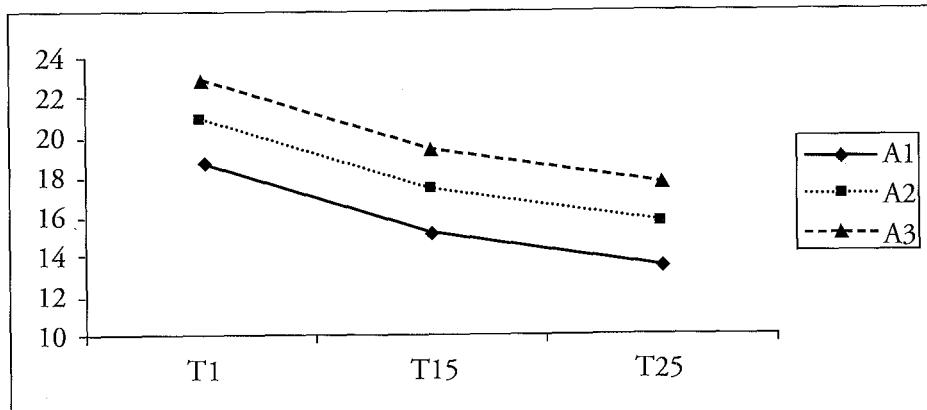
Figura 2.1. Interacción de dos factores (Terapia x Días)



Si los factores no interactuasen se esperaría que los efectos fuesen nulos, y, por tanto, todas las casillas iguales. Consideremos la siguiente tabla factorial y su representación gráfica que aparece en la Figura 2.2:

	T1	T15	T25	
A1	18,67	15,17	13,5	15,78
A2	20,83	17,33	15,67	17,94
A3	22,83	19,33	17,67	19,94
	20,78	17,28	15,61	17,89

Figura 2.2. Representación gráfica de dos factores que no interactúan

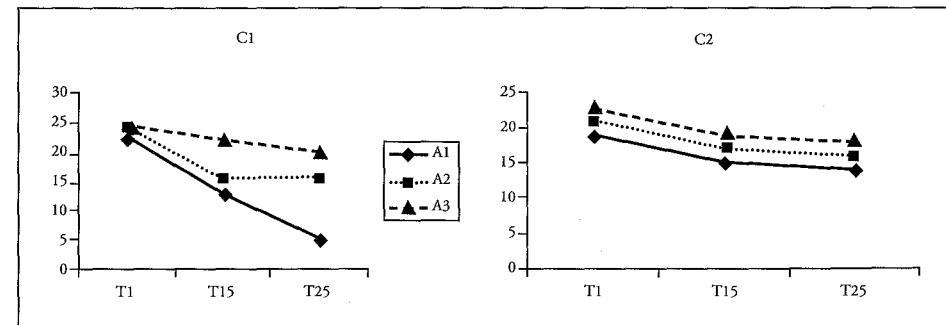


Ahora, las líneas son paralelas, y los efectos de la interacción serían:

	T1	T15	T25
A1	0	0	0
A2	0	0	0
A3	0	0	0

La interacción de factores se manifiesta gráficamente en forma de líneas no paralelas, sea que se crucen físicamente o no. Aritméticamente se traduce en que los efectos de interacción son no nulos. Por tanto, podemos decir que la interacción de factores implica que los efectos de un factor dependen del nivel del otro factor que se considere. La interacción implica que las medias de casilla no son explicables sumando los efectos principales de los factores. Cuando hay más de dos factores pueden aparecer interacciones de orden superior. Por ejemplo, con tres factores (sean A, B y C), pueden aparecer tres interacciones de primer orden ( $A \times B$ ,  $A \times C$  y  $B \times C$ ), y una de segundo orden ( $A \times B \times C$ ). La interacción de segundo orden implica que las interacciones de primer orden dependen de los niveles del tercer factor. Por ejemplo, la interacción  $A \times B$  es diferente en distintos niveles del factor C. La Figura 2.3. presenta un ejemplo de una interacción de segundo orden, en el que se aprecia que la interacción  $A \times B$  existe en el primer nivel de C, pero no en el segundo.

A continuación presentaremos el análisis global de varianza de los diseños de dos factores, junto con el análisis de comparaciones, principalmente orientado al análisis detallado de la interacción de factores.

Figura 2.3. Interacción de segundo orden. La interacción  $A \times B$  cambia de un nivel de C a otro

### 2.2.1. Análisis global de varianza de los diseños de dos factores

En función de si las variables independientes se han manipulado entre grupos o intrasujetos, podemos encontrar tres tipos de diseños factoriales diferentes. Cuando todas las variables se han manipulado entre grupos, el diseño se denomina factorial entre grupos y se caracteriza porque cada sujeto es medido

sólo una vez en el experimento, por lo que tendremos tantos grupos de sujetos como condiciones experimentales resulten de la mezcla factorial. Por ejemplo, si manipulamos entre grupos los dos factores que venimos empleando en la exemplificación, el tipo de terapia y el número de días de terapia, tendríamos, como hemos visto, nueve condiciones experimentales, cada una de las cuales sería aplicada a un grupo diferente de sujetos. En el otro extremo, si ambos factores fuesen manipulados intrasujetos (lo que no resultaría en un buen diseño), cada sujeto del experimento sería medido nueve veces, esto implica que recibiría los tres tipos de terapia y los tres niveles de la variable días en cada una de las terapias. El factorial se llamaría entonces intrasujetos, o de medidas repetidas en dos factores. Finalmente, si una variable fuese manipulada entre grupos (pongamos el tipo de terapia) y la otra intrasujetos (los días de terapia), tendríamos un diseño de medidas parcialmente repetidas, o mixto. Ahora habría tantos grupos de sujetos como niveles tenga el factor entre grupos, y cada sujeto de esos grupos sería medido tantas veces en la variable dependiente como niveles tenga el factor intrasujeto.

Las diferencias entre los tres tipos de diseños factoriales son importantes respecto del análisis. Recordemos que una buena estrategia para saber cuáles son las fuentes de variabilidad de un diseño consiste en preguntarse acerca de cómo puede obtenerse la puntuación de cualquier sujeto en cualquier condición experimental. Denotemos, como siempre, a los sujetos por el subíndice  $i$ , a los niveles del primer factor (A, Terapia en nuestro ejemplo), mediante el subíndice  $j$  y emplearemos el subíndice  $k$  para el factor restante. Así, tendremos en el diseño factorial entre grupos que:

$$\text{Puntuación}_{ijk} - \text{Media Total} = \text{Efecto de } A_j + \text{Efecto } B_k + \text{Interacción}_{jk} + \text{Error}_{ijk}$$

Lo importante aquí es caer en la cuenta de que aparte de los efectos de las variables y su interacción, sólo es computable la diferencia entre los sujetos dentro del mismo grupo, esto es, como hemos visto antes, el error intragrupo. Por tanto, tendremos que:

$$SC_{\text{TOTAL}} = SC_A + SC_B + SC_{AxB} + SC_{\text{ERROR IG}}$$

Por otra parte, en el diseño factorial intrasujeto hay que considerar más efectos, puesto que ahora podemos calcular diferencias entre los sujetos con independencia de las condiciones experimentales, y, por otro lado, los sujetos pueden interactuar con cada una de las variables e incluso con la interacción de las mismas. Lo cual implica que:

$$\text{Puntuación}_{ijk} - \text{Media Total} = \text{Efecto } A_j + \text{Interacción de Sujetos con A} + \text{Efecto } B_k + \text{Interacción de Sujetos con B} + \text{Interacción}_{jk} + \text{Interacción de Sujetos con AxB} + \text{Efecto de Sujetos}$$

Lo que en términos de sumas de cuadrados se traduce en:

$$SC_{\text{TOTAL}} = SC_A + SC_{AxS} + SC_B + SC_{BxS} + SC_{AxB} + SC_{AxBxS} + SC_{ES}$$

Finalmente, en el caso del diseño mixto, los sujetos de cada grupo pueden interactuar únicamente con el factor repetido, por tanto, tendremos que:

$$\begin{aligned} \text{Puntuación}_{ijk} - \text{Media Total} &= \text{Efecto } A_j + \text{Error } IG_{ij} + \text{Efecto } B_k + \text{Interacción}_{jk} + \text{Puntuación}_{ijk} \\ \text{Media Total} &= \text{Efecto } A_j + \text{Interacción de Sujetos con A} + \text{Efecto } B_k + \text{Interacción de Sujetos con B} + \text{Interacción}_{jk} + \text{Interacción de Sujetos con AxB} \end{aligned}$$

Y en términos de sumas de cuadrados:

$$SC_{\text{TOTAL}} = SC_A + SC_{\text{ERROR IG}} + SC_B + SC_{AxB} + SC_{BxS} \text{ SEGÚN GRUPO}$$

Para completar el análisis global es necesario computar los grados de libertad asociados a cada suma de cuadrados. Los grados de libertad de cualquier suma de cuadrados, cualquiera que sea el diseño, pueden obtenerse aplicando un pequeño conjunto de reglas generales. En primer lugar, los grados de libertad de un efecto principal son el número de niveles del efecto menos 1. Por tanto, en todos los diseños tendremos que los grados de libertad del factor A serán  $a-1$ , y los del factor B,  $b-1$ . Los grados de libertad de una interacción son el producto de los grados de los factores que interactúan. Así, los de la interacción  $AxB$  serán  $(a-1)(b-1)$ . Los de  $AxBxC$  serán  $(a-1)(b-1)(c-1)$ . Respecto de los errores, las reglas son también sencillas. En primer lugar, si el error es entre grupos, sus grados de libertad son número de sujetos menos 1 por el número de grupos. Así, el error intragrupo en el caso del diseño de dos factores entre grupos serán  $(n-1)ab$ , puesto que tenemos  $n$  sujetos en cada grupo y  $ab$  grupos. En el diseño mixto, sin embargo, los grados de libertad del error intragrupo serán  $(n-1)a$ , puesto que sólo tenemos  $a$  grupos. En segundo lugar, si el error es intrasujeto, la fuente principal del error es la interacción de los sujetos con los factores intrasujeto, por tanto, el error será computado como en cualquier interacción. Por ejemplo, en el diseño de dos factores intrasujeto tenemos tres errores, dos asociados a cada factor y uno a la interacción, por tanto, los grados de libertad del error asociado a A serán  $(a-1)(n-1)$ , los del asociado a B  $(b-1)(n-1)$ , y los asociados a la interacción  $(a-1)(b-1)(n-1)$ . Por otra parte, en el caso del diseño mixto los sujetos sólo pueden interactuar con el factor intrasujeto, B, luego, en cada grupo tendremos  $(b-1)(n-1)$ , y puesto que hay  $a$  grupos, en total tendremos  $(b-1)(n-1)a$  grados de libertad.

Las medias de cuadrados se obtienen dividiendo las sumas de cuadrados por sus grados de libertad respectivos. Finalmente, las F de cada efecto son cocientes entre medias de cuadrados. En el numerador tendremos una media de cuadrados de efecto y en el denominador su error correspondiente. Cuál sea la media de

cuadrados de error es algo fácil de determinar, puesto que si el numerador contiene una fuente entre grupos, el error será intragrupo, mientras que si la fuente contiene algún factor intrasujeto, el error será la interacción de los sujetos con esa fuente. Por ejemplo, en el caso del diseño de dos factores entre grupos, todos los efectos son entre grupos, y su error es común, el intragrupo. Así, tendremos:

$$F_A = MC_A/MC_{IG}$$

$$F_B = MC_B/MC_{IG}$$

$$F_{AxB} = MC_{AxB}/MC_{IG}$$

En el diseño de dos factores intrasujeto, todas las fuentes son intrasujeto, y su error será específico. Concretamente:

$$F_A = MC_A/MC_{AxS} \quad (\text{la interacción de los sujetos con A})$$

$$F_B = MC_B/MC_{BxS} \quad (\text{la interacción de los sujetos con B})$$

$$F_{AxB} = MC_{AxB}/MC_{AxBxS} \quad (\text{la interacción de los sujetos con AxB})$$

Finalmente, en el diseño mixto:

$$F_A = MC_A/MC_{S/A} \quad (\text{el error intragrupo})$$

$$F_B = MC_B/MC_{BxS/A} \quad (\text{la interacción de los sujetos con B en cada grupo})$$

$$F_{AxB} = MC_{AxB}/MC_{BxS/A} \quad (\text{la interacción de los sujetos con B en cada grupo})$$

Es importante caer en la cuenta de que un mismo conjunto de datos puede llevar a conclusiones diferentes según estén manipuladas las variables. En general, el diseño menos potente (y el más costoso) es el entre grupos, seguido del mixto y del intrasujeto. Esto implica que es más fácil declarar efectos significativos en el último que en los dos anteriores. El análisis mediante SPSS en cada uno de estos diseños puede realizarse con las siguientes secuencias de comandos:

Diseño factorial entre grupos:

Analizar → Modelo lineal general → *Univariante* → Dependientes: BDI; Factores fijos: Terapia; Días → Opciones → Estimaciones del tamaño del efecto; Potencia observada → Continuar → Aceptar (análisis)

Diseño factorial intrasujetos:

Analizar → Modelo lineal general → *Medidas repetidas* → Nombre del factor intrasujetos: Terapia; Número de niveles: 3 → Añadir → Nombre del factor intrasujetos: Días; Número de niveles: 3 → Añadir → Medida: BDI → Añadir → Definir → Variables intrasujetos: A1T1; A1T15; A1T25; A2T1; A2T15; A2T25; A3T1; A3T15; A3T25 → Opciones → Estimaciones del tamaño del efecto; Potencia observada → Continuar → Aceptar (análisis)

Diseño factorial mixto (Terapia es el factor entre y Días el intra):

Analizar → Modelo lineal general → *Medidas repetidas* → Nombre del factor intrasujetos: Días; Número de niveles: 3 → Añadir → Medida: BDI → Añadir → Definir → Variables intrasujetos: T1; T15; T25 → Factores intersujetos: Terapia → Opciones → Estimaciones del tamaño del efecto; Potencia observada → Continuar → Aceptar (análisis)

La salida de SPSS para cada diseño es similar a las que hemos visto antes respecto de los diseños de un solo factor. Para exemplificar, consideremos la salida que aparece en el Cuadro 2.3. (véase página siguiente), obtenida con los datos de la tabla factorial que hemos venido empleando como ejemplo. Puesto que tenemos un factor intrasujeto, SPSS presentará resultados tanto en una aproximación multivariada, como en la univariada para todos los efectos en que ese factor aparezca, por tanto, tendremos el efecto principal de Días y su interacción con el factor entre grupos, la Terapia. El análisis multivariado indica que tanto uno como otro efecto son significativos. Además, la magnitud del efecto de Días (medida mediante eta cuadrado) es un poco mayor que la de la interacción, pero la potencia observada es idéntica y total en ambos casos. En segundo lugar, la prueba de esfericidad de Mauchly no es significativa, lo que implicaría, con las salvedades mencionadas más arriba que se cumple el supuesto. Esta conclusión puede corroborarse examinando los valores de épsilon, que oscilan entre 0,844 y 1, respectivamente para Greenhouse-Geisser y Huynh-Feldt. Es evidente que se alejan sustancialmente del límite inferior, lo que implicaría que efectivamente no hay una gran desviación de la esfericidad y que, por tanto, el análisis univariado de varianza es aplicable con garantías. A continuación aparece la tabla resumen del ANOVA para las fuentes que incluyen al factor días. De nuevo puede apreciarse que tanto su efecto principal como su interacción con la terapia son significativos, aun en el caso de que se produjese una muy grave violación del supuesto (véase la significatividad para el límite inferior de épsilon). A esta misma conclusión puede llegarse examinando el análisis de contrastes ortogonales (de tendencias) que indica que tanto el componente lineal (una comparación del primer nivel de días con el último) como el cuadrático (una comparación del segundo nivel de días con primero y último juntos) son significativos. Finalmente, el análisis de las fuentes entre grupos, indica que el factor terapia es significativo, siendo su magnitud de efecto bastante elevada y también su potencia. En conclusión, tenemos efectos significativos de los efectos principales de cada factor y de la interacción de ambos.

## Análisis univariado de varianza

Cuadro 2.3. Análisis de varianza de un diseño mixto con el factor Terapia manipulado entre grupos a 3 niveles y el factor Días de terapia manipulado intrasujeto a 3 niveles

Contrastes multivariados									
	Valor	F	Gl de la hipótesis	Gl del error gl	Sig.	Eta cuadrado	Parámetro de no centralidad	Potencia observada	
DÍAS	Traza de Pillai	,959	129,502	2,000	,000	,959	259,004	1,000	
	Lambda de Wilks	,041	129,502	2,000	,000	,959	259,004	1,000	
	Traza de Hoteling	23,546	129,502	2,000	,000	,959	259,004	1,000	
	Raíz mayor de Roy	23,546	129,502	2,000	,000	,959	259,004	1,000	
"DÍAS*TERAPIA	Traza del Pillai	1,546	20,417	4,000	24,000	,000	,773	81,667	1,000
	Lambda de Wilks	,043	20,971	4,000	22,000	,000	,792	83,883	1,000
	Traza de Hoteling	8,522	21,306	4,000	20,000	,000	,810	85,223	1,000
	Raíz mayor de Roy	6,386	38,317	2,000	12,000	,000	,865	76,635	1,000

Prueba de esfericidad de Mauchly									
	Medida: BDI								
	W. de Mauchly	Chi-cuadrado aprox	gl	Sig.	Epsilon				
Efecto intrasujetos	,815	2,257	2	,324	Greenhouse-Geisser	Huynt-Feldt	Límite-inferior		
DÍAS					,844	,1,000	,500		

Pruebas de efectos intrasujetos									
	Medida: BDI								
	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta cuadrado	Parámetro de no centralidad	Potencia observada <sup>a</sup>	
DÍAS	Esféricidad asumida	834,444	2	417,222	172,644	,000	,935	345,287	1,000
	Greenhouse-Geisser	834,444	1,687	494,614	172,644	,000	,935	291,260	1,000
	Huynt-Feldt	834,444	2,000	417,222	172,644	,000	,935	345,287	1,000
	Límite-inferior	834,444	1,000	834,444	172,644	,000	,935	172,644	1,000
DÍAS*TERAPIA	Esféricidad asumida	288,889	4	72,222	29,885	,000	,8333	119,540	1,000
	Greenhouse-Geisser	288,889	3,374	85,619	29,885	,000	,8333	100,836	1,000
	Huynt-Feldt	288,889	4,000	72,222	29,885	,000	,8333	119,540	1,000
	Límite-inferior	288,889	2,000	144,444	29,885	,000	,8333	59,770	1,000
Error (DÍAS)	Esféricidad asumida	58,000	24	2,417					
	Greenhouse-Geisser	58,000	20,245	2,865					
	Huynt-Feldt	58,000	24,000	2,417					
	Límite-inferior	58,000	12,000	4,833					

Pruebas de contrastes intrasujetos									
	Medida: BDI								
	Fuente	DÍAS	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta cuadrado	Parámetro de no centralidad
DÍAS	Lineal	800,833	1	800,833	234,390	,000	,951	234,390	1,000
		33,611	1	33,611	23,725	,000	,664	23,725	,994
DÍAS*TERAPIA	Lineal	251,667	2	125,833	36,829	,000	,860	73,659	1,000
		37,222	2	18,611	13,137	,001	,686	26,275	,985
DÍAS	Cuadrático	41,000	12	3,417					
		17,000	12	1,417					

Pruebas de los efectos intersujetos									
	Medida: BDI								
	Variable transformada:	Promedio	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta cuadrado	Parámetro de no centralidad
Intercept	14,400,556	1	14,400,556	3200,123	,000	,996	3,200,123		1,000
TERAPIA	521,111	2	260,556	57,901	,000	,906	115,802		1,000
Error	54,000	12	4,500						

<sup>a</sup> Calculado con alfa = ,05

## 2.2.2. Análisis de contrastes

El análisis de contrastes en los diseños factoriales puede tratar de esclarecer dos tipos de efectos, los principales de cada factor y los de la interacción de factores. Cuando la interacción es significativa, lo más frecuente es intentar determinar de qué forma se concreta, en lugar de orientarse hacia los efectos principales. Hay dos razones importantes para esa preferencia. En primer lugar, un investigador que diseña un experimento factorial espera que la interacción sea significativa. Por tanto, el análisis posterior debe conformarse a esa expectativa. En segundo lugar, en general, cuando una interacción es significativa, la descripción de la realidad que puede realizarse estudiándola de forma detallada es más completa y adecuada que si el investigador se centra en el estudio de los efectos principales. Por ejemplo, el efecto principal de la variable días (véase el análisis de tendencias del Cuadro 2.3.) indica que parece haber una mejora a medida que transcurren los días de terapia. Esta afirmación se realiza sin tener en cuenta la terapia, o lo que es lo mismo, se afirma como correcta en todas las terapias. Sin embargo, esa afirmación puede no ser del todo cierta, puesto que es fácil apreciar (véase Figura 2.1.) que la mejora es considerable si se emplea la terapia 1, pero prácticamente no existe si se emplea el control. Por tanto, la cuestión más importante a resolver cuando la interacción es significativa es cómo se concreta ese efecto.

Una de las formas de estudiar la interacción, podríamos decir que la más frecuente con mucho, consiste en examinar el efecto de una variable en cada uno de los niveles de la otra variable. Esta aproximación permite responder a preguntas del tipo siguiente: ¿qué efecto producen los días de terapia en la terapia A1? ¿Y en la A2? ¿Y en la A3? Pero también a preguntas en la otra dirección: ¿Qué efecto producen las terapias en T1? ¿Qué efecto en T15? ¿Y en T25? Este tipo de análisis de efectos simples se realiza de una manera muy sencilla, puesto que, conceptualmente se traduce en reducir el diseño factorial a un conjunto de diseños de un solo factor. Aunque en sentido estricto esta reducción es inadecuada si se cumplen los supuestos de homogeneidad de varianzas (en los diseños entre grupos) o de esfericidad (en los diseños que incluyen variables intrasujeto), viene prácticamente obligada por el hecho de que los paquetes estadísticos no permiten, en general, el uso de los términos error obtenidos en el análisis global. Por tanto, en la práctica, un diseño como el mixto que estamos ejemplificando se descompondrá en tres diseños de un solo factor entre grupos, en los que el factor sería la terapia, y los distintos días serían una constante en cada diseño, y en tres diseños intrasujetos, en los que el único factor serían los días y la terapia sería una constante. El Cuadro 2.4. (véase página siguiente) contiene los datos correspondientes a estos seis diseños. Para realizar su análisis global y de contrastes utilizaríamos las secuencias de comandos de SPSS que hemos presentado más arriba para los diseños de un solo factor.

Cuadro 2.4. El diseño factorial como un agregado de diseños unifactoriales: Análisis de efectos simples de la interacción de primer orden Terapia x Días

1) Diseños entre grupos:

Efecto de Terapia en T1			Efecto de Terapia en T15			Efecto de Terapia en T25		
A1	A2	A3	A1	A2	A3	A1	A2	A3
25	24	24	15	16	21	5	14	22
24	21	25	14	14	23	7	15	20
18	26	26	11	16	23	5	13	21
26	25	23	13	14	20	4	16	19
22	24	22	12	15	23	4	17	18

2) Diseños intrasujetos:

Efecto de Días en A1			Efecto de Días en A2			Efecto de Días en A3		
T1	T15	T25	T1	T15	T25	T1	T15	T25
25	15	5	24	16	14	24	21	22
24	14	7	21	14	15	25	23	20
18	11	5	26	16	13	26	23	21
26	13	4	25	14	16	23	20	19
22	12	4	24	15	17	22	23	18

El segundo tipo de análisis de la interacción es el que se conoce como interacción parcial. Como su nombre indica, se trata de reducir la complejidad de la estructura factorial, pero reduciendo el número de niveles en uno o más factores. La forma de reducir niveles en un factor es realizar comparaciones entre ellos. Por ejemplo, si antes de realizar el experimento hubiésemos previsto que la diferencia entre las terapias sería mayor al final del tratamiento que al principio, habríamos planificado una comparación entre los niveles T1 y T25 de la variable días, pero manteniendo intacto el factor terapia. Así, los coeficientes del contraste serían los siguientes:

	T1	T15	T25
A1	1	0	-1
A2	1	0	-1
A3	1	0	-1

La reducción de niveles en los dos factores implica que se realizan simultáneamente comparaciones sobre ambos. Por ejemplo, podríamos haber previsto que la diferencia entre T1 y T15 sería la misma en A1 que en A2. En este caso excluimos la condición de control de terapia (A3) y el nivel T25 de la variable días. Los coeficientes del contraste de interacción serían el producto cartesiano de los coeficientes del contraste en Terapia (1, -1, 0) con los del contraste en la

variable días (1, -1, 0). Los coeficientes resultantes indican claramente que el contraste de interacción puede realizarse eliminando la condición T25 y la condición A3, lo que pone de manifiesto claramente que estamos calculando una parte de la interacción:

	T1	T15	T25
A1	1	1	0
A2	-1	-1	0
A3	0	0	0

Desafortunadamente, SPSS no permite el análisis de contrastes de interacción o parciales definidos por el usuario (si exceptuamos el análisis de la interacción parcial de tendencias). Otros paquetes estadísticos como STATISTICA los incorporan de manera natural. Sin embargo, al usuario de SPSS siempre le queda el recurso de excluir del análisis las condiciones intrasujeto y de seleccionar los grupos de sujetos que entran en la interacción parcial, siempre que el contraste sea simple (no implique la mezcla de condiciones). Por ejemplo, para realizar el contraste anterior, primero seleccionamos los sujetos de los grupos A1 y A2 con la siguiente secuencia de comandos:

Datos → Seleccionar casos → Si se satisface la condición... → Si → Tera-pia >< 3 → Continuar → Aceptar

Ahora los sujetos del grupo A3 aparecen marcados como excluidos. SPSS crea una nueva variable filter\_\$, que adjudica valor 1 a los sujetos que van a entrar en los sucesivos análisis, y 0 a los sujetos excluidos. El análisis continuaría con la secuencia de comandos habitual:

Analizar → Modelo lineal general → Medidas repetidas → Nombre del factor intra-sujetos: Días; Número de niveles: 2 → Añadir → Definir → Variables intrasujeto: T1, T15 → Factores intersujetos: Terapia → Aceptar (análisis)

Declarando que el factor Días tiene dos niveles e indicando que éstos son T1 y T15 excluimos, como pretendíamos, el tercer nivel. Los resultados de SPSS son comprensibles ya para el lector. Si el lector realiza el análisis se percibirá de que la hipótesis no era incorrecta y que las diferencias entre T1 y T15 son iguales en A1 que en A2. Pero debe caer en la cuenta de que también puede interpretarse en la otra dirección, es decir, que las diferencias entre A1 y A2 son las mismas en T1 que en T15. Como ejercicio el lector puede intentar resolver si la diferencia entre T15 y T25 es mayor en A1 que en A2. El contraste de interacción significativo que obtendrá puede interpretarse en el sentido de que la diferencia entre T15 y T25 no es la misma en A1 que en A2. Examinando las medias podrá comprobar que lo que ocurre realmente es que la diferencia es mayor en A1 que en A2.

### 2.3. INTRODUCCIÓN A LOS DISEÑOS CON COVARIANTES

La actividad de diseño de experimentos puede definirse mediante dos términos clave: manipulación y control. Manipulación de variables independientes para producir efectos en las variables dependientes. Control de variables relevantes o extrañas, para intentar lograr que la variabilidad observada en las variables dependientes pueda ser atribuida únicamente a nuestras manipulaciones. En muchas ocasiones el control puede realizarse a través del diseño. Por ejemplo, eliminando variables relevantes, asignando aleatoriamente los sujetos a los tratamientos, etc. Sin embargo, no siempre este control ideal y directo es posible, particularmente cuando la investigación se produce en contextos aplicados o en los que no es posible la aleatorización. En esas situaciones el investigador puede intentar realizar un control estadístico. Los diseños en los que se introduce un control estadístico se denominan, en general, diseños de covarianza. El procedimiento que se emplea para lograr el objetivo de control es muy sencillo: las diferencias entre los sujetos después de recibir tratamientos diferentes son ajustadas teniendo en cuenta las diferencias entre ellos anteriores a la aplicación de las variables independientes. Naturalmente el ajuste sólo puede producirse cuando se conoce de cada sujeto, además de sus puntuaciones en las variables dependientes, su puntuación en las mismas o en otras variables relacionadas con ellas y medidas antes del tratamiento. Estas variables no contaminadas por las variables independientes se llaman covariantes o covariados. La lógica sería, pues, la siguiente: si los sujetos son diferentes en el covariado que está relacionado con las variables dependientes, se esperaría que fuesen diferentes de manera proporcional en las variables dependientes. Por el contrario, si los sujetos fuesen iguales en el covariado, esperaríamos que fuesen también iguales en las variables dependientes. Lo importante es que las diferencias entre los sujetos en las variables dependientes que son esperadas a partir del covariado pueden descartarse. De esta forma, el efecto principal del análisis de covarianza será reducir las diferencias entre los sujetos (por suprimir las esperadas desde el covariado), o, lo que es lo mismo, reducir el tamaño del error.

Para que el análisis de covarianza pueda realizarse de manera efectiva es necesario que covariados y variables dependientes estén relacionados linealmente entre sí. Quizá el caso más claro es aquel en el que covariado y variable dependiente son la misma variable. Por ejemplo, si se realiza una medición del BDI antes de aplicar la terapia y otra tras finalizarla. El covariado podría ser la medición pretratamiento y la variable dependiente la medición postratamiento. Puesto que se asume una relación lineal entre ambos, es posible usar el análisis de regresión lineal para predecir la variabilidad de la variable dependiente a partir del covariado.

El análisis de covarianza puede entenderse, pues, como un análisis de varianza en el que se suprime la parte de la variabilidad de cada fuente (efectos y errores) que es predecible a partir de la relación lineal entre covariados y variables dependientes. A continuación exemplificaremos el análisis de covarianza supo-

niendo que se ha manipulado una variable entre grupos, la terapia a tres niveles A1, A2 y A3, y que se ha realizado una medición de la depresión mediante BDI antes de aplicar el tratamiento (PRE o covariado) y otra tras 25 días de terapia (POST o variable dependiente). Los datos ficticios aparecen en la tabla siguiente:

A1		A2		A3	
PRE	POST	PRE	POST	PRE	POST
28	19	23	20	23	20
27	18	22	18	21	18
30	21	21	18	24	21
29	20	19	16	21	18
28	19	20	17	20	17

Cuadro 2.5. Comparación del análisis de varianza sobre la variable post con el análisis de covarianza de la variable post usando como covariado la variable pre

#### 1) Análisis de varianza

##### Pruebas de los efectos intersujetos

Variable dependiente: POST

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Modelo corregido	6,533 <sup>a</sup>	2	3,267	1,581	,246
Intersección	5226,667	1	5226,667	2529,032	,000
TERAPIA	6,533	2	3,267	1,581	,246
Error	24,800	12	2,067		
Total	5258,000	15			
Total corregido	31,333	14			

<sup>a</sup> R cuadrado = ,209 (R cuadrado corregido = ,077)

#### 2) Análisis de covarianza

##### Pruebas de los efectos intersujetos

Variable dependiente: POST

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Modelo corregido	30,572 <sup>a</sup>	3	10,191	147,198	,000
Intersección	,794	1	,794	11,469	,006
PRE	24,038	1	24,38	347,222	,000
TERAPIA	16,044	2	8,022	115,876	,000
Error	,762	11	6,923E-02		
Total	5258,000	15			
Total corregido	31,333	14			

<sup>a</sup> R cuadrado = ,976 (R cuadrado corregido = ,969)

Si no tuviésemos en cuenta el covariado, realizaríamos un análisis de varianza para un diseño con un solo factor entre grupos, lo que, como hemos visto antes, puede hacerse mediante el programa ANOVA del grupo Comparar medias o mediante el programa univariante del grupo Modelo lineal general. Los resultados (véase Cuadro 2.5. en página anterior) indicarían en ese caso que no hay efecto diferencial de la terapia. Sin embargo, si tenemos en cuenta el covariado las cosas pueden cambiar de manera sustancial, puesto que, como puede apreciarse inspeccionando la tabla de datos, parece que los grupos no son iguales *a priori*: los sujetos de A1 parecen puntuar más en el pre que el resto de los grupos. El análisis mediante SPSS puede realizarse con el grupo de programas del modelo lineal general mediante la siguiente secuencia de comandos:

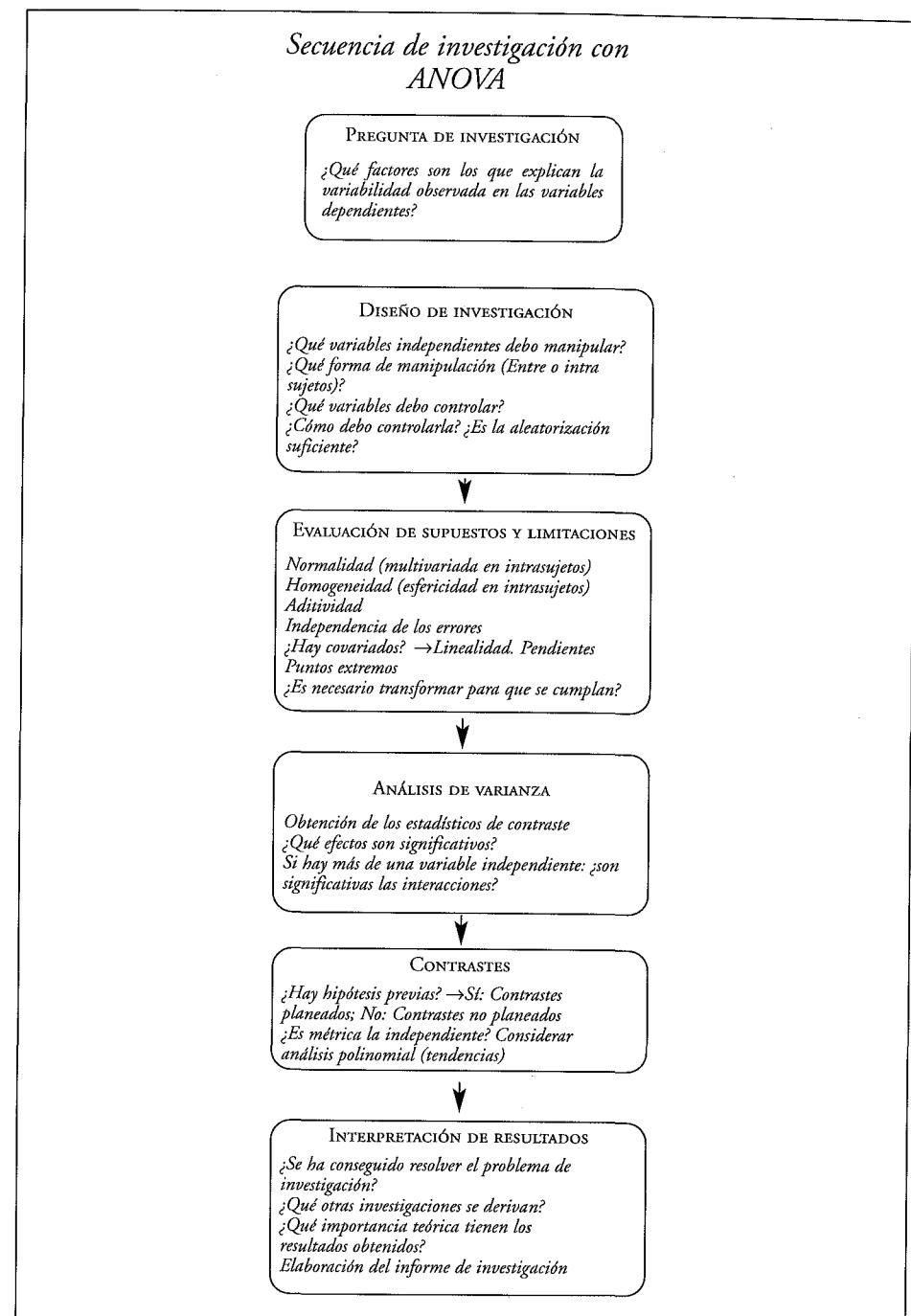
Analizar → Modelo lineal general → Univariante → Dependientes: Post, Factores fijos: Terapia; Covariables: Pre → Aceptar (análisis)

SPSS proporciona como resultado una tabla resumen de análisis de varianza en la que podemos apreciar (véase Cuadro 2.5.) que hay diferencias significativas entre las terapias. Si comparamos las dos tablas del Cuadro 2.5. comprobaremos que tener en cuenta el covariado se ha traducido en dos cosas, en primer lugar se ha incrementado la suma de cuadrados del efecto, y en segundo lugar, ha disminuido de modo importante la suma de cuadrados de error. Tener en cuenta el covariado permite detectar diferencias entre tratamientos que de otro modo podrían quedar enmascaradas debido a que los grupos no son equivalentes entre sí antes del experimento.

### 3. LA SECUENCIA DE INVESTIGACIÓN ORIENTADA AL ANÁLISIS MEDIANTE ANOVA

ANOVA suele ser la técnica analítica seleccionada por defecto en investigaciones en las que se manipulan una o más variables independientes, sean o no métricas, y se miden una o más variables dependientes métricas, cuando el interés recae, lo que suele ocurrir prácticamente siempre en esas situaciones, en determinar si la manipulación puede interpretarse como el agente causal que produce al menos parte de la variabilidad observada en las variables dependientes medidas (véase Cuadro 2.6.). Como hemos señalado a lo largo del capítulo, esta elección no es inadecuada, siempre y cuando las variables dependientes no correlacionen entre sí, y se cumplan los supuestos y eviten las limitaciones inherentes a la técnica. Por tanto, un aspecto fundamental del proceso de análisis, como en el resto de técnicas, debe ser la evaluación de los supuestos. En este sentido es importante recordar que hay dos diferencias importantes entre los diseños que incluyen sólo variables manipuladas entre grupos y los que incluyen alguna variable manipulada intrasujetos: los supuestos de normalidad y de homogeneidad.

Cuadro 2.6. La secuencia de investigación orientada al análisis mediante ANOVA



La normalidad es multivariada cuando hay variables intrasujeto, pero no cuando sólo hay variables entre grupos. Además, el supuesto de homogeneidad hace referencia a las varianzas y covarianzas (más estrictamente a la esfericidad) en el primer caso, pero sólo a varianzas en diseños entre grupos. El no cumplimiento de estos supuestos, particularmente del segundo, puede acarrear consecuencias graves sobre la toma de decisiones acerca de los efectos de las manipulaciones, por lo que es recomendable que el investigador trate de corregir sus efectos. Una de las estrategias consiste en realizar transformaciones que normalicen, incrementen la homogeneidad, o protejan contra los cambios en la forma de la distribución del estadístico de contraste bajo la hipótesis nula (por ejemplo: correcciones de los grados de libertad). La otra consiste en utilizar estrategias de análisis alternativas, como las pruebas no paramétricas o el análisis multivariado de varianza (en diseños intrasujetos).

En el diseño de investigaciones de este tipo es sumamente importante tener en cuenta que la habilidad para inferir relaciones causales es una función directa del grado de control que se haya ejercido sobre posibles causas alternativas, esto es, sobre las variables extrañas. En general, cuando la aleatorización es posible (en diseños entre grupos), si el número de sujetos es suficientemente grande el grado de control puede considerarse suficiente. Sin embargo, es importante no confiar en el azar para controlar variables extrañas cuyos efectos sean conocidos o potencialmente grandes. En esos casos es preferible eliminar esas variables (manteniéndolas constantes a lo largo del experimento) o bloquearlas (una forma explícita de hacer que aparezcan con los valores que el investigador considere oportunos). Por otro lado, es importante tener presente que en los diseños intrasujetos el control de variables de sujeto es, en general, facilitado por el diseño, puesto que el investigador no necesita ocuparse de aquellas variables que permanezcan estables en el sujeto a lo largo de la investigación.

Finalmente, en nuestra opinión es importante recordar que la calidad de una investigación se mide por el interés teórico de las preguntas que en ella se pretende responder, y que la formulación de preguntas importantes es una función directa del conocimiento que el investigador tenga del dominio en que su investigación se desarrolla. Pero no por ello debe olvidarse que la pregunta no es un indicador perfecto de las variables que se van a manipular y medir. La misma pregunta puede responderse con diferentes manipulaciones, pero no todas ellas proporcionarán una información con suficientes garantías. En este sentido creemos que es necesario evitar una tentación frecuente entre los investigadores noveles, la de manipular un gran número de variables en cada experimento, con la idea subyacente de atender a más de una hipótesis alternativa de investigación a la vez. Aunque no siempre, es frecuente que los resultados de una investigación semejante sean tan complejos que su interpretación sea farragosa y a menudo inútil. Es preferible, en ese caso, realizar una serie experimental en la que de modo sucesivo vayan resolviéndose las hipótesis en conflicto.

#### 4. LECTURAS RECOMENDADAS

- ANGUERA, M. T. y cols., *Métodos de investigación en Psicología*, Madrid, Síntesis, 1995.  
 ARNAU, J., *Psicología experimental: Un enfoque metodológico*, México, Trillas, 1978.  
 — *Diseños experimentales en Psicología y Educación*, vol. I, México, Trillas, 1981.  
 LEÓN, O. G. y MONTERO, I., *Diseño de investigaciones. Introducción a la lógica de la investigación en psicología y educación*, Madrid, McGraw-Hill, 1997.  
 PASCUAL, J.; FRÍAS, D. y GARCÍA, F., *Manual de Psicología Experimental*, Barcelona, Ariel, 1996.  
 PELEGRIÑA, M. y SALVADOR, F., *La investigación experimental en psicología: Fundamentos científicos y técnicas*, Málaga, Aljibe, 1999.

#### LECTURAS COMPLEMENTARIAS

- BISHOP, T. A., «Heteroscedastic ANOVA, MANOVA, and multiple comparisons», Phd Thesis, Ohio State Univ. Columbus, no publicada, 1976.  
 BRYK, A. S. y RAUDENBUSH, S. W., «Heterogeneity of variance in experimental studies: A challenge to conventional interpretations», *Psychological Bulletin*, 104, 1988, págs. 396-404.  
 BOEHNKE, K., «F- and H-test assumptions revisited», *Educational and Psychological Measurement*, 44, 1984, págs. 609-617.  
 COHEN, J., *Statistical power analysis for the behavioral sciences*, Nueva York, Academic Press, 1988.  
 COX, D. R., *Planning of experiments*, Nueva York, Wiley, 1958.  
 HAASE, R. F.; ELLIS, M. V. y LADANY, N., «Multiple criteria for evaluating the magnitude of experimental effects», *Journal of Counseling Psychology*, 36, 1989, págs. 511-516.  
 HOWELL, D. C., *Statistical Methods for Psychology*, Boston, MA, Duxbury Press, 1989.  
 JUDD, Ch. M.; McCLELLAND, G. H. y CULHANE, S. E., «Data analysis: Continuing issues in the everyday analysis of Psychological data», *Annual reviews of Psychology*, 46, 1995, págs. 433-465.  
 RASMUSSEN, J. L., «An evaluation of parametric and nonparametric test on modified and non-modified data», *British Journal of Mathematical and Statistical Psychology*, 39, 1986, págs. 213-220.  
 RASMUSSEN, J. L. y DUNLAP, W. P., «Dealing with nonnormal data: Parametric analysis of transformed data vs nonparametric analysis», *Educational and Psychological Measurement*, 51, 1991, págs. 809-820.  
 RIBA, M.ª D., *Modelo lineal de análisis de la variancia*, Barcelona, Herder, 1990.  
 RUTHERFORD, A., *Introducing Anova and Ancova*, Londres, Sage, 2000.  
 TOMARKEN, A. J. y SERLIN, R. C., «Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures», *Psychological Bulletin*, 99, 1986, págs. 90-99.  
 WILCOX, R. R., «New design in analysis of variance», *Annual Review of Psychology*, 38, 1987, págs. 29-60.

## 5. EJERCICIOS

A continuación encontrará una serie de problemas, trate de resolverlos mediante SPSS empleando las técnicas analíticas adecuadas. Encontrará las soluciones en la página siguiente.

1. Un defensor de la hipótesis de que el recuerdo del material aprendido depende sobre todo del tipo de organización del material en la fase de estudio utilizó 30 sujetos (15 estudiantes universitarios y 15 personas de edad equivalente que no han conseguido terminar bachiller) para ponerla a prueba. Este investigador piensa que una lista de palabras organizadas de manera conceptual se recordará mejor que organizadas en función de su semejanza fonológica, y ambas serán preferibles a una presentación aleatoria del material. Cada sujeto fue asignado al azar a uno de tres tratamientos, con la restricción de que el número de universitarios en cada condición fuese igual al número de no universitarios. En la primera condición los sujetos reciben una lista de palabras organizadas semánticamente; los sujetos de la segunda la reciben organizada fonológicamente y los de la tercera la reciben sin ningún tipo de organización. Los sujetos pasaron la tarea de forma individual. Los resultados, en número de palabras recordadas correctamente, fueron los siguientes:

Sujetos	CONDICIÓN		
	SEMÁNTICA	FONOLÓGICA	NO ORGANIZ.
1	13	7	12
2	15	4	11
3	12	4	4
4	10	1	9
5	12	10	5
6	6	7	10
7	7	5	2
8	6	9	8
9	10	5	3
10	6	8	6

- a) ¿Qué diseño de investigación se ha empleado (Especifique las variables)?
- b) ¿Qué controles de variables relevantes se han utilizado?
- c) ¿Se cumplen los supuestos subyacentes al diseño?
- d) ¿Produce efecto significativo la variable independiente?
- e) ¿Lleva razón el investigador en su primera hipótesis?
- f) ¿Puede mantenerse la segunda hipótesis?

2. En el estudio anterior, con respecto a la variable nivel de estudios, otro investigador piensa que debería interactuar con el tipo de organización. Los primeros 5 sujetos de cada condición son universitarios y los cinco últimos tienen estudios primarios. ¿Lleva razón? ¿Qué potencia tiene la prueba? ¿Qué magnitud tiene cada efecto?
3. Suponga ahora que sólo se emplearon 10 sujetos en el estudio relatado en 1, y que cada uno de los sujetos recibió todos los tipos de organización. La secuencia en que recibieron los tres niveles de la variable fue seleccionada al azar para cada sujeto.
  - a) ¿Qué diseño se ha empleado?
  - b) ¿Se cumple el supuesto de esfericidad?
  - c) ¿Produce efectos diferenciales la variable independiente?
  - d) Compare los resultados de este ejercicio con los obtenidos en el experimento 1 y discuta qué efectos produce el cambio en la forma de manipular la variable independiente tipo de organización.
4. Imagine ahora que, teniendo en cuenta el cambio de manipulación mencionado en la pregunta 3, los sujetos 1 a 5 eran estudiantes universitarios y el resto eran no universitarios.
  - a) ¿Qué diseño tendría?
  - b) ¿Qué efectos son significativos?
  - c) ¿Es la diferencia entre organización semántica y física la misma en ambos niveles de estudio?

## RESPUESTAS A LOS EJERCICIOS

1.
  - a) El diseño es claramente un unifactorial entre grupos. El factor tipo de organización del material está manipulado entre grupos a tres niveles. La variable dependiente es el recuerdo correcto.
  - b) Aparece controlada por bloqueo la variable nivel de estudios, y el resto están controladas debido a la asignación aleatoria de los sujetos a los grupos.
  - c) Independencia: los sujetos de cada grupo son diferentes y pasaron individualmente la tarea, no hay razones para pensar que los datos sean dependientes.

Normalidad: Primero comprobaremos el supuesto de normalidad. SPSS ofrece dos pruebas, Shapiro-Wilk y Kolmogorov-Smirnov. Se encuentran en Analizar → Estadísticos descriptivos → Explorar → Gráficos → Gráficos con pruebas de normalidad.

Debemos comprobar que la distribución es normal en cada grupo. La desviación de la normal no es significativa según el estadístico de Shapiro-Wilk en ninguno de los grupos. Los estadísticos son: 0,899, 0,968 y 0,955. Esta conclusión puede confirmarse con el test de Kolmogorov-Smirnov, y examinando la asimetría y la curtosis.

**Homogeneidad de varianzas:** La prueba de Levene,  $F(2,27)=0,764$ ,  $p>0,45$ , indica que las varianzas de grupo son homogéneas. Esta prueba se realiza mediante el comando Comparar Medias → AÑOVA de un factor → Opciones → Homogeneidad de varianzas

- d) La variable tipo de organización del material produce efectos significativos, puesto que  $F(2,27)=3,609$ , con una media de cuadrados de error  $MSE=10,152$ , y nivel de significación  $p<=0,05$ .
- e) Los coeficientes para el contraste de la primera hipótesis son (1 -1 0) y para la segunda (1 1 -2). Se introducen con la siguiente secuencia de comandos: Analizar → AÑOVA de un factor → Contrastes → 1 → Añadir → -1 → Añadir → 0 → Añadir → Siguiente → 1 → Añadir → 1 → Añadir → -2 → Añadir → Continuar. Hay diferencias significativas entre la organización semántica y la fonológica,  $t(27)=2,579$ ,  $p<0,05$ , pero ambas juntas no son superiores a la aleatoria,  $t(27)=0,689$ ,  $p>0,45$ . Esta situación no cambia cuando asumimos que las varianzas no sean homogéneas. Un método alternativo para comprobar las diferencias sería realizar comparaciones *a posteriori*. En ese caso se podría determinar si cada condición difiere del resto. Por ejemplo, usando la prueba de Tukey, comprobaríamos que ninguna de las dos organizaciones es superior a la aleatoria, aunque entre ellas sí hay diferencias, como ha quedado indicado más arriba.
2. Si la variable nivel de estudios se considera un factor, entonces el diseño se convierte en un factorial entre grupos, puesto que en cada tipo de organización se incluyen dos subgrupos con diferente nivel de estudios. El análisis de varianza a aplicar sería pues el incluido en el grupo de programas Modelo lineal general → Univariante. El AÑOVA factorial indica que son significativos los efectos principales de ambos factores,  $F(2,24)=4,874$ ,  $p<0,02$  y  $F(1,24)=4,262$ ,  $p<=0,05$ , respectivamente para organización y estudios, y de su interacción,  $F(2,24)=4,102$ ,  $p<0,03$ . La media de cuadrados de error es común,  $MSE=7,517$ .
3.
  - a) Se trataría de un diseño de medidas repetidas de un único factor.
  - b) El supuesto de esfericidad es contrastado por defecto en SPSS cuando se emplea el comando Medidas Repetidas del grupo

Modelo lineal general. Tanto la W de Mauchly (0,963;  $\chi^2(2)=0,298$ ,  $p=0,862$ ) como los valores de épsilon de Greenhouse-Geisser (0,965) y de Huynh-Feldt (1,00) indican que el supuesto de esfericidad parece cumplirse.

- c) El análisis de varianza realizado con el comando medidas repetidas del grupo Modelo lineal general indica que hay efectos significativos del tipo de organización,  $F(2,18) = 3,635$ ,  $MSE=10,078$ ,  $p<0,05$ .
- d) Aunque no de manera rotunda, parece claro que los diseños intrasujetos son más potentes que los entre grupos, la media de cuadrados de error tiende a ser menor, a igualdad de otras circunstancias, en los primeros.
4.
  - a) El diseño sería de medidas parcialmente repetidas o mixto, puesto que tendremos una variable que sólo puede ser manipulada entre grupos (el nivel de estudios) y otra manipulada intrasujetos (la organización del material).
  - b) El análisis de varianza indica que efectos significativos del factor organización,  $F(2,16)=4,895$ ,  $MSE=7,483$ ,  $p<0,05$ , y de la interacción estudios por organización,  $F(2,16)=4,120$ ,  $MSE=7,483$ ,  $p<0,05$ , pero sólo marginalmente del factor nivel de estudios,  $F(1,8)=4,224$ ,  $MSE=7,583$ ,  $p=0,074$ .
  - c) La cuestión de si una diferencia entre dos niveles de un factor es la misma en un nivel u otro de un segundo factor se resuelve mediante interacciones parciales. La forma de hacerlas en SPSS consiste en seleccionar sólo los niveles adecuados del factor que se compara. Por tanto, realizaríamos un análisis como en b, pero excluyendo ahora el nivel tercero de la variable organización. Encontramos que la interacción de organización (con dos niveles) y estudios es significativa,  $F(1,8)=10,938$ ,  $MSE=5,6$ ,  $p<0,05$ . Por tanto, la diferencia no es la misma. Examinando las medias comprobaremos que la diferencia semántica-física es mayor en el nivel universitario.

## CAPÍTULO III

### Análisis de datos categóricos

#### OBJETIVOS DE APRENDIZAJE

- 1) Distinguir el análisis de frecuencias de múltiples vías de otras técnicas orientadas a la descripción de relaciones entre variables, como la regresión múltiple, o la regresión logística.
- 2) Conocer en qué contextos de investigación es adecuado el análisis de frecuencias.
- 3) Conocer la clase de problemas de investigación que el análisis de frecuencias ayuda a resolver.
- 4) Conocer los fundamentos lógicos y matemáticos del análisis.
- 5) Saber calcular y conocer el significado de los parámetros.
- 6) Conocer los diferentes tipos de modelos (jerárquicos y no jerárquicos).
- 7) Conocer las limitaciones y supuestos del análisis.
- 8) Saber interpretar los resultados de los análisis en relación al problema de investigación.

#### 1. CONCEPTOS FUNDAMENTALES DEL ANÁLISIS DE DATOS CATEGÓRICOS

En la investigación psicológica es relativamente frecuente la manipulación de variables que están medidas en una escala nominal u ordinal (variables categóricas). Es menos frecuente, sin embargo, que las variables de respuesta sean medidas en una escala semejante. Sin embargo, no siempre estamos interesados en estudiar relaciones entre variables manipuladas y variables de respuesta, o dependientes en sentido estricto. En ocasiones la investigación se plantea con el objetivo de determinar si hay relaciones entre variables sin que haya manipulación explícita alguna. Supongamos, por ejemplificar, que se ha pasado una

encuesta a una muestra representativa de la población española y que entre otras, se ha hecho una pregunta sobre el nivel educativo (primaria, secundaria, bachiller, universitaria) y otra sobre el interés con el que se siguen los acontecimientos políticos (bajo, medio, alto). Una cuestión interesante podría ser si existe algún tipo de relación entre ambos ítem. Evidentemente, para resolver esa cuestión parece lógico construir una tabla que contenga el número de personas que han estudiado primaria y tienen interés bajo, medio y alto en política, también para secundaria, etc. Esa tabla tendría dos entradas, una por ítem, con cuatro filas y dos columnas, y podría tener, supongamos, el siguiente aspecto:

		Interés político	
Educación	Bajo	Alto	
Primaria	90	10	
Secundaria	82	33	
Bachiller	73	47	
Universidad	60	90	

Lo que nos interesaría sería encontrar un modelo, una explicación de las frecuencias que aparecen en las casillas de la tabla. En un caso tan simple como éste debe resultar claro al lector que esa explicación, ese modelo, en lo esencial, podría basarse en tres cosas diferentes. Es obvio que los elementos de la explicación han de ser las variables que hemos medido (una, la otra, o ambas sumadas), y posiblemente la interacción entre ambas. Como se recordará la interacción implica que los puntuaciones (las frecuencias en este caso) de una variable son una función de la otra variable. Por ejemplo, el interés político puede ser bajo con mayor frecuencia que alto en personas con educación primaria, pero al revés en personas con bachiller o estudios universitarios.

El análisis de frecuencias de múltiples vías es una técnica orientada al estudio de relaciones entre variables de tipo categórico (o categorizadas artificialmente). Una variable continua puede ser categorizada fácilmente. Por ejemplo, si en la encuesta preguntamos el sueldo y después dividimos a los sujetos según tengan un sueldo bajo, medio o alto, hemos categorizado una variable métrica. El objetivo fundamental del análisis es encontrar un modelo que combina las fuentes explicativas (las variables medidas y su interacción) para explicar la relación entre esas variables. La forma más frecuente de realizar el análisis consiste en construir tablas de múltiples entradas (tantas como variables medidas) que contienen tantas casillas como el producto de los niveles de las variables medidas. Por ejemplo, si hemos medido dos variables, una con cuatro valores y otra con dos, el número de casillas de la tabla es  $4 \times 2$ . Si añadimos una tercera variable, también con dos valores, entonces el número de entradas se eleva a 3 y el de casillas se eleva a  $4 \times 2 \times 2$ . Las casillas contienen el número de sujetos que tienen el valor adecuado de cada una de las variables que definen esa casilla. A continuación, se intenta estudiar la asociación entre variables. Una forma frecuente consiste en construir un modelo lineal que incluya el menor número de factores explicativos necesarios. Habitualmente se comienza introduciendo

do en el modelo todos los posibles (por ejemplo, en el caso de dos variables, cada una de ellas más la interacción) y a continuación se intenta eliminar el máximo posible sin que la capacidad de explicar las frecuencias observadas disminuya significativamente. En el caso de dos variables, podemos continuar eliminando la interacción y comprobando si eso repercute negativamente en la capacidad explicativa del modelo. En principio no es necesario que el papel explicativo de las variables esté definido. Esto es, todas pueden considerarse variables medidas sin más. Sin embargo, en algunas investigaciones es posible tratar a algunas de las variables como independientes y a las otras como dependientes. Por ejemplo, podríamos considerar que el nivel educativo es una posible causa del interés político. En este sentido el análisis de frecuencias puede entenderse como un análisis explicativo, más que predictivo, puesto que se emplearía para contrastar la hipótesis de que la variable independiente es la causa de los cambios observados en la dependiente.

Aún hay una tercera aproximación, de carácter descriptivo, en la que la existencia de relaciones se deriva de la comparación entre proporciones. Por ejemplo, en la tabla de arriba podríamos calcular la probabilidad de que el interés sea bajo dado que se ha realizado estudios primarios ( $90/(90+10) = 0,9$ ) y compararla con la probabilidad de que el interés sea bajo dado que se han realizado estudios universitarios ( $60/(60+90) = 0,3$ ). Lógicamente si no hubiese asociación, esperaríamos que las probabilidades fuesen iguales. Una forma diferente consiste en trabajar con razones. Así, la razón de interés bajo respecto de interés alto cuando se han realizado estudios primarios es  $9/(=90/10)$ , mientras que cuando se han realizado estudios universitarios la razón se reduce a  $0,67 (=60/90)$ . Parece que es más frecuente tener bajo interés político si se tienen estudios primarios que universitarios. Esto puede apreciarse mejor si calculamos el cociente entre ambas proporciones:  $9/0,67 = 13,5$ . La razón interés bajo/interés alto es más de 13 veces mayor con estudios primarios que universitarios. De nuevo, si las variables no estuvieran asociadas lo que esperaríamos es que las razones fuesen iguales con independencia del nivel de estudios.

## 2. LAS PREGUNTAS QUE EL ANÁLISIS DE FRECUENCIAS AYUDA A RESOLVER

El análisis de frecuencias tiene el objetivo fundamental de comprobar la existencia de relaciones entre variables categóricas (o categorizadas). Esto implica que nuestro interés reside en hacer predicciones sobre las frecuencias de las casillas. Nótese que esto implica que las frecuencias son consideradas la variable dependiente, la variable a explicar, mientras que los factores y su interacción son los posibles agentes explicativos, las independientes. Por ejemplo, si sabemos que el nivel de estudios es primaria ¿podemos predecir si el interés político será alto o bajo? Cuando tenemos más de dos variables medidas, discriminar cuáles están asociadas con cuáles otras será, por tanto, un objetivo

prioritario. El análisis de frecuencias completo intenta, por tanto, proporcionar información sobre los efectos en una variable dependiente, los parámetros necesarios para formular el modelo, la magnitud de los efectos y la fuerza de la asociación entre variables. Veamos cada uno de estos objetivos concretos.

## 2.1. PARÁMETROS

Los factores explicativos contribuyen en alguna medida a las frecuencias observadas. La contribución de cada factor es capturada por un conjunto de coeficientes, parámetros, tantos como valores tenga. La combinación de parámetros permitirá explicar las frecuencias observadas.

## 2.2. IMPORTANCIA DE LOS NIVELES

La comparación entre los tamaños de los parámetros nos permitirá hacernos una idea bastante cabal de la importancia de cada factor para la explicación. Es frecuente emplear no coeficientes directos, sino estandarizados, para facilitar el proceso de comparación. De este modo los mayores coeficientes son indicadores de una mayor importancia de los niveles de un factor.

## 2.3. MAGNITUD DE LA RELACIÓN

El hecho de que un modelo proporcione una explicación significativa de las frecuencias observadas no es un buen indicador del grado de relación entre variables.

## 2.4. INFLUENCIA DE VARIABLES INDEPENDIENTES SOBRE UNA VARIABLE DEPENDIENTE

Como ya hemos dicho, la variable dependiente en el análisis estándar son las frecuencias de la tabla. Sin embargo, también lo hemos dicho, hay investigaciones en las que una de las variables funciona como dependiente, es la variable que debe ser explicada por las otras, que son independientes. En estos casos el análisis de frecuencias es apropiado, pero no es la única alternativa disponible. Puede ser más apropiado el uso de técnicas que expondremos más adelante, como la regresión logística (véase Capítulo XI). De cualquier modo, debe entenderse que en este sentido se trata, en la práctica de realizar un análisis semejante al de varianza, aunque con variables no métricas. Esto es interesante porque nos pone sobre la pista de una última cuestión relevante en el análisis de frecuencias. Hemos podido comprobar que nuestra tabla ejemplo tiene

una variable con cuatro niveles, y otra con dos. Cuando sólo hay dos no hay necesidad de realizar comparaciones adicionales. Sin embargo, cuando hay más de dos, uno puede preguntarse cómo difieren unos niveles de otros. Por tanto, se estaría interesado en realizar comparaciones entre pares de niveles. Las preguntas que resuelven las comparaciones están claras, por ejemplo, ¿difieren universitarios y bachilleres en el interés político?

## 3. EL ANÁLISIS DE TABLAS DE FRECUENCIA: MODELOS LOGARÍTMICO-LINEALES

Cuando el objetivo es construir un modelo de efectos, el proceso se divide habitualmente en tres etapas: Primero se comprueban los efectos. Este paso es imprescindible siempre que no se haya hipotetizado un modelo concreto, puesto que puede ahorrar gran parte de trabajo posterior que podría ser inútil si no hay efectos. Si el investigador supone que hay un modelo concreto que puede explicar las frecuencias, entonces debe proceder directamente a su estudio en los pasos segundo y tercero. En segundo lugar se construye el modelo, y en tercer lugar se evalúa. Los pasos dos y tres deben repetirse para contrastar modelos alternativos más simples pero con la misma potencia predictiva. Veamos cada uno de estos pasos aplicado a los datos de la siguiente tabla, que recoge los datos obtenidos en una investigación ficticia de encuesta en la que se ha preguntado por el interés político (bajo, alto), el nivel educativo (primaria, secundaria, bachiller, universidad) y la integración familiar (baja, media, alta). Sean las siguientes frecuencias:

Tabla 3.1.

		Educación				
Integración	Interés	Primaria	Secundaria	Bachiller	Universidad	Totales
		15	10	9	8	42
Alta	Bajo	1	2	5	9	17
	Totales	16	12	14	17	59
Media	Bajo	4	5	6	6	21
	Alto	2	3	3	5	13
	Totales	6	8	9	11	34
Baja	Bajo	3	4	7	5	19
	Alto	1	2	2	3	8
	Totales	4	6	9	8	27
	Totales	26	26	32	36	120

### 3.1. COMPROBACIÓN DE EFECTOS

Lo primero que debemos considerar es si hay algún tipo de efecto en la tabla. Parece claro que si las variables no estuviesen relacionadas en modo alguno las frecuencias de las casillas deberían ser iguales entre sí. Por el contrario, las casillas tendrían diferente frecuencia si hubiese algún tipo de efecto. Así, la primera comprobación que podemos hacer consiste en computar las frecuencias esperadas de las casillas ( $F_e$ ) bajo el supuesto de que no hubiese efectos. Está claro que esto implica repartir el conjunto de individuos entrevistados entre las casillas ( $c$ ), por tanto, la frecuencia de cada casilla debería ser:

$$F_e = N/c = 120/(4 \times 3 \times 2) = 5$$

A continuación parece lógico que la relación entre las frecuencias de casilla observadas y las esperadas sea una medida adecuada de si hay efecto. Esta idea queda capturada en el estadístico de razón de verosimilitud, que se define como:

$$G^2 = 2 \sum (f_o) \ln \left( \frac{f_o}{F_e} \right) \quad (\text{Ec. 3.1})$$

donde i, j, k corresponden a cada factor y  $f_o$  son las frecuencias observadas. Este estadístico se distribuye como una chi-cuadrado con (número de casillas-1) grados de libertad. Aplicando a nuestro ejemplo tendríamos que:

$$G^2 = 2[(15)\ln(15/5) + (10)\ln(10/5) + \dots + (3)\ln(3/5)] = 72,94,$$

que es mayor que la chi-crítica con 23 grados de libertad a un nivel de significación de 0,05 (35,17). Por tanto, alguno de los efectos ha de ser significativo. Los pasos siguientes van orientados en consecuencia hacia la identificación de cuáles pueden ser esos efectos significativos. Es importante proceder a una exploración sistemática, comenzando por los efectos principales de los factores, continuar luego con las interacciones de primer orden (de dos variables) y finalizar con las de segundo orden (tres variables). La estrategia en todos los casos será similar a la que hemos empleado para la prueba global: Determinar las frecuencias esperadas y comprobar la separación respecto de las observadas.

Los efectos principales son los debidos a un factor sin considerar al resto, por tanto, se obtienen sumando frecuencias a través de los demás factores; además, tendremos tantos efectos como factores. Podemos comenzar el análisis por la educación recibida. Las frecuencias correspondientes a cada nivel serían las siguientes:

Primaria	Secundaria	Bachiller	Universidad
26	26	32	36

Las frecuencias esperadas serían  $F_{e(\text{EDUCACIÓN})} = 120/4=30$ , luego

$$G^2_{\text{EDUCACIÓN}} = 2[(26)\ln(26/30) + \dots + (36)\ln(36/30)] = 2,37,$$

que no es significativo con 3 grados de libertad y nivel de significación, p, de 0,05.

De la misma manera procederíamos con los efectos principales de Integración familiar ( $G^2_{\text{INTEGRACIÓN}} = 13,57$ ,  $p<0,05$ ) e Interés político ( $G^2_{\text{INTERÉS}} = 16,52$ ,  $p<0,05$ ).

Los efectos de las interacciones de primer orden serían nuestro siguiente objetivo. Tengamos presente que con tres variables podemos tener tres interacciones de este tipo: EDUCACIÓN x INTEGRACIÓN; EDUCACIÓN x INTERÉS e INTEGRACIÓN x INTERÉS. La cuestión reside en obtener una tabla que contenga las mismas frecuencias marginales observadas que esperadas, pero en la que no hay interacción de primer orden presente. Para conseguir esto es necesario realizar un proceso iterativo de tres ciclos (uno por interacción). Sin embargo, la complejidad del proceso es enorme puesto que han de considerarse también las interacciones de orden superior. En nuestro caso la de segundo orden (tres variables), implicaría un proceso iterativo que requeriría diez pasos. Es obvio, pues que el cálculo manual no es, ni mucho menos aconsejable. Desafortunadamente SPSS carece de un subprograma capaz de evaluar cada efecto por separado (principales, de primer orden, de segundo, etcétera). Por ello, recomendamos al lector que emplee STATISTICA (versión 6.0), que sí permite esa evaluación. La secuencia de comandos sería la siguiente:

```
Statistics → Advanced linear/Non linear models → Log-linear analysis of frequency tables → Variables: Educación, Integración, Interés → Ok → Ok → Test of all marginal & partial associations models
```

STATISTICA proporciona dos test, uno de asociación marginal y otro de asociación parcial. Respecto de los efectos principales, el primero compara un modelo en el que se han incluido todos menos el que nos interesa estudiar contra otro en el que se han incluido también el que nos interesa estudiar. Esto es, si queremos estudiar el efecto de Educación, se compara un modelo que incluye Interés e Integración frente a otro que incluye esos dos más Educación. Naturalmente si la diferencia en el ajuste entre ambos modelos es significativa, eso implica que el efecto principal es necesario para mejorar el ajuste. El test de asociación parcial es idéntico al anterior para los efectos principales. Respecto de las interacciones de primer orden, el test de asociación marginal compara un modelo en el que se han incluido todos los efectos principales contra un modelo en el que se incluye además la interacción que nos interesa evaluar. Por otro

lado, el test de asociación parcial compara un modelo en el que se han incluido los efectos principales y todas las interacciones de primer orden, excepto la de interés contra un modelo que incluye también la de interés. La interpretación es semejante a la de los efectos principales.

Cuadro 3.1. Test de asociación parcial y asociación marginal en STATISTICA

Effect	Tests of Marginal and Partial Association (Multiway.sta)				
	Degrs. of Freedom	Prt. Ass. Chi-sqr.	Prt. Ass. p	Mrg. Ass. Chi-sqr.	Mrg. Ass. p
1	3	2,16043	0,539785	2,16043	0,539785
2	2	12,37264	0,002057	12,37264	0,002057
3	1	14,95110	0,000110	14,95110	0,000110
12	6	1,95314	0,923957	2,15644	0,904764
13	3	6,29570	0,098077	6,49898	0,089703
23	2	0,73098	0,693857	0,93422	0,626810

NOTA: Los números que aparecen en la columna efecto se refieren a los factores del siguiente modo:  
1=Educación, 2=Integración, 3=Interés. Así, 12=Educación por Integración

Si el segundo modelo ajusta significativamente mejor que el primero, entonces la interacción de interés es necesaria para predecir las frecuencias observadas. El Cuadro 3.1. presenta los resultados obtenidos para nuestros datos (las frecuencias observadas de la Tabla 1 pueden introducirse como fichero de datos en STATISTICA) con la secuencia de comandos especificada anteriormente. Nótese que tanto uno como otro test parecen indicar que un buen modelo explicativo debería incluir a los factores Integración, Interés y, posiblemente, la interacción Educación x Interés, puesto que su probabilidad se sitúa por debajo de 0,1, un buen criterio para introducir un efecto (STATISTICA lo usa por defecto cuando se le pide encontrar el mejor modelo). En esta aproximación queda por evaluar si la interacción de segundo orden (de los tres factores) hace o no una aportación significativa. Una aproximación razonable puede consistir en calcular las frecuencias esperadas mediante el programa de tabulación cruzada de SPSS. La secuencia de comandos sería

Analizar → Estadísticos descriptivos → Tablas de contingencia → Filas: Educación → Columnas: Interés → Capa 1 de 1: Integración → Casillas → Frecuencias: Esperadas → Continuar → Aceptar (análisis)

Y computar después  $G^2$  según la ecuación 3.1. En nuestro ejemplo, obtendríamos que  $G^2 = 11,08$ , que no es significativo a  $(4-1)(3-1)(2-1)$  grados de libertad y nivel de significación de 0,05.

La evaluación de los posibles efectos significativos permite, al menos, hacerse una idea del modelo que sería necesario construir para predecir las frecuencias de las casillas de la tabla. Sin embargo, debe tenerse en cuenta que no siempre la interpretación de estos análisis es sencilla, que no siempre permiten tomar decisiones claras sobre qué efectos incluir, y que, en cualquier caso, uno puede estar interesado más en encontrar el mejor modelo de predicción, aunque no resulte significativo.

### 3.2. CONSTRUCCIÓN DEL MODELO DE PREDICCIÓN

La construcción de un modelo de predicción es un proceso de decisión complejo en el que trata de establecer cuál es el modelo con el menor número de efectos que es capaz de ajustar los datos de la tabla de frecuencias. Si *a priori* se había definido un modelo posible, el investigador está legitimado para comenzar por contrastarlo. Por el contrario, cuando no hay un modelo previo, y se ha realizado la tarea de comprobar qué efectos (factores e interacciones entre ellos) hacen una aportación significativa, puede comenzarse por incluirlos en el modelo. En el caso, no recomendable, de que no se haya realizado examen alguno de efectos, el investigador se verá obligado a contrastar todos los modelos posibles. En cualquier caso, se trata de establecer una ecuación de predicción en la que la variable dependiente es el logaritmo de las frecuencias de casilla y las independientes son los efectos. Dicho formalmente, asumiendo tres factores, sean A, B y C:

$$\ln(F_{ijk}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC} + \varepsilon_{ijk} \quad (\text{Ec. 3.2})$$

Donde i, j y k son los subíndices de las casillas, correspondientes, respectivamente a los factores A, B y C; los  $\lambda$  son los parámetros del modelo, asociados a cada factor; m es la constante y  $\varepsilon_{ijk}$  es el error de predicción cometido en cada casilla. Como el lector habrá deducido, el modelo queda especificado cuando se conocen los parámetros de cada efecto, pero debe tenerse en cuenta que cada efecto tiene un parámetro por cada una de sus condiciones. Esto es, si el factor A tiene a niveles, tendremos que computar a parámetros, si C tiene c niveles, habrá que obtener c parámetros, por tanto, la interacción AxC tendrá asociados a·c parámetros. En nuestro ejemplo habrá 4 parámetros para Educación, 3 para Integración, 2 para Interés, 4x3 para Educación x Integración, etc. Además, la suma de los parámetros de cualquier efecto siempre es cero.

### 3.2.1. Tipos de modelos

Los modelos pueden clasificarse en función de dos criterios, el número de efectos que incluyen y su estructura.

#### 3.2.1.1. «Número de efectos»

En esencia pueden construirse dos tipos de modelos, los llamados saturados y los incompletos. Los modelos saturados o completos contienen todos los efectos posibles. Esto siempre se traduce en que se explica de manera perfecta el conjunto de las frecuencias de la tabla, y, por ende, el error de predicción es nulo. Estos modelos no interesan en general, puesto que *a priori* sabemos su poder explicativo. El segundo tipo de modelos, los incompletos, se caracterizan porque contienen solamente algunos de los efectos posibles. Por ejemplo, pueden incluirse algunos de los factores, no todos, y algunas de las interacciones, no todas. Pongamos por caso, podríamos incluir en el modelo, en nuestro ejemplo, los factores Integración e Interés, y su interacción.

El objetivo del análisis es, pues, simple. Se trata de encontrar un modelo incompleto que contenga el mínimo número de efectos necesarios para realizar una explicación satisfactoria de las frecuencias observadas. En otras palabras, un modelo cuyo error de predicción no sea significativo. Puesto que el estadístico de razón de verosimilitud,  $G^2$ , compara precisamente los valores observados con los predichos, eso implica que se necesita un modelo cuya razón de verosimilitud no sea significativa.

#### 3.2.1.2. «Estructura»

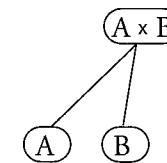
Atendiendo a su estructuración los modelos se clasifican en jerárquicos o anidados y no jerárquicos. Los modelos jerárquicos son los que cuando incluyen un efecto de nivel superior (por ejemplo, una interacción) también incluyen todos los componentes de ese término (por ejemplo, los efectos principales). Esto implica que su estructura es piramidal, como refleja la Figura 3.1. (véase página siguiente). Nótese cómo el modelo 1a) (A, B, AB) es un subconjunto del modelo más completo 1b) (A, B, C, AB, AC), es decir, 1a) está anidado dentro de 1.b), pero no ocurre lo mismo con el modelo 2a) y 2b).

El contraste de modelos difiere entre jerárquicos y no jerárquicos. Cuando se trata de modelos jerárquicos, suele compararse un modelo de orden superior con un modelo de orden inferior. Si la diferencia entre las razones de verosimilitud de ambos, que es otra razón de verosimilitud, no es significativa, entonces es preferible el modelo de nivel inferior. En otras palabras, si comparamos el modelo 1b) respecto al 1a), y no hay diferencias entre ellos, el mode-

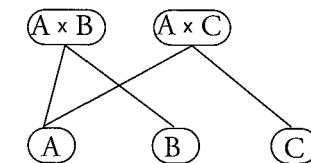
Figura 3.1. Modelos log-lineales según su estructura

#### 1) Modelos jerárquicos o anidados

a) modelo (A, B, AB)

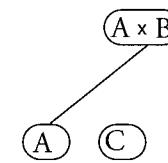


b) Modelo (A, B, C, AB, AC)

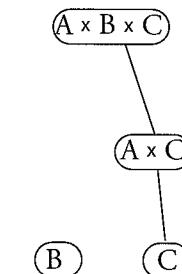


#### 2) Modelos no jerárquicos

a) Modelo (A, C, AB)



b) Modelo (B, C, ABC)



lo 1a) es una explicación más simple y preferible que 1b). El cálculo de la  $G^2$  de cada modelo puede hacerse fácilmente mediante SPSS a través del programa HILOGLINEAR, que aparece en el menú como Selección de modelos. Ilustraremos, primero, el del modelo más completo 1b), suponiendo que A=Integración, B= Interés y C=Educación. La secuencia de comandos sería la siguiente:

Analizar → Log-lineal → Selección de modelo → Factores: Integración → Definir rango → Mínimo: 1; Máximo: 3 → Continuar → Factores: Interés → Definir rango → Mínimo: 1; Máximo: 2 → Continuar → Factores: Educación → Definir rango → Mínimo: 1; Máximo: 4 → Continuar → Modelo → Construir términos: Interacción → Factores: Integración, Interés → Factores: Integración, Educación → Continuar → Opciones: Frecuencias, Residuos → Continuar → Construcción de modelos: Utilizar la eliminación hacia atrás → Aceptar (análisis)

La salida de SPSS aparece en el Navegador de Resultados en formato de texto. Puede copiarse y trasladarse a un procesador de texto o examinarse completa incrementando el tamaño del cuadro de texto. Es importante tener pre-

sente varias cuestiones. Primero, en HILOGLINEAR cuando pedimos una interacción el programa automáticamente incluye todos los términos de nivel inferior, de este modo, al introducir la interacción Integración por Interés, quedan incluidos los efectos principales de esos dos factores. Segundo, el uso del método de eliminación hacia atrás garantiza que se probarán los modelos de nivel inferior al que inicialmente hemos solicitado. Por tanto, el programa seleccionará el mejor modelo de nivel inferior que no difiera del solicitado por nosotros (si es que lo hay). El proceso de prueba iterativo aparece en el Cuadro 3.2.

Cuadro 3.2. Análisis log-lineal jerárquico. Modelo contrastado: (Integración, Interés, Integración x Interés, Integración x Educación)

***** HIERARCHICAL LOG LINEAR *****						
Backward Elimination (p = .050) for DESIGN 1 with generating class						
EDUCACION*INTEGRAC INTEGRAC*INTERES						
Likelihood ratio chi square = 11.75790 DF = 9 P = .227						
If Deleted Simple Effect is	DF	L.R.	Chisq	Change	Prob	Iter
(Si se elimina)						
EDUCACION*INTEGRAC	6		2.490	.8695	2	
INTEGRAC*INTERES	2		.934	.6268	2	
<b>Step 1</b>						
The best model has generating class						
INTEGRAC*INTERES						
EDUCACION						
Likelihood ratio chi square = 14.24826 DF = 15 P = .507						
If Deleted Simple Effect is	DF	L.R.	Chisq	Change	Prob	Iter
INTEGRAC*INTERES	2		.934	.6268	2	
EDUCACION	3		2.375	.4983	2	
<b>Step 2</b>						
The best model has generating class						
EDUCACION						
INTEGRAC						
INTERES						
Likelihood ratio chi square = 15.18251 DF = 17 P = .582						
If Deleted Simple Effect is	DF	L.R.	Chisq	Change	Prob	Iter
EDUCACION	3		2.375	.4983	2	
INTEGRAC	2		13.586	.0011	2	
INTERES	1		16.516	.0000	2	
<b>Step 3</b>						
The best model has generating class						
INTEGRAC						
INTERES						
Likelihood ratio chi square = 17.55764 DF = 20 P = .617						
If Deleted Simple Effect is	DF	L.R.	Chisq	Change	Prob	Iter
INTEGRAC	2		13.586	.0011	2	
INTERES	1		16.516	.0000	2	
<b>Step 4</b>						
The best model has generating class						
INTEGRAC						
INTERES						
Likelihood ratio chi square = 17.55764 DF = 20 P = .617						
The final model has generating class						
INTEGRAC						
INTERES						

### Análisis de datos categóricos

Nótese cómo se comienza examinando por separado la contribución de los términos de nivel superior del modelo (las interacciones de primer orden) y eliminando el que menos contribución realiza. En este paso es excluida la interacción Educación x Integración, algo que podíamos prever a partir del examen de contribuciones que habíamos realizado en la etapa anterior (véase Cuadro 3.1.), y, por tanto, quedan la interacción Integración x Interés, y los efectos principales de los tres factores.

A continuación se evalúan la interacción y el factor Educación. La interacción es la que menos aporta, por lo que es eliminada del modelo. Inmediatamente después se elimina el factor Educación, porque es el que menos aporta.

Cuadro 3.3. Frecuencias observadas (OBS), esperadas (EXP) y residuos (Residuals) según el modelo (Integración, Interés)

Factor	Code	OBS count	EXP count	Residual	Std Resid
EDUCACION	1				
INTEGRAC	1				
INTERES	1	15.0	10.1	4.92	1.55
INTERES	2	1.0	4.7	-3.67	-1.70
INTEGRAC	2				
INTERES	1	4.0	5.8	-1.81	-.75
INTERES	2	2.0	2.7	-.69	-.42
INTEGRAC	3				
INTERES	1	3.0	4.6	-1.61	-.75
INTERES	2	1.0	2.1	-1.14	-.78
EDUCACION	2				
INTEGRAC	1	10.0	10.1	-.08	-.02
INTERES	1	2.0	4.7	-2.67	-1.24
INTERES	2				
INTEGRAC	2				
INTERES	1	5.0	5.8	-.81	-.34
INTERES	2	3.0	2.7	.31	.19
INTEGRAC	3				
INTERES	1	4.0	4.6	-.61	-.29
INTERES	2	2.0	2.1	-.14	-.09
EDUCACION	3				
INTEGRAC	1	9.0	10.1	-1.08	-.34
INTERES	1	5.0	4.7	.33	.15
INTERES	2				
INTEGRAC	2				
INTERES	1	6.0	5.8	.19	.08
INTERES	2	3.0	2.7	.31	.19
INTEGRAC	3				
INTERES	1	7.0	4.6	2.39	1.11
INTERES	2	2.0	2.1	-.14	-.09
EDUCACION	4				
INTEGRAC	1				
INTERES	1	8.0	10.1	-2.08	-.65
INTERES	2	9.0	4.7	4.33	2.00
INTEGRAC	2				
INTERES	1	6.0	5.8	.19	.08
INTERES	2	5.0	2.7	2.31	1.41
INTEGRAC	3				
INTERES	1	5.0	4.6	.39	.18
INTERES	2	3.0	2.1	.86	.59

Finalmente se contrastan los dos factores restantes, Integración e Interés, pero ambos hacen una aportación significativa, por tanto, no pueden ser excluidos del modelo. Así, el modelo final resultante sería

$$\ln(F_{ijk}) = \mu + \lambda_i^A + \lambda_j^B + \varepsilon_{ijk}$$

Es importante caer en la cuenta de que el análisis resumido en el Cuadro 3.1. indicaba claramente que estos dos serían muy probablemente los efectos que habrían de considerar para explicar las frecuencias observadas.

A continuación, podremos obtener las predicciones que se derivan del modelo usando los estimados de los parámetros de Integración (A) y de Interés (B). Desafortunadamente SPSS no los ofrece como parte de la salida de HILO-GLINEAR para modelos no saturados, por lo que no queda otro remedio que deducirlos a partir de la tabla de frecuencias esperadas que SPSS proporciona.

Cuadro 3.4. Frecuencias esperadas según el modelo (Integración, Interés)

		Interés	
Educación	Integración	Bajo	Alto
Primaria	Alta	10,1	4,7
	Media	5,8	2,7
	Baja	4,6	2,1
Secundaria	Alta	10,1	4,7
	Media	5,8	2,7
	Baja	4,6	2,1
Bachiller	Alta	10,1	4,7
	Media	5,8	2,7
	Baja	4,6	2,1
Universidad	Alta	10,1	4,7
	Media	5,8	2,7
	Baja	4,6	2,1

Esta tabla aparece en el Cuadro 3.3., junto con las observadas, los errores de predicción (residuos) y los residuos estandarizados, y en el Cuadro 3.4. (arriba) reorganizada para facilitar los cálculos. Debe tenerse en cuenta que la tabla es la que se espera cuando no hay más que efectos principales de los factores Integración e Interés, y puede obtenerse por el procedimiento iterativo mencionado más arriba. Esto es importante porque SPSS sí proporciona todos los parámetros cuando se le pide un modelo saturado. Pero la tabla esperada entonces es idéntica a la observada y la estimación de parámetros será diferente. El primer paso consiste en obtener los logaritmos naturales de las probabilidades esperadas de cada casilla ( $P_{ijk}$ ), que se obtienen como el logaritmo de la frecuencia esperada de la casilla menos el logaritmo del total de observaciones, esto es:

$$\ln(P_{ijk}) = \ln(F_{ijk}) - \ln(N) \quad (\text{Ec. 3.2})$$

Y, en nuestro ejemplo, redondeando al segundo decimal, tendríamos que la primera casilla sería:

$$\ln(P_{ijk}) = \ln(10,1) - \ln(120) = -2,47$$

Aplicando la ecuación 3.2 para el resto de las casillas, tendríamos:

-2,47	-3,24
-3,03	-3,79
-3,26	-4,05
-2,47	-3,24
-3,03	-3,79
-3,26	-4,05
-2,47	-3,24
-3,03	-3,79
-3,26	-4,05
-2,47	-3,24
-3,03	-3,79
-3,26	-4,05

Ahora, el resto de los pasos son sencillos, aunque prolijos. Recordemos, para empezar, que un efecto se define siempre como una diferencia de medias. De modo que para calcular los efectos principales de factor y de las interacciones necesitamos obtener la media total de las probabilidades esperadas, las medias de cada factor, las medias de las casillas de las interacciones de primer orden y así sucesivamente. Comencemos por la media total ( $m_G$ ), que obtendríamos sumando todas las probabilidades esperadas y dividiendo por el número total de casillas, que como hemos dicho es el producto del número de niveles de los factores,  $a \cdot b \cdot c$ , esto es:

$$m_G = S \ln(P_{ijk}) / (a \cdot b \cdot c) = [(-2,47) + \dots + (-4,05)] / (3 \cdot 2 \cdot 4) = -3,3076$$

Las medias de los niveles del factor A se calculan sumando a través de los niveles de B y de C, esto es:

$$m_{A(i)} = \sum_{jk} \ln(P_{ijk}) / (b \cdot c)$$

que en el caso del nivel 1 (alta integración) será:

$$m_{\text{INTEGRACIÓN (alta)}} = [-2,47 - 3,24 - 2,47 - 3,24 - 2,47 - 3,24] / (2 \cdot 4) = -2,857$$

en el de media:

$$m_{\text{INTEGRACIÓN}(\text{media})} = \frac{[-3,03 - 3,79 - 3,03 - 3,79 - 3,03 - 3,79 - 3,03 - 3,79]}{(2 \cdot 4)} = -3,412$$

y en el de baja:

$$m_{\text{INTEGRACIÓN}(\text{baja})} = \frac{[-3,26 - 4,05 - 3,26 - 4,05 - 3,26 - 4,05 - 3,26 - 4,05]}{(2 \cdot 4)} = -3,653$$

lo mismo haríamos para el resto de factores hasta obtener

$$\begin{aligned} m_{\text{INTERÉS}(\text{bajo})} &= -2,922 \quad \text{y} \quad m_{\text{INTERÉS}(\text{alto})} = -3,693 \quad \text{y} \\ m_{\text{EDUCACIÓN}(\text{primaria})} &= -3,308; \quad m_{\text{EDUCACIÓN}(\text{secundaria})} = -3,308; \\ m_{\text{EDUCACIÓN}(\text{bachiller})} &= -3,308; \quad \text{y} \quad m_{\text{EDUCACIÓN}(\text{universidad})} = -3,308 \end{aligned}$$

De una forma semejante se obtienen los logaritmos de las probabilidades para las interacciones de primer orden. Por ejemplo, las casillas de AxB serían:

$$m_{AxB(ij)} = \sum_k \ln(P_{ijk})/c$$

por tanto, en la interacción Integración x Interés tendremos:

$$m_{\text{INTEGRACIÓN}(\text{alta}) \times \text{INTERÉS}(\text{bajo})} = (-2,47 - 2,47 - 2,47 - 2,47) / 4 = -2,47$$

y así sucesivamente hasta obtener el resto de casillas:

		Interés	
Integración		Bajo	Alto
y	Alta	-2,47	-3,24
	Media	-3,03	-3,79
	Baja	-3,26	-4,05

El cálculo de los parámetros de los efectos se realiza a continuación según reglas muy simples. Concretamente, la constante es siempre la media global del modelo. Los parámetros de los efectos principales se computan como las medias de nivel menos la media total, es decir:

$$\begin{aligned} \lambda_i^A &= m_i - m_G; \quad \text{para los efectos de A} \\ \lambda_j^B &= m_j - m_G; \quad \text{para los efectos de B} \\ \text{y} \quad \lambda_k^C &= m_k - m_G; \quad \text{para los efectos de C} \end{aligned}$$

Por tanto, en los efectos que nos interesan, tendremos:

Efecto de A (Integración):

$$\begin{aligned} \text{Alta: } \lambda_{\text{alta}}^{\text{INTEGRACIÓN}} &= -2,857 - (-3,308) = 0,451; \\ \text{Media: } \lambda_{\text{media}}^{\text{INTEGRACIÓN}} &= -0,104; \\ \text{y Baja: } \lambda_{\text{baja}}^{\text{INTEGRACIÓN}} &= -0,346 \end{aligned}$$

Respecto de los efectos del factor Interés, tendremos:

$$\begin{aligned} \text{y Bajo: } \lambda_{\text{bajo}}^{\text{INTERÉS}} &= -2,922 - (-3,308) = 0,386; \\ \text{y Alto: } \lambda_{\text{alto}}^{\text{INTERÉS}} &= -0,386. \end{aligned}$$

Nótese que la suma de los parámetros de cualquier efecto es siempre cero. Queda para el lector la comprobación de que, salvo errores de redondeo, los parámetros de otros efectos, como los principales de C (Educación) o las interacciones de primer orden tienen valores despreciables.

El cómputo de parámetros es importante porque contribuye a clarificar cuál es la importancia de cada efecto respecto de la predicción de las frecuencias observadas. Sin embargo, es más adecuado evaluar su importancia teniendo en cuenta el error estándar asociado a cada parámetro. Esto es, pasándolos a puntuaciones z. El error estándar de un parámetro es una función de una suma de cocientes en los que se divide el cuadrado del recíproco del número de niveles del factor por la frecuencia correspondiente a cada nivel. Por ejemplo, el error estándar de los parámetros de A ( $SE_{\lambda(A)}$ ), es:

$$\begin{aligned} SE^2_{\lambda(A)} &= \sum (1/a)^2/F_{o(i)} \Rightarrow (1/3)^2/59 + (1/3)^2/34 + (1/3)^2/27 = 0,09. \\ \text{y} \quad SE_{\lambda(A)} &= 0,09^{1/2} = 0,0963 \end{aligned}$$

Por tanto, los parámetros de A estandarizados serán:

$$\begin{aligned} Z_{\lambda A(\text{alta})} &= 0,451/0,0963 = 4,67 \\ Z_{\lambda A(\text{media})} &= -0,104/0,0963 = -1,08 \\ Z_{\lambda A(\text{baja})} &= -0,346/0,0963 = -3,59 \end{aligned}$$

Si se desea decidir si la contribución es significativa sólo es necesario recordar que el z crítico para un nivel de significación de 0,05 (dos colas) es 1,96. Fácilmente se aprecia que alta y baja contribuyen significativamente, pero no media. El parámetro más importante es el correspondiente a alta, seguido por baja y, por último, media. Respecto del factor Interés, tendríamos que

$$\begin{aligned} Z_{IB(\text{bajo})} &= 0,386/0,1094 = 3,525 \\ \text{y} \quad Z_{IB(\text{bajo})} &= -0,386/0,1094 = -3,525 \end{aligned}$$

por lo que cabe concluir que la contribución es claramente significativa. Con

fines de completitud los parámetros asociados a la interacción AxB, por ejemplo, el de Integración alta e Interés bajo, tendrían el error estándar siguiente:

$$\begin{aligned} SE_{\lambda(AxB)}^2 &= \sum (1/(a \cdot b))^2 / F_{o(ij)}, \text{ luego} \\ SE_{\lambda(AxB)} &= (1/(3 \cdot 2))^2 / 42 + (1/(3 \cdot 2))^2 / 17 + \dots + (1/(3 \cdot 2))^2 / 8 = 0,0107, \\ \text{luego } SE_{\lambda(AxB)} &= 0,1034. \end{aligned}$$

Una vez que los parámetros del modelo están computados podemos comprobar que efectivamente obtenemos de forma correcta las frecuencias esperadas. Para ello, debemos recordar que

$$\ln(F_{e(ijk)}) = \mu + \lambda_i^A + \lambda_j^B$$

y que la frecuencia esperada es

$$F_{e(ijk)} = N \cdot \exp(\mu + \lambda_i^A + \lambda_j^B)$$

Así, la casilla primera de la tabla global corresponde a Educación Primaria, Integración alta, Interés bajo, por tanto:

$$\ln(F_{e(111)}) = -3,308 + 0,451 + 0,386 = -2,471$$

y la frecuencia esperada será

$$F_{e(111)} = 120 (e^{-2,471}) = 10,1$$

El lector puede comprobar que, salvo errores de redondeo, es posible reproducir la tabla de frecuencias observadas del Cuadro 3.3., proporcionada por SPSS.

La evaluación del ajuste obtenido requiere la consideración de los errores de predicción, los residuos. En el Cuadro 3.3. (véase página 109) aparecen estandarizados (columna Std Resid) y sin estandarizar. La estandarización se realiza dividiendo los residuos por su error estándar. Es habitual el uso de los residuos estandarizados porque permite tomar decisiones de forma relativamente automática, puesto que se trata de puntuaciones z. Es bien conocido que un valor de  $p < 0,05$  tiene una puntuación z de 1,96 (dos colas), por tanto, si adoptamos ese nivel de significación como criterio automáticamente podemos decidir que el ajuste no es adecuado en las casillas cuyo residuo sea igual o mayor que el criterio. Es fácil ver que en ningún caso el residuo estandarizado lo supera, por tanto, el ajuste puede considerarse aceptable.

Para terminar, SPSS ofrece otros dos programas más para el contraste de modelos: Log-Lineal General y Logit. El modelo Logit se emplea cuando una de las variables puede considerarse dependiente, en el sentido de que puede ser predicha a partir del resto de variables medidas (que serían consideradas inde-

pendientes). Sin embargo, es preferible utilizar regresión logística (véase Capítulo XI) en este tipo de análisis.

Por último, el programa LogLineal General permite el contraste de modelos concretos. Por ejemplo, a partir de nuestra evaluación inicial podríamos examinar directamente la idea de que los efectos principales de Integración e Interés son suficientes para explicar la tabla de frecuencias observada. El programa ofrece ahora las estimaciones de los parámetros, pero tiene el inconveniente de que no podemos evaluar el ajuste global del modelo, sino solamente respecto de la tabla que resultaría de la interacción de los dos factores que hemos introducido. Sin embargo, el analista puede realizar a mano la evaluación global teniendo en cuenta que las frecuencias predichas a partir de esos parámetros han de ser divididas por el número de niveles del factor (o factores) no introducidos en el modelo. Por ejemplo, empleemos la siguiente secuencia de comandos para realizar el análisis general:

Analizar → Loglineal → General → Factores: Integración, Interés → Personalizado → Construir Términos: Efectos principales → Factores y covariantes: Integración, Interés 4 → Continuar → Opciones: Frecuencias, Residuos, Estimaciones → Continuar → Aceptar (análisis)

Parte de los resultados, los que hacen referencia a las frecuencias y a las estimaciones de los parámetros, aparecen en el Cuadro 3.5, que se encuentra en la página siguiente. Nótese que la tabla de observaciones corresponde a la interacción Integración x Interés, y que los parámetros no coinciden con los que hemos estimado anteriormente. Sin embargo, sí su evaluación. Podemos reproducir la tabla global de frecuencias predichas, teniendo en cuenta que el factor no incluido tiene cuatro niveles, de la siguiente forma:

$$F_{e(ijk)} = \exp(\mu + \lambda_i^A + \lambda_j^B)/c,$$

por ejemplo, la casillas primera y segunda serán

$$\begin{aligned} F_{e(111)} &= \exp(2,1459 + 0,7817 + 0,7691)/4 = 10,1 \\ F_{e(112)} &= \exp(2,1459 + 0,7817 + 0)/4 = 4,68 \end{aligned}$$

y así sucesivamente con el resto de casillas.

Cuadro 3.5. Análisis Log-lineal general

a) Información de la tabla: frecuencias observadas y esperadas						
Factor	Value	Observed		Expected		
		Count	%	Count	%	
INTEGRAC	1	42.00	(35.00)	40.32	(33.60)	
	INTERES	17.00	(14.17)	18.68	(15.57)	
INTEGRAC	2	21.00	(17.50)	23.23	(19.36)	
	INTERES	13.00	(10.83)	10.77	(8.97)	
INTEGRAC	3	19.00	(15.83)	18.45	(15.38)	
	INTERES	8.00	(6.67)	8.55	(7.13)	
b) Parámetros estimados						
	Constant	Estimate				
	1	2.1459				
Note: Constant is not a parameter under multinomial assumption. Therefore, standard errors are not calculated.						
Parameter	Estimate	SE	Z-value	Asymptotic 95% CI		
				Lower	Upper	
				.33	1.24	
				-.27	.74	
				.38	1.15	
				.	.	
				.	.	

#### 4. LIMITACIONES Y SUPUESTOS DE LOS MODELOS LOG-LINEALES

Los modelos log-lineales son extraordinariamente interesantes como técnica de análisis no sometida a las restricciones que caracterizan a las técnicas paramétricas. Además pueden ser aplicados incluso cuando las variables son métricas, bien porque realicemos una transformación de escala, categorizándolas, bien porque empleemos el análisis log-lineal general, que permite el uso de variables continuas. En este último caso todavía seguimos contando con la ventaja de la no asunción de supuestos sobre sus características. Desde un punto de vista aplicado, los requisitos más importantes que los datos deben cumplir hacen referencia a la independencia de las observaciones, el número de variables en relación con el de sujetos y que en cada casilla haya un número suficiente de casos clasificados. Veamos brevemente cada una de ellas.

#### 4.1. INDEPENDENCIA DE LAS OBSERVACIONES

El requisito imprescindible es que cada sujeto pueda ser asignado solamente a una de las casillas de la tabla. O dicho de otro modo, las variables definen grupos de sujetos diferentes, y, por tanto, su combinación también.

#### 4.2. NÚMERO DE VARIABLES Y DE SUJETOS

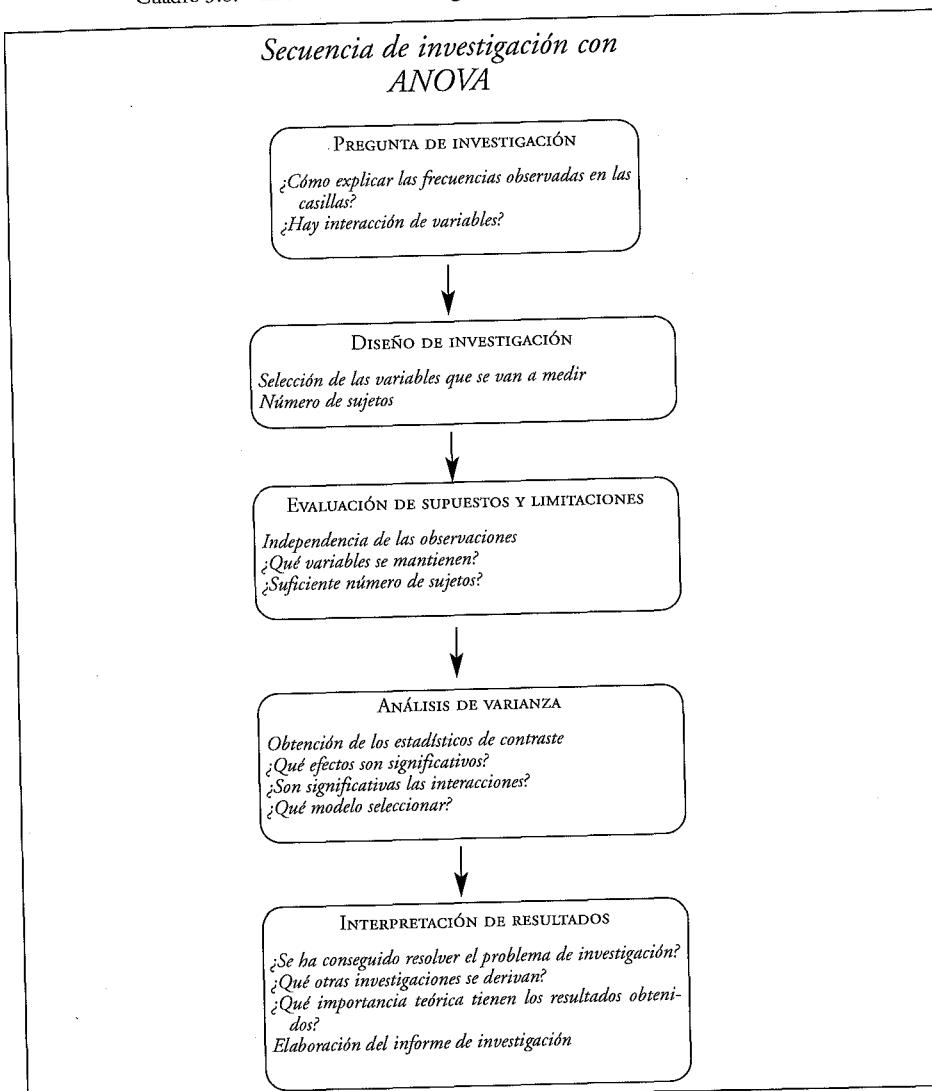
Si el número de sujetos es insuficiente en relación al número de variables es frecuente que un elevado número de casillas de la tabla carezcan de observaciones. En este caso, es bastante probable que no se encuentre una solución satisfactoria mediante el método de estimación máximo verosímil. No es éste, sin embargo, el único caso en que pueden plantearse problemas en la estimación. Cuando el número de casillas que tienen una frecuencia muy baja es elevado, las frecuencias marginales pueden ser demasiado bajas para que la estimación sea fiable. La solución es simple, es conveniente introducir un número elevado de sujetos. Aunque no hay una regla fija, puede ser suficiente con incluir al menos cinco casos por condición resultante de la mezcla de las variables medidas. Esto es, en nuestro ejemplo, con 24 condiciones ( $4 \times 3 \times 2$ ), el número mínimo recomendable es de  $24 \times 5 = 120$ , que es justamente el empleado. Además, es importante que el investigador se cerciore de que no ha incluido niveles de las variables que produzcan frecuencias anormalmente bajas. Por ejemplo, si en el factor educación hubiésemos introducido un quinto nivel de doctorado, hubiese sido bastante probable que las casillas correspondientes hubiesen tenido 1 o menos de frecuencia. Es preferible, en esos casos, eliminar el nivel, quizás agrupándolo con el nivel más semejante, educación universitaria. Otras alternativas son la supresión de la variable, cuando es ella la que de forma completa tiene asociadas las bajas frecuencias o reducir la potencia, admitiendo niveles de significación más restringidos.

#### 5. LA SECUENCIA DEL ANÁLISIS LOG-LINEAL

La secuencia de investigación orientada al análisis de datos de frecuencias aparece en el Cuadro 3.6. (véase página siguiente). Como ya hemos mencionado anteriormente, es importante caer en la cuenta de que, aunque estas investigaciones no exigen aparentemente una gran inversión de esfuerzo en su diseño, la selección de variables que puedan estar relacionadas entre sí, y la elección de una muestra suficientemente grande de sujetos son cruciales para que la investigación pueda tener alcance teórico suficiente. Tanto la determinación de los parámetros como la de la relación entre variables depende de modo especial de esas dos cuestiones. Finalmente, debe tenerse presente que estos análisis

pueden ser aplicados a variables que no son categóricas en origen, sino que han sido categorizadas *a posteriori*, sea por razones teóricas o prácticas. En cualquiera de los casos, es importante caer en la cuenta de que los modelos no permiten inferir relaciones causales entre variables.

Cuadro 3.6. La secuencia de investigación orientada al análisis de frecuencias



## 6. LECTURAS RECOMENDADAS

- ATO, M., *Investigación en ciencias del comportamiento. I: Fundamentos*, Barcelona, PPU, 1991.
- ATO, M. y LÓPEZ, J. J., *Análisis estadístico para datos categóricos*, Madrid, Síntesis, 1996.
- WICKENS, T. D., «Categorical data analysis», *Annual Review of Psychology*, 49, 1998, págs. 537-557.

## LECTURAS COMPLEMENTARIAS

- AGRESTI, A., *An Introducción to Categorical Data Analysis*, Nueva York, John Wiley, 1996.
- CHRISTENSEN, R., *Log-Linear Models and Logistic Regression*, Nueva York, Springer-Verlag, 1997.
- LE, C. T., *Applied Categorical Data Analysis*, Nueva York, Wiley, 1998.
- LLOYD, C. L., *Statistical analysis of categorical data*, Nueva York, Wiley, 1999.
- TABACHICK, B. G. y FIDELL, L. S., *Using Multivariate Statistics*, Nueva York, Harper Collins, 2000.
- WICKENS, T. D., *Multiway contingency tables analysis for the social sciences*, Hillsdale, N.J., LEA, 1989.

## 7. EJERCICIOS

- En la tabla siguiente aparecen las frecuencias obtenidas en una investigación en la que se pretendía investigar la relación entre el área geográfica (N, C y S), el género (V y M) y el nivel de ganancias mensual (A y B):

Ganancias	Género	N	C	S
A	V	52	120	30
	M	200	147	62
B	V	52	63	75
	M	156	103	85

Comprobar cuál es el modelo, con el mínimo número de variables que mejor ajusta los datos.

2. Si en el ejercicio anterior hubiese medido el partido al que los encuestados votan y creyese que el voto puede predecirse a partir de las tres variables (área geográfica, género y ganancias)
  - a) ¿Qué análisis categórico podría emplear?
  - b) ¿Qué otra alternativa analítica tendría?
3. Si el modelo más adecuado hubiese sido (área geográfica, ganancias, área x ganancias) ¿Se trataría de un modelo jerárquico?
4. ¿Cuáles de los siguientes residuos estandarizados cree que indican una falta de ajuste en el modelo? (1,07, -2,67, 0,32, -1,12)

#### RESPUESTAS A LOS EJERCICIOS

1. El mejor modelo es un modelo saturado, que incluya por tanto todos los efectos.
2. a) Un análisis logit, en el que se trata de explicar una variable, considerada dependiente, el voto, a partir de un conjunto de variables, consideradas independientes.  
b) Un análisis de regresión logística con predictores categóricos
3. Sería un modelo no jerárquico, puesto que la interacción no incluye los efectos principales implicados en el modelo.
4. Claramente -2,67 es una desviación significativa.

#### CAPÍTULO IV

#### Análisis factorial y de componentes principales

##### OBJETIVOS DE APRENDIZAJE

- 1) Conocer en qué contextos de investigación se emplea análisis factorial y en cuáles es preferible el análisis de componentes principales.
- 2) Distinguir estas técnicas de reducción de datos de otras, como el análisis de *cluster*, también orientadas a la agrupación de variables.
- 3) Distinguir el análisis factorial exploratorio (EFA) del análisis factorial confirmatorio (CFA).
- 4) Conocer la relación entre CFA y los modelos lineales estructurales (SEM).
- 5) Comprender cuáles son las operaciones fundamentales que es preciso realizar para llevar a cabo el análisis factorial: Extracción, rotación, comunidades, varianza y covarianza.
- 6) Conocer las diferentes alternativas entre las que se debe elegir en la extracción y en la rotación de factores.
- 7) Conocer las limitaciones y supuestos subyacentes al análisis.
- 8) Conocer los aspectos fundamentales del diseño de investigaciones de naturaleza factorial.

#### 1. CONCEPTOS FUNDAMENTALES DEL ANÁLISIS FACTORIAL Y DE COMPONENTES PRINCIPALES

El análisis factorial y el análisis de componentes principales son dos técnicas de reducción de datos que comparten un mismo objetivo general, reducir un conjunto de variables observadas a un conjunto menor de variables no observadas o subyacentes. La idea central es, pues, que en cualquier investigación es posible distinguir entre dos tipos de variables o atributos: superficiales

y latentes. Las variables superficiales u observadas, son las que el investigador mide a sus sujetos en el curso de su investigación (la amplitud de dígitos hacia adelante, el número de problemas resueltos en un tiempo limitado, etc.). Las variables superficiales se consideran una manifestación de variables no observables directamente (por ejemplo, la habilidad numérica, el razonamiento espacial, etc.), denominadas también atributos latentes, o factores. Esta forma de razonar es bastante común en ciencia, por ejemplo, los aspectos del comportamiento de los objetos celestes (rotación, etc.), las variables superficiales, se entienden bien cuando se recurre al concepto de gravedad, la variable latente o factor.

Lo importante es que los factores son una explicación de las variables superficiales en varios sentidos. Primero, cada factor puede influir en una o más variables superficiales. Por ejemplo, la habilidad numérica puede explicar las puntuaciones en test aritméticos, el rendimiento en matemáticas y física, al menos parcialmente. Segundo, distintos factores influyen en variables superficiales diferentes. Por ejemplo, la habilidad verbal puede explicar la fluidez verbal, pero no la velocidad de procesamiento del espacio. Tercero, las diferencias observadas entre individuos en una misma variable superficial, la variabilidad de la variable, dependen, al menos parcialmente, de las diferencias entre ellos en el factor. Por ejemplo, las diferencias entre distintos niños en la velocidad de resolución de problemas de sumas son el resultado, en parte, de que su habilidad numérica no es la misma. Cuarto, dos variables superficiales que sean influidas por el mismo factor deben tener una alta correlación. Por ejemplo, la correlación entre el número de problemas de sumar y de multiplicar debe ser alta, puesto que ambos dependen de la habilidad numérica. Sin embargo, la correlación entre variables superficiales que son influidas por diferentes factores debe ser baja. Por ejemplo, el tiempo que se invierte en recorrer un laberinto y la fluidez verbal, que dependen de habilidades espaciales y verbales, respectivamente, no deben tener una alta correlación entre sí. En pocas palabras, la existencia de factores explica, por lo menos en parte, la variabilidad (varianza) de las variables superficiales y las correlaciones entre ellas.

El análisis factorial y el de componentes principales son, pues, técnicas orientadas al descubrimiento de las variables subyacentes a un conjunto de variables superficiales. Las técnicas agrupan juntas a variables superficiales que correlacionan alto entre sí, puesto que se supone que dependen de una misma variable latente. La investigación dirigida a la identificación de factores procede a través de una serie de pasos bien ordenados. En primer lugar, es preciso identificar el dominio en el cual la investigación se inserta. Los dominios, conjuntos de fenómenos relacionados, son bastante diferentes, personalidad, inteligencia, memoria, etc. A continuación es necesario seleccionar un buen número de variables superficiales que van a ser medidas a un elevado número de sujetos. Estas variables son atributos superficiales. Cuando los atributos superficiales han sido medidos, hay varias cuestiones que son aparentes. Los individuos puntúan de forma bastante diferente unos de otros, y algunos atributos

correlacionan alto entre sí, mientras que entre otros la correlación es baja. Estos patrones de variación-correlación son aprovechados para identificar los factores subyacentes. Como hemos dicho, las variables que correlacionan entre sí son agrupadas en un factor. Además, las que no contribuyan a la definición de ningún factor son descartadas. A continuación, aunque no necesariamente, pueden volver a medirse las variables, y a realizar la extracción de factores hasta que se obtenga una solución estable. El conjunto de variables medidas forma una batería.

Es conveniente, en este punto, distinguir entre diferentes tipos de factores (véase Figura 4.1., en página siguiente). Los factores comunes, de los que hemos venido hablando hasta aquí, son los que influyen sobre más de una variable superficial. Los factores específicos son los que influyen solamente en una de las variables superficiales. Lo interesante es que si las pruebas de una batería se cambian, los factores específicos pueden cambiar también, desapareciendo los de la batería anterior y apareciendo nuevos asociados a las variables medidas en la batería actual. Sin embargo, debe tenerse presente que lo que en una batería es un factor específico, por ejemplo, si solamente tenemos una prueba de habilidad espacial, puede ser un factor común en otra, por ejemplo, si introducimos varias pruebas espaciales. De este modo, la puntuación observada para un sujeto depende de la influencia de factores comunes y de factores específicos. Los factores comunes son los responsables de la correlación entre variables, mientras que ambos son responsables de las diferencias entre los sujetos observadas en cada variable, las varianzas. Finalmente, es importante notar que las puntuaciones pueden depender de los errores de medida, que son considerados como factores adicionales, aunque es mejor verlos como indicadores de la existencia de variables que han influido en los sujetos durante la medición de las variables superficiales. Estos errores pueden deberse, por tanto, a características transitorias de los sujetos (su motivación, su fatiga, etc.), o de las situaciones de prueba (condiciones de ruido, etc.), que van a afectar a la fiabilidad con la cual se van a medir las variables superficiales, de ahí la necesidad de utilizar pruebas fiables o, de controlar esas posibles influencias extrañas. Es frecuente combinar los factores específicos con los factores de error de medida para obtener lo que se denominan factores únicos. Cada factor único influye en una sola variable superficial.

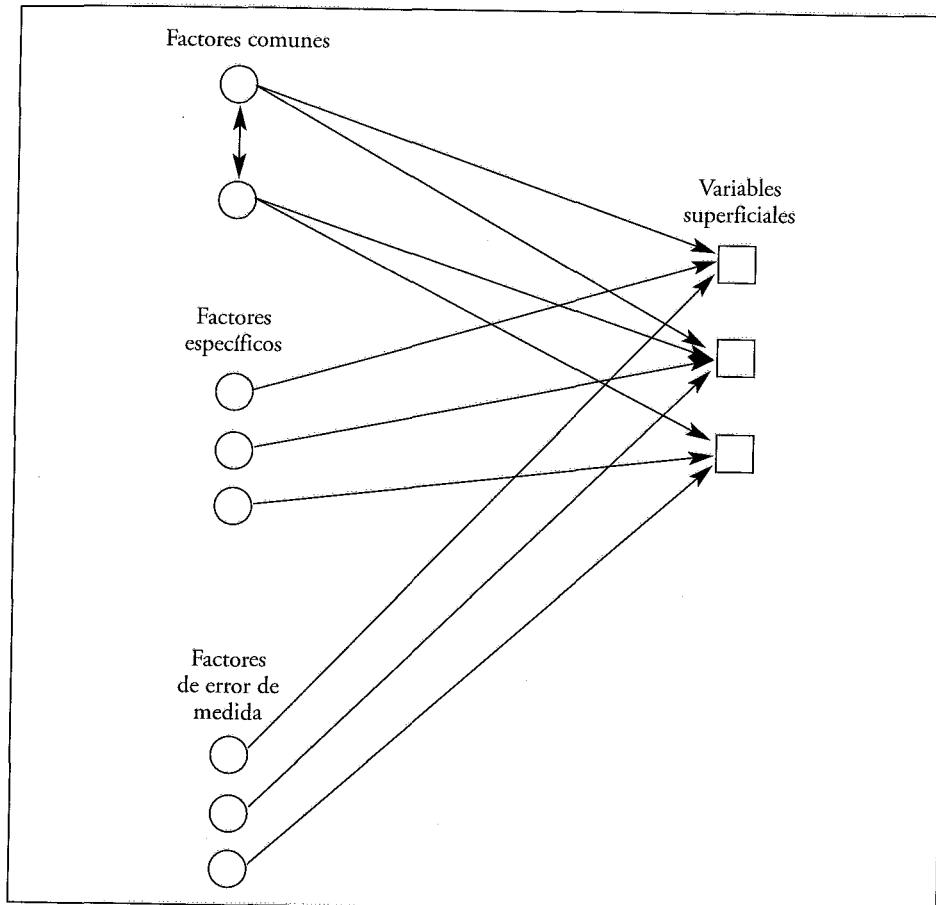
La puntuación de cualquier sujeto en una variable superficial es una suma de influencias de factores comunes, de un factor específico y de un factor de error de medida, o de factores comunes y de un factor único. Las Figuras 4.1. y 4.2. (véase páginas 124 y 125) presentan un esquema de estos conceptos y sus relaciones.

Por tanto, parte de la varianza de una variable superficial es debida a factores comunes, es la varianza común o comunalidad, y parte es debida al factor único, la varianza única. Podemos expresar esto de manera formal como:

$$X_{ij} = F_{1i}a_{i1} + F_{2i}a_{i2} + F_{3i}a_{i3} + \dots + F_{pi}a_{ip} + U_i$$

Donde  $X_{ij}$  es la puntuación del sujeto  $i$  en la variable  $j$ , los  $F$  son los coeficientes o cargas factoriales correspondientes a los  $p$  diferentes factores comunes (puede haber tantos como variables), las  $a$  son las puntuaciones del sujeto en cada factor y  $U$  es el factor único.

Figura 4.1. La teoría de los factores comunes

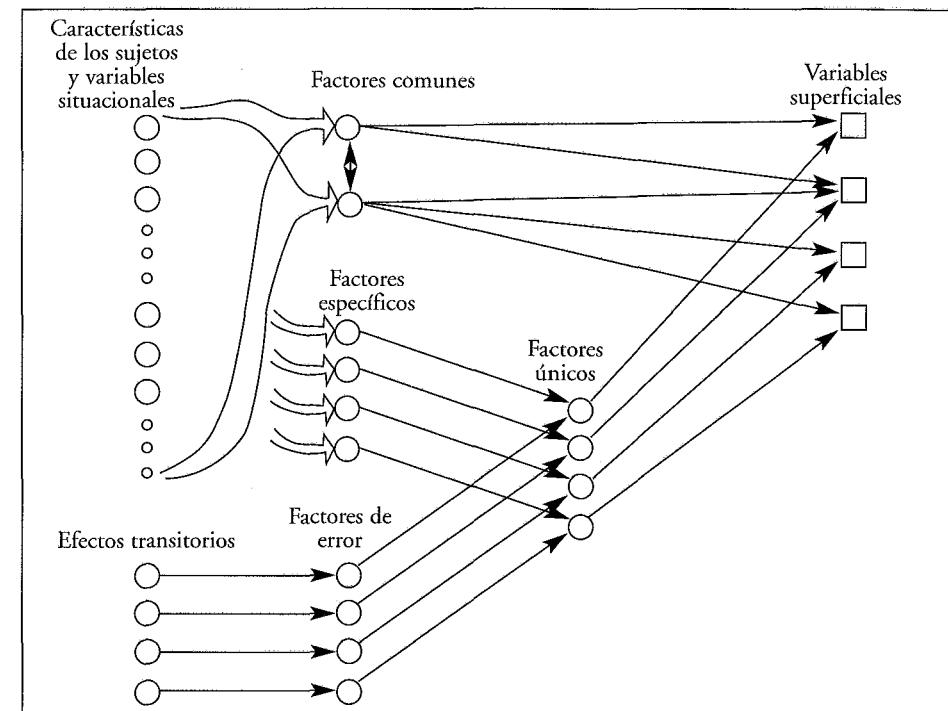


Por otro lado, la correlación entre variables sólo puede ser debida a los factores comunes, puesto que los factores únicos afectan solamente a una variable. En otras palabras, dos variables correlacionan si comparten, son influidas por, un factor común. Un último punto importante es que los factores comunes pueden estar a su vez relacionados entre sí. Por ejemplo, la habilidad numérica y la habilidad espacial pueden estar relacionadas entre sí.

Conceptualmente, el análisis factorial difiere del análisis de componentes en que el primero asume que los factores son entidades realmente existentes en los individuos, llenas de contenido teórico, la habilidad, la inteligencia, etc.,

mientras que el segundo trata sólo de agrupar las variables que correlacionan entre sí y separarlas de variables con las que no correlacionan. Los factores en el análisis de componentes principales no son interpretados teóricamente, sino que son sólo agrupaciones de variables. En otras palabras, en el análisis factorial se asume que hay una relación de causalidad entre factores y variables superficiales, pero los componentes principales son sólo agrupaciones empíricas de variables. Más adelante volveremos sobre este punto.

Figura 4.2. La teoría completa de factores comunes

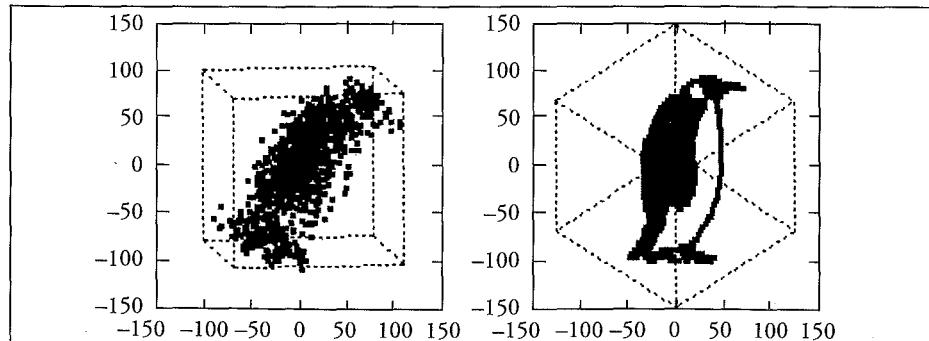


En el análisis factorial el concepto clave, por lo expuesto anteriormente, es el de matriz de correlaciones. Esta matriz se construye computando las correlaciones entre todos los pares de variables superficiales medidas. Es, por tanto, una matriz cuadrada de  $p$  filas por  $p$  columnas, asumiendo que  $p$  es el número de variables medidas. Suele llamarse matriz de correlaciones observadas. Tras el análisis, la combinación lineal de los factores permite obtener una nueva matriz de correlaciones, que denominaremos matriz de correlaciones predichas. La diferencia entre la observada y la predicha es la matriz de correlaciones residual. El objetivo del análisis es que la matriz residual contenga elementos lo más pequeños posibles.

Hay dos formas fundamentales de análisis factorial: exploratorio y confirmatorio. El análisis factorial exploratorio (EFA) se realiza cuando el investiga-

dor no tiene hipótesis *a priori* sobre cuáles pueden ser los factores que influyan en las variables medidas. Suele realizarse en las etapas iniciales de un proyecto de investigación. Permite identificar factores que pueden luego contrastarse en un análisis confirmatorio. El análisis factorial confirmatorio (CFA) se realiza, por tanto, cuando se tiene una idea clara de qué factores pueden extraerse. En general se lo considera como un caso particular de los modelos de ecuaciones estructurales (SEM). En cualquiera de los casos la secuencia del análisis está bien definida. En primer lugar se seleccionan las variables y se miden a un número elevado de sujetos. En segundo lugar se computa la matriz de correlaciones observada, y se procede a realizar la extracción de factores. Hay métodos diferentes para realizar la extracción, pero es importante dejar establecido desde el principio que solamente ofrecen resultados realmente diferentes cuando el número de sujetos es insuficiente. Con un elevado número de sujetos las soluciones (los factores extraídos) coinciden. A continuación se evalúan los factores para determinar qué número es necesario retener para explicar la matriz de correlaciones observada (que la matriz residual sea mínima). Prácticamente siempre es necesario rotar los factores con el objetivo de facilitar su interpretación. También en este caso se dispone de un gran número de alternativas, agrupadas en dos clases: rotación ortogonal, los factores son independientes, y oblicua, los factores correlacionan. La elección de una técnica concreta dentro de cada clase es, en general, irrelevante, puesto que prácticamente se va a alcanzar la misma solución si el número de sujetos es elevado. Es importante que el lector sea consciente de que la rotación no altera la estructura de la solución, sino solamente la cercanía de cada variable superficial a cada factor. La Figura 4.3. presenta una idea intuitiva del efecto de la rotación. A la izquierda, identificar una forma, una estructura en la nube compleja de puntos resulta difícil, casi imposible. Sin embargo, rotando en tres dimensiones los mismos puntos pueden percibirse muy claramente organizados. La rotación no ha cambiado la relación entre los puntos, sólo ha cambiado nuestro punto de vista sobre los mismos. Y cambiar el punto de vista no cabe duda de que ha facilitado enormemente la identificación de la estructura subyacente.

Figura 4.3. Efecto de la rotación. A la izquierda, antes de rotar; a la derecha, después



En este punto estamos en condiciones de predecir las puntuaciones de los sujetos en los factores, a partir de las variables medidas y de la importancia de cada variable en cada factor. Una vez identificados los factores es importante interpretarlos. Para ello se tienen en cuenta las variables superficiales asociadas a cada factor. Recuérdese que estos tienen un contenido teórico cuando realizamos análisis factorial, pero no cuando realizamos análisis de componentes principales. Por último, la solución factorial puede ser contrastada, validada, con respecto a un criterio externo. Por ejemplo, podemos emplear el rendimiento académico para validar una batería de pruebas de habilidades intelectuales.

En las páginas que siguen vamos a presentar los problemas de investigación que el análisis puede ayudar a resolver, la forma en que lo hace, y concluiremos con las limitaciones y supuestos que deben considerarse para que sea adecuado. Con fines expositivos utilizaremos el ejemplo que sigue. El lector debe tener en cuenta que, para facilitar la comprensión se emplean un bajo número de variables y un pequeño número de sujetos, pero que en estas circunstancias el análisis no es recomendable. Supongamos, hecha la salvedad, que queremos construir una batería de pruebas que nos permita medir memoria operativa. Como se sabe éste es un concepto que hace referencia al mantenimiento de la información activa durante un corto período de tiempo. Sin embargo, incluye, al menos, dos constructos teóricos fundamentales, el ejecutivo central y la memoria a corto plazo. El ejecutivo se supone encargado de una gran variedad de funciones de coordinación de procesamiento, de control de la respuesta, etc. (véase Capítulo primero), mientras que la memoria a corto plazo se concibe con un almacén temporal de información. Decidimos incluir una prueba de dígitos hacia delante (DA), otra de dígitos hacia atrás (DB), una prueba de capacidad de memoria (CM), y otra de bloques de Corsi (CB). Aplicamos las pruebas a diez sujetos y obtenemos los siguientes resultados:

Sujetos	DA	DB	CM	CB
1	7	5	6	6
2	6	5	6	5
3	8	6	7	7
4	7	6	5	5
5	8	4	5	7
6	8	5	5	7
7	9	5	5	8
8	7	4	5	6
9	6	5	5	6
10	7	5	6	5



## 2. LOS TIPOS DE PREGUNTAS QUE RESUELVE EL ANÁLISIS FACTORIAL

El objetivo fundamental del análisis factorial y del análisis de componentes principales consiste en reducir el gran conjunto de variables medidas a un conjunto menor de factores, normalmente con intenciones de obtener un modelo de las relaciones entre las variables superficiales y los factores. En algunos casos, cuando se ha hipotetizado una determinada relación, el análisis sirve para realizar el contraste del modelo. En este sentido, el análisis permite resolver una gran variedad de preguntas de investigación.

### 2.1. NÚMERO Y NATURALEZA DE LOS FACTORES

En primer lugar, el análisis nos permite resolver cuestiones relativas al número de factores que es necesario retener para reproducir la matriz de correlaciones observada. ¿Es necesario un solo factor, puesto que todo son medidas de memoria operativa? ¿Se necesitan dos factores, puesto que unas medidas pueden implicar más al ejecutivo central y otras más al almacén a corto plazo? Pero también en relación con su interpretación. ¿Qué significado tiene un único factor? ¿Cómo interpretaremos una solución de dos factores? Naturalmente las respuestas dependerán de los pesos (las cargas) de cada variable en cada factor y del patrón de correlaciones que observemos. Por ejemplo, si encontramos que DB y CM correlacionan entre sí más que con respecto a DA y CB, que también correlacionan alto entre sí, parece lógico pensar en una solución de dos factores. Si la solución es razonable, DB y CM tendrán una alta carga en un factor y DA y CB en otro. A partir de ahí podremos deducir que DB y CM comparten algo, probablemente sean medidas que implican al ejecutivo central, mientras que DA y CB implican más al almacén a corto plazo.

### 2.2. IMPORTANCIA DE LOS FACTORES

La variabilidad de los datos debe ser explicada por los factores identificados. ¿Cuánta varianza explican realmente? ¿Qué factores son los que explican mayor cantidad de varianza? Como veremos en el apartado 3, los métodos de extracción de factores tienden a garantizar que el primer factor extraído es el que mayor cantidad de varianza explica, después el segundo, y así sucesivamente. Evidentemente, cuanta mayor varianza explique un factor mayor es su importancia en la solución obtenida.

### 2.3. CONTRASTE DE MODELOS

En nuestro ejemplo hemos partido de la idea de la memoria operativa como un sistema dual. Si hubiésemos seleccionado las pruebas para tratar de medir los componentes, esperaríamos obtener dos factores, que estarían relacionados entre sí, puesto que la teoría asume que el ejecutivo controla al almacén. Optaríamos entonces por extraer dos factores y realizar una rotación oblicua. Nuestros problemas serían relativos a si la solución obtenida se ajusta a la que esperábamos, y si es capaz de explicar las correlaciones observadas. En este caso utilizaríamos análisis factorial confirmatorio, y muy probablemente una buena opción sería emplear modelos de ecuaciones estructurales (véase Capítulo XII).

### 2.4. ESTIMACIÓN DE LAS PUNTUACIONES FACTORIALES

Lo ideal sería poder medir los factores directamente, pero eso no es posible, tenemos que contentarnos con medir las variables superficiales. Pero sí cabe preguntarse qué puntuaciones en cada factor tendría cada sujeto. En este caso emplearemos la matriz de coeficientes para el cálculo. Estas puntuaciones pueden ser utilizadas para comparar sujetos entre sí, o para proseguir con más análisis.

## 3. EL ANÁLISIS FACTORIAL

El análisis factorial procede con una serie de pasos ordenados que presentaremos a continuación de la forma más intuitiva posible. En cada paso iremos presentando los métodos alternativos disponibles, sin embargo, seleccionaremos uno de ellos con fines expositivos, mientras que el resto serán descritos a un nivel intuitivo. Las etapas del análisis son: examen de la matriz de correlaciones, extracción de los factores, rotación factorial, importancia de los factores, interpretación de la solución obtenida y cálculo de los coeficientes para obtener las puntuaciones de cada sujeto en cada factor.

### 3.1. MATRIZ DE CORRELACIONES

El primer paso del análisis consiste en la obtención de una matriz que contiene las correlaciones entre todos los pares de variables superficiales medidas. Ésta es la matriz de correlaciones observada, que denominaremos R. Para obtenerla es preciso computar primero en cada variable las distancias de cada sujeto a la media de la variable. De esta forma obtenemos la matriz Y (véase Cuadro 4.1.) que cuando es premultiplicada por su traspuesta nos lleva a obtener una

matriz de sumas de cuadrados (en la diagonal positiva, la que va de la esquina superior izquierda a la esquina inferior derecha de la matriz) y de sumas de productos entre las variables, fuera de la diagonal. Dividiendo por el número de observaciones menos 1 obtenemos la matriz de varianzas-covarianzas (que puede ser empleada en el análisis en lugar de la de correlaciones, pero que tiene el inconveniente de depender de la escala de medida de las variables). La correlación la obtenemos dividiendo todos los elementos de esa matriz por las desviaciones típicas correspondientes a las variables cuya correlación queremos obtener.

Cuadro 4.1. Cálculo de la matriz de correlaciones

Y'				Y			
				-0,3	0	0,5	-0,2
				-1,3	0	0,5	-1,2
				0,7	1	1,5	0,8
-0,3	-1,3	0,7	-0,3	0,7	0,7	1,7	-0,3
0	0	1	1	-1	0	0	-1
0,5	0,5	1,5	-0,5	-0,5	-0,5	-0,5	0,5
-0,2	-1,2	0,8	-1,2	0,8	0,8	1,8	-0,2
							-0,2
							-1,2

Correlaciones Matriz R			
SCPC	Covarianzas	DA	DB
8,1 0 -0,5 7,4	0,9 0 -0,1 0,82	DA 1	0 -0,083 0,84
0 4 2 -1	0 0,44 0,22 -0,1	DB 0	1 0,4714 -0,2
-0,5 2 4,5 -1	-0,1 0,22 0,5 -0,1	CM -0,083	0,4714 1 -0,2
7,4 -1 -1 9,6	0,82 -0,1 -0,1 1,07	CB 0,8392	-0,161 -0,152 1

El Cuadro 4.1. resume este conjunto de pasos. Un examen simple de la matriz R indica que hay variables que correlacionan muy alto entre sí (DA con CB), otras que correlacionan de forma moderada (por encima de 0,30, CM y DB) y que el resto de correlaciones son muy bajas. Esto parece indicar que el número de factores que será posible extraer es de dos. Sin embargo, una inspección visual no es adecuada cuando el número de variables medidas, que es lo frecuente, es muy elevado. Es necesario, por tanto, tener índices que permitan saber si hay correlaciones altas en la matriz que permitan extraer factores. Hay varias pruebas utilizables en este sentido: el determinante de la matriz, el test de esfericidad de Bartlett, la prueba de Kaiser-Meyer-Olkin, y la correlación anti-imagen. Veámoslas.

El determinante de la matriz se emplea como índice del tamaño de las correlaciones. Cuando su valor es elevado, las correlaciones dentro de la matriz son bajas. Por el contrario, un determinante bajo indica que hay algunas correlaciones altas en la matriz. El determinante puede calcularse con la instrucción MDETERM(matriz) en Excel, pero los paquetes estadísticos, como SPSS, lo proporcionan entre sus opciones.

La prueba de esfericidad de Bartlett está diseñada para contrastar la hipótesis de que los elementos de fuera de la diagonal positiva (las correlaciones) de la matriz de correlaciones son cero (los diagonales son siempre 1). Una matriz que cumple esta propiedad se denomina matriz identidad. El test se computa a partir del determinante, de hecho es una transformación a chi-cuadrado del mismo:

$$\chi^2 = -\left[n - 1 - \frac{1}{6} (2p + 5)\right] \ln |R| \quad (\text{Ec. 4.1})$$

que se distribuye como una chi-cuadrado con  $(p^2-p)/2$  grados de libertad. En nuestro ejemplo, el determinante de R es 0,2079, por tanto tendremos que:

$$\chi^2 = -\left[10 - 1 - \frac{1}{6} (2 \cdot 4 + 5)\right] \ln (0.2079) = -(9 - \frac{13}{6}) \ln (0.2079) = 10,733$$

que está próximo a la significatividad con  $(4^2-4)/2 = 6$  grados de libertad ( $p=0,097$ ).

El índice de Bartlett tiene el importante problema de que depende del número de sujetos. Un índice alternativo es el KMO (Kaiser-Meyer-Olkin) que tiene en cuenta las correlaciones y las correlaciones parciales entre variables. Es un cociente entre la suma de los cuadrados de las correlaciones y la suma de las sumas de cuadrados de las correlaciones y de las correlaciones parciales (se obtienen para cada par de variables sustrayendo la influencia del resto). La idea subyacente es que si las correlaciones parciales son bajas las correlaciones observadas entre variables pueden emplearse para agrupar, puesto que la correlación desaparece cuando se elimina la influencia de los factores comunes. SPSS proporciona este índice junto con los dos anteriores. Es aconsejable obtener valores de más de 0,60 para que el análisis factorial pueda realizarse con garantías.

Un indicador complementario del anterior es la matriz anti-imagen, que contiene el complementario de las correlaciones parciales. Naturalmente la interpretación es simétrica de la anterior. Si en la matriz anti-imagen hay valores altos las correlaciones parciales son bajas y el análisis factorial puede realizarse con garantías. Por el contrario valores anti-imagen bajos son indicadores de que las correlaciones parciales son altas y de que probablemente no hay factores comunes que extraer.

El Cuadro 4.2. (véase página siguiente) contiene los índices de Bartlett, KMO y anti-imagen. Obsérvese que en nuestro ejemplo prácticamente todos indican que probablemente el análisis factorial no será muy exitoso, puesto que KMO se sitúa por debajo de 0,60 (que es un valor mediocre, Comrey y Lee, 1993), y encontramos una proporción elevada de anti-imágenes bajas.

Una vez que se ha decidido si los datos son analizables mediante análisis factorial puede procederse al siguiente paso, si la decisión es positiva, la extracción de factores.

Cuadro 4.2. Índices para evaluar la adecuación del análisis factorial a partir de las matrices de correlación

KMO y prueba de Bartlett				
Medida de adecuación muestral de Kaiser-Meyer-Olkin			,483	
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado gl Sig.	10,733 6 ,097		
Matrices anti-imagen				
	DA	DB	CM	CB
Correlación anti-imagen DA	,477 <sup>a</sup>	-,242	3,736E-02	-,849
DB	-,242	,429 <sup>a</sup>	-,453	,258
CM	3,763E-02	-,453	,549 <sup>a</sup>	1,442E-02
CB	-,849	,258	1,442E-02	,489 <sup>a</sup>

<sup>a</sup> Medida de adecuación muestral.

### 3.2. EXTRACCIÓN DE FACTORES

La extracción de factores es uno de los aspectos fundamentales del análisis, puesto que es precisamente donde se trata de reducir la información contenida en las variables superficiales a un pequeño número de variables latentes. El número de métodos existente para realizar la extracción es extraordinario (SPSS ofrece hasta siete alternativas), por lo que nos centraremos en el método de los componentes principales y luego explicaremos las diferencias que caracterizan al resto de métodos.

#### 3.2.1. Análisis factorial por componentes principales

El objetivo fundamental del análisis es reducir la matriz de correlaciones a un conjunto menor de variables latentes o componentes principales. Los componentes principales se caracterizan por no estar correlacionados entre sí, son ortogonales. Por tanto, si computamos la matriz de correlaciones de los componentes deberíamos tener unos en la diagonal principal y ceros fuera de ella. Una matriz con números distintos de cero en la diagonal y cero fuera de ella se denomina matriz diagonal. Esto es, la correlación de una variable latente consigo misma es la unidad, y su correlación con cualquier otra es cero. Considerando nuestro ejemplo, la idea es transformar la matriz original de correlaciones en una matriz diagonal, diagonalizar la matriz. Pero teniendo en la diagonal

la variabilidad explicada por cada componente principal. El Cuadro 4.3. contiene los cálculos necesarios para obtener la matriz de correlaciones con los datos de nuestro ejemplo, así como el objetivo a conseguir. La matriz de correlaciones se obtiene fácilmente computando las puntuaciones diferenciales para cada sujeto en cada variable y premultiplicando la matriz obtenida (sea X) por su traspuesta ( $X'$ , las filas se cambian por las columnas). La matriz  $X'X$  contiene sumas de cuadrados y productos cruzados entre pares de variables, de modo que si dividimos por los grados de libertad asociados (número de sujetos menos 1) tendremos la matriz de varianzas-covarianzas. Las varianzas en la diagonal positiva y las covarianzas entre pares de variables fuera de ella. Dividiendo las covarianzas por el producto de las desviaciones típicas de las variables correspondientes obtendremos las correlaciones. Llamaremos R a la matriz de correlaciones y D a la matriz diagonal. Para diagonalizar R es preciso obtener tanto sus vectores propios o autovectores (llamémoslos V) como sus valores propios o autovalores (sean  $\lambda_i$  cada uno de ellos). Recordemos que cada vector propio tiene asociado un valor propio y viceversa.

Un vector propio es una columna de pesos asociados cada uno de ellos a una variable medida. Por tanto, en nuestro ejemplo contendrá cuatro valores. Lo importante es que las cargas factoriales se obtienen multiplicando cada uno de los elementos del autovector por su autovalor. Las cargas factoriales indican la importancia que cada factor tiene en la determinación de la puntuación del individuo. La suma de los cuadrados de las cargas de un factor (dividido por la suma de los autovalores) indica la proporción de la varianza de las variables superficiales explicadas por ese factor. Lo interesante es que el autovalor asociado a ese factor es justamente la suma de cuadrados de sus cargas. Esto implica, lógicamente, que cuanto mayor sea el valor propio de un factor más varianza explica, mayor importancia tiene para explicar la variabilidad de las variables superficiales.

Asumiendo que D contiene los autovectores en la diagonal, tendremos que  $D = V'RV$ . Operando sobre esta expresión tendríamos que  $R = VDV'$  y, reorganizando los términos, tendríamos:

$$R = V \sqrt{D} \sqrt{D} V' \quad (\text{Ec. 4.2})$$

si llamamos F a las cargas factoriales, los pesos de cada factor, tendremos que

$$\begin{aligned} F &= V \sqrt{D} \\ F' &= \sqrt{D} V' \end{aligned} \quad (\text{Ec. 4.3})$$

y, por tanto,

$$R = FF' \quad (\text{Ec. 4.4})$$

que es la ecuación fundamental del análisis factorial e indica que la matriz de correlaciones puede obtenerse multiplicando una matriz de cargas por su traspuesta. Las cargas indican la importancia del factor en cada variable, y pueden obtenerse correlacionando factores y variables superficiales.

Como puede comprobarse, uno de los problemas más importantes del análisis consiste en derivar los vectores propios y los valores propios, puesto que su producto son las cargas factoriales. La derivación puede realizarse a partir de la matriz de covarianzas, pero entonces debemos tener presente la diferencia entre las escalas de las variables superficiales. Por ello, en las páginas que siguen presentaremos su derivación a partir de la matriz de correlaciones. Una de las formas de obtenerlos consiste en resolver dos sistemas de ecuaciones determinantes, uno para los autovalores y otro para los autovectores. Para los primeros tendríamos

$$|R - \lambda I| = 0, \quad (\text{Ec. 4.5})$$

donde  $R$  es la matriz de correlaciones  $I$  es la matriz identidad (que tiene unos en la diagonal positiva y ceros fuera de ella) y  $\lambda$  es el autovalor que buscamos. El segundo sistema sería

$$|(R - \lambda I) v| = 0, \quad (\text{Ec. 4.6})$$

donde  $v$  el autovector correspondiente a  $\lambda$ . El problema con esta aproximación es que produce un sistema de ecuaciones homogéneo, que debe resolverse adjudicando valores arbitrarios a algunos de los elementos de  $v$ , y que se vuelve relativamente inmanejable cuando el número de variables superficiales es elevado. Una solución más razonable consiste en emplear un algoritmo iterativo ideado por Hotelling (1933). Su ventaja, además, es que hace explícita una idea central del método de los componentes principales, a saber: que los componentes se van extrayendo de manera sucesiva, y que la extracción se realiza de manera que la mayor proporción de variabilidad es explicada por el primero, después por el segundo, y así sucesivamente. El algoritmo comienza suponiendo un conjunto de valores para el primer vector propio (que dará lugar al primer componente). Este vector es comparado con un criterio, de modo que si diverge del mismo, el vector es modificado para producir un segundo vector, que es comparado con el anterior. Si la divergencia entre ambos es demasiado grande (por ejemplo, difieren en el segundo decimal), se genera un tercer vector, que es comparado con el segundo, etc., hasta que se produce la convergencia (el vector  $n$  no difiere del vector  $n-1$  de modo significativo, por ejemplo, difieren a partir del cuarto decimal), o, dicho de otro modo, sucesivas iteraciones no producen ningún cambio en el vector. Este método iterativo permite el cálculo del valor propio a partir del vector propio.

Una vez que se ha obtenido el primer componente principal, se obtiene el segundo, pero ahora ya no se emplea la matriz de correlaciones original, sino la matriz residual, que se define como la diferencia entre la original,  $R$ , y la predicha a partir del primer componente extraído,  $R_1$ . Esto es:  $R_{\text{RES}1} = R - R_1$ . Este proceso se continúa hasta que se han extraído el resto de componentes. Es importante caer en la cuenta de que la varianza explicada es menor cuanto mayor es el orden del componente, de forma que los últimos en ser extraídos probablemente expliquen una porción irrelevante de variabilidad y deban ser descartados. El Cuadro 4.4. presenta una adaptación del algoritmo iterativo, aplicada para la obtención de los dos primeros componentes principales. El proceso puede implementarse muy fácilmente en un programa de ordenador. Obsérvese que se parte de la matriz de correlaciones y que para obtener el primer vector de prueba se emplean las sumas de las columnas de la matriz (realmente esto es innecesario, cualquier vector de prueba serviría). El lector puede intentarlo con el siguiente: (1 1 1 1). Este primer vector se obtiene dividiendo las sumas de las columnas por la raíz cuadrada de las sumas de cuadrados de las sumas. A continuación computamos el segundo vector multiplicando el primero por la matriz de correlaciones, y dividiendo los productos por la raíz de sus sumas de cuadrados, y así sucesivamente hasta que la diferencia entre vectores consecutivos es prácticamente nula. Por ejemplo, la diferencia entre  $V_{21}$  y  $V_{22}$  en el primer componente está por debajo de la milésima en cada elemento.

Cuadro 4.3. Obtención de la matriz de correlaciones y objetivo del análisis de componentes principales

Sujeto	Puntuación del sujeto en cada variable - media de la variable				X
	DA	DB	CM	CB	
1	7	5	6	6	-0,3 0 0,5 -0,2
2	6	5	6	5	-1,3 0 0,5 -1,2
3	8	6	7	7	0,7 1 1,5 0,8
4	7	6	5	5	-0,3 1 -0,5 -1,2
5	8	4	5	7	0,7 -1 -0,5 0,8
6	8	5	5	7	0,7 0 -0,5 0,8
7	9	5	5	8	1,7 0 -0,5 1,8
8	7	4	5	6	-0,3 -1 -0,5 -0,2
9	6	5	5	6	-1,3 0 -0,5 -0,2
10	7	5	6	5	-0,3 0 0,5 -1,2
Media	7,3	5	5,5	6,2	

SCPC	Covarianzas				Correlaciones (R)	Matriz Diagonal							
	DA	DB	CM	CB			C1	C2	C3	C4			
8,1	0	-0,5	7,4		DA	1	0	-0,1	0,84	$\lambda_1$	0	0	0
0	4	2	-1		DB	0	1	0,47	-0,2	0	$\lambda_2$	0	0
-0,5	2	4,5	-1		CM	-0,1	0,47	1	-0,2	0	0	$\lambda_3$	0
7,4	-1	-1	9,6		CB	0,84	-0,2	-0,2	1	0	0	0	$\lambda_4$
				ln	0,9	0	-0,1	0,82					
					0	0,44	0,22	-0,1					
					-0,1	0,22	0,5	-0,1					
					0,82	-0,1	-0,1	1,07					

Como puede comprobarse en el Cuadro 4.4. (véase página siguiente), el vector propio del primer componente sería  $V1 = (0,628, -0,269, -0,301, 0,664)$ ,

con un valor propio de  $\lambda_1 = 1,9276$ ; mientras que el segundo componente tendría  $V_2 = (0,3462, 0,6592, 0,6285, 0,2246)$  y un valor propio  $\lambda_2 = 1,3944$ . El vector del tercer componente sería  $V_3 = (-0,0603, -0,687, 0,7166, 0,1042)$ , con  $\lambda_3 = 0,5328$ , y el del cuarto componente sería  $V_4 = (-0,6941, 0,145, -0,0218, 0,7048)$ , con  $\lambda_4 = 0,1452$ .

Ahora, aplicando la Ec. 4.3. podemos obtener la matriz F de cargas factoriales (véase Cuadro 4.5., pág. 138), y demostrar que  $FF'$  producen R, salvo errores de redondeo, puesto que estamos empleando, todavía, todos los componentes principales.

Cuadro 4.4. Cómputo del primer componente principal

	R			
	DA	DB	CM	CB
DA	1	0	-0,08282	0,899177
DB	0	1	0,471405	-0,16137
CM	0,08282	0,471405	1	-0,15215
CB	0,899177	0,16137	-0,15215	1
Suma= U1	1,75636	1,31003	1,236442	1,525658
Vector: V1= U1/Ö(SCU1)	0,596925	0,445233	0,420223	0,518517
U2 = V1R <sup>-1</sup>	0,997251	0,559653	0,501783	0,883659
V2 = U2	0,651873	0,365828	0,328	0,577622
U3	1,109436	0,427236	0,358585	1,01572
V3	0,691556	0,266314	0,22352	0,633139
U4	1,204361	0,26951	0,19546	1,136494
V4	0,713034	0,159561	0,115721	0,672853
U5	1,268093	0,105531	0,029516	1,22786
V5	0,717031	0,059672	0,016689	0,694282
...	...	...	...	...
U19	1,214675	-0,51159	-0,57516	1,283584
V19	0,630156	-0,26541	-0,29838	0,665905
U20	1,21368	-0,51353	-0,577	1,282945
V20	0,629636	-0,26641	-0,29934	0,665569
U21	1,212957	-0,51492	-0,57833	1,282479
V21	0,629258	-0,26713	-0,30003	0,665326
U22	1,212432	-0,51593	-0,57929	1,282141
V22	0,628985	-0,26766	-0,30053	0,665149

$$\lambda = \sqrt{3,715643} = 1,927$$

Cuadro 4.4. (Continuación) Cómputo del segundo componente principal

Matriz de correlaciones predicha por el primer componente: R1 = v1v1 <sup>-1</sup> = (v1*v1, v1*v2, v1*v3, ..., v4*v1, ..., v4*v4)				
0,395622	-0,16835	-0,18903	0,418368	
<b>R1 = v1v1<sup>-1</sup> = (v1*v1, v1*v2, v1*v3, ..., v4*v1, ..., v4*v4)</b>				
-0,16835	0,071639	0,080437	-0,17803	
<b>R1 = v1v1<sup>-1</sup> = (v1*v1, v1*v2, v1*v3, ..., v4*v1, ..., v4*v4)</b>				
-0,18903	0,080437	0,090316	-0,19989	
0,418368	-0,17803	-0,19989	0,442423	
Matriz Residual: RRES = R - R1				
0,604378	0,168351	0,106209	0,420809	
0,168351	0,928361	0,390967	0,016656	
0,106209	0,390967	0,909684	0,047749	
0,420809	0,016656	0,047749	0,557577	
Suma RRES=U21 V21=U21/Ö(25,1)				
3,597776	0,526442	0,356622	3,472947	25,40967
0,713731	0,104436	0,070747	0,688967	
U22 = V21RRES V22				
0,746383	0,256247	0,213891	0,689614	1,144067
0,697808	0,239571	0,199971	0,644734	
U23 V23				
0,75462	0,428806	0,380474	0,666671	1,342536
0,651276	0,370082	0,328368	0,575372	
...				
U2,20 V2,20				
0,479701	0,920568	0,877882	0,310017	1,944345
0,34402	0,66019	0,629578	0,22233	
U2,21 V2,21				
0,479487	0,920658	0,877984	0,309791	1,944346
0,343867	0,660255	0,629651	0,222168	
$\lambda = \sqrt{1,944346} = 1,394$				

### 3.2.2. Número de componentes que se pueden extraer

En principio pueden obtenerse tantos componentes principales como variables superficiales tenemos. Sin embargo, conceptualmente resultaría curioso, cuando menos, que intentando reducir el número de variables, retuviésemos al final del análisis el mismo número total que teníamos inicialmente. Para ese viaje, casi no hacen falta alforjas. Parece más interesante retener solamente algunos de ellos. El problema reside en saber cuántos son necesarios y cuáles son los criterios de selección. Una aproximación razonable consiste en considerar el significado de los autovalores, que como hemos dicho anteriormente indican la proporción de varianza explicada. En nuestro ejemplo, el total de varianza es la suma de autovalores, por tanto tendremos:

$$\text{Varianza total} = 1,9276 + 1,3944 + 0,5328 + 0,1452 = 4$$

Por tanto, el primer factor explica un

$$1,9276/4 * 100 = 48,19\% \text{ de varianza,}$$

el segundo un 34,86 por 100, el tercero un 13,32 por 100 y el cuarto un 3,63 por 100, lo que indica que los dos primeros factores explican conjuntamente un 83,05 por 100 de la varianza total. Parece sensato retener esos dos factores y descartar los otros dos, de modo que reducimos de cuatro variables a dos componentes. En términos generales, suele admitirse que los factores cuyo autovalor sea inferior a la unidad pueden descartarse. Naturalmente esto conlleva una pérdida de la capacidad explicativa. Lo importante es que esa pérdida no sea significativa. En nuestro ejemplo, la matriz de correlaciones predicha empleando solamente dos factores aparece en el Cuadro 4.5. Compare con la matriz observada y obtenga la matriz residual, para obtener los errores de predicción. Obsérvese que, en general, el pronóstico es bueno, aunque falla un poco en la predicción de las relaciones entre las variables DB y CM.

Cuadro 4.5. Cálculo de cargas factoriales y de R

Autovectores, Matriz V:	C1	C2	C3	C4
	0,628268	0,346357	-0,06024	-0,694037
	-0,26902	0,659195	-0,68705	0,1450766
	-0,30183	0,628461	0,716562	-0,021795
	0,664684	0,224801	0,104254	0,7048339

Matriz $\sqrt{D}$ , con la raíz cuadrada de los autovalores en la diagonal
1,388382
0
0
0
0
1,180844
0
0
0
0,729956
0
0,3810079

Matriz $F = V \sqrt{D}$
0,872276
-0,37351
-0,41905
0,922835
0,408993
0,778407
0,742114
0,2654555
-0,04397
-0,50152
0,523058
0,0761006
0,264433
0,0552753
-0,008304
0,2685473

Matriz $F'$ , traspuesta de F
0,872276
0,408993
-0,04397
0,264433
-0,37351
0,778407
0,742114
0,2654555
-0,41905
0,523058
0,0761006
0,922835
0,0552753
-0,008304
0,2685473

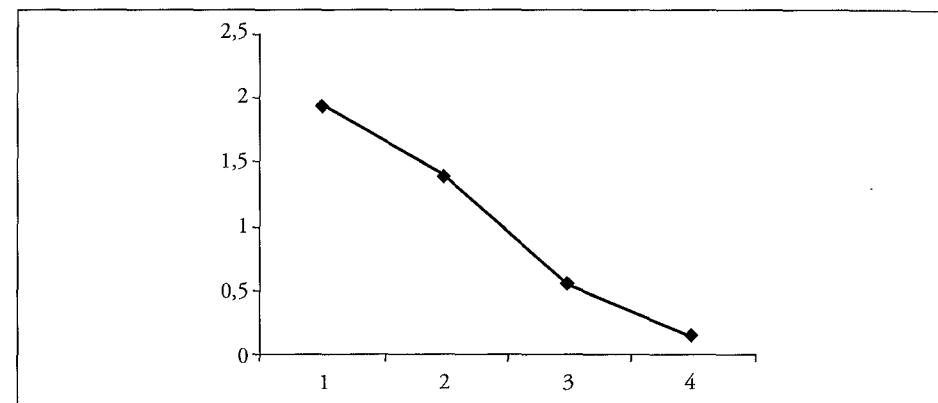
  

Matriz $R = FF'$
1
0
-0,08282
0,839177
0
1
0,471405
-0,161374
-0,08282
1
-0,152145
0,839177
-0,16137
-0,15215
1

Matriz R predicha por los dos primeros factores
0,9281
-0,0074
-0,0620
0,9135
-0,0074
0,7454
0,7342
-0,1381
-0,0620
0,7342
0,7263
-0,1897
0,9135
-0,1381
-0,1897
0,9221

Un método alternativo, conocido como *scree plot* consiste en representar gráficamente los autovalores (eje de ordenadas) para cada factor (eje de abscisas). La Figura 4.4. presenta los datos correspondientes del ejemplo. La inspección visual suele concentrarse en la búsqueda de un punto de inflexión en la gráfica, que habitualmente se produce con valores por debajo de uno.

Figura 4.4. La gráfica *scree plot*

### 3.2.3. Comunalidades

La comunalidad (que suele denominarse como  $h^2$ ) es también varianza explicada. Sin embargo, con este término se hace referencia a la varianza de una variable explicada por los componentes principales o los factores comunes. Esto implica que su cómputo se realiza a partir de la matriz que contiene las cargas factoriales, y que habrá una comunalidad para cada variable superficial. La comunalidad de la variable es la suma de las cargas factoriales asociadas a ella elevadas al cuadrado. En el caso de los componentes principales, cuando se retienen todos los componentes, la comunalidad es siempre uno, lo que hace que sea de poca utilidad. Cuando se emplean otros métodos de extracción de factores, se suele tener interés en computar la comunalidad inicial, que es el coeficiente de correlación múltiple entre cada variable y todas las demás, debido a que puede emplearse para evaluar la adecuación de los datos al análisis factorial. Pero es la comunalidad final la que suele tener mayor interés, puesto que indica qué variabilidad de cada variable superficial es explicada por el conjunto de factores retenidos y, por tanto, cuánta variabilidad es debida al factor único (recuérdese, un compuesto de factores específicos y factores de error). El Cuadro 4.6. presenta la comunalidades de las variables cuando se adopta una solución de cuatro componentes (arriba) o de dos (abajo). Obsérvese que en el primer caso el 100 por 100 de la varianza de cada variable es explicado por los componentes, mientras que en el segundo se explica más del 92 por 100 de la varianza de las variables DA y CM, pero menos del 75 por 100 de las variables DB y CB.

Cuadro 4.6. Comunalidad de cada variable reteniendo cuatro y dos componentes

	C1	C2	C3	C4	$h^2 =$ $F_1^2 + F_2^2 + \dots + F_p^2$	$h^2 (4C)$	$h^2 (2C)$	% de Var. Explicado
DA	0,873	0,409	-0,044	-0,264	=0,873 <sup>2</sup> +... =(-0,374) <sup>2</sup> +... =(-0,419) <sup>2</sup> +... =0,923 <sup>2</sup> +	1 1 1 1	0,928 0,745 0,726 0,922	92.8 74.5 72.6 92.2
DB	-0,374	0,778	-0,502	0,055				
CM	-0,419	0,742	0,523	-0,008				
CB	0,923	0,265	0,076	0,269				

### 3.2.4. Rotación de factores

Las soluciones factoriales no rotadas son difíciles de interpretar más allá de meras soluciones algebraicas. La razón fundamental de esto reside en que los métodos de extracción de factores están orientados a extraer la máxima cantidad de varianza en cada paso, de manera que a los últimos factores les queda muy poca varianza por explicar. Pero eso implica también que los primeros factores no son factores puros. Esto es, el primer factor que se extrae explica varianza propia y varianza que debería ser explicada por los demás factores. Por ejemplo, si extraemos, como en nuestro ejemplo, dos factores, el primero explica varianza propia y del segundo, mientras que al segundo sólo le queda por explicar parte de la varianza propia. Eso hace que las cargas factoriales del segundo tiendan a ser inferiores que las del primero. La rotación consigue obtener los pesos esperados de cada factor extraído, haciendo que la solución sea, entonces sí, interpretable.

La rotación consiste en una transformación de la matriz de cargas factoriales original de forma que los ejes factoriales (cada uno representa un factor) se aproximen lo máximo posible a las variables en las que tienen alta saturación (alto peso). La matriz resultante se denomina matriz factorial rotada. La cantidad de varianza explicada no cambia. El problema reside, lógicamente, en cómo conseguir esa transformación. Algebraicamente, la cuestión reside en computar la matriz rotada,  $V$ , multiplicando la matriz de cargas original,  $F$ , por una matriz de transformación,  $\Lambda$ . Esto es:

$$V = FA \quad (\text{Ec. 4.7})$$

de manera que  $R = VV'$ . Lo que reduce el problema a determinar la matriz  $\Lambda$ . Hay varios modos de obtenerla, pero quizás nos interesa más en este punto, atender al principio para seleccionar una matriz de transformación determinada. Este principio fue definido por Thurstone (1947) como de estructura simple, y puede entenderse como una concreción del principio general de parsimonia en la explicación científica, que establece que de todas las soluciones posibles debe preferirse la más sencilla. Según el principio de estructura simple, la rotación de factores debe cumplir varios criterios:

1. Cada factor debe contener cargas altas y cargas próximas a cero, esto es, debe tener importancia en la explicación de algunas de las variables superficiales, no de todas.
2. Cada variable debe ser explicada por un solo factor.
3. Factores con la misma distribución de cargas son redundantes, por tanto, factores diferentes deben tener distribución de cargas distinta.

El segundo problema importante es decidir qué rotación va a emplearse. Hay dos clases generales diferentes: ortogonal y oblicua. Veamos cada una de ellas por separado. Es importante tener presente que las rotaciones se realizan por parejas de factores, de modo que cuando se tienen más de dos, el proceso requiere varias iteraciones.

#### 3.2.4.1. «Rotación ortogonal»

Cuando los factores no se espera que estén correlacionados la rotación de elección es la ortogonal, lo que implica que los ejes (factores) son rotados el mismo número de grados en la misma dirección. En otras palabras, los ejes siguen, después de la rotación, siendo perpendiculares. Hay varios métodos disponibles: Varimax, Cuartimax, Ecuamax y Ortomax (SPSS incluye sólo los tres primeros).

##### 3.2.4.1.1. Rotación Varimax

El método Varimax es el más frecuentemente empleado. Fue diseñado por Kaiser (1958), y, como su nombre indica está orientado a la maximización de la varianza de los factores. Lógicamente, una forma de conseguir esto es que algunos factores tengan cargas altas y otros próximas a cero, lo que tiene como consecuencia que bastantes variables tengan saturaciones bajas en un factor.

Por trigonometría elemental sabemos que para rotar un eje tenemos que multiplicar por coseno y seno del ángulo de rotación, de modo que debemos primero identificar el ángulo que deseamos rotar los ejes. En nuestro ejemplo, la inspección de la Figura de la izquierda del Cuadro 4.7. indica que rotando en sentido antihorario alrededor de 24 grados lograremos el propósito de maximizar las varianzas de cada factor. Si realizamos un proceso iterativo podemos aproximarnos un poco más. Kaiser (1958) propuso un criterio de selección que es empleado por paquetes estadísticos como SPSS. Concretamente, el ángulo de rotación seleccionado  $j$ , será aquel que minimice la expresión siguiente:

$$V = \frac{\sum_{k=1}^c \sum_{j=1}^p g_{kj}^4 - \left( \sum_{k=1}^c \sum_{j=1}^p g_{kj}^2 \right)^2}{p^2} \quad (\text{Ec. 4.8})$$

donde  $g_{kj}$  son los elementos de la matriz de pesos transformada,  $k$  hace referencia a los componentes principales,  $j$  a las variables,  $c$  es el número de componentes y  $p$  el de variables. Llamaremos  $Q$  al primer sumando del numerador y  $C$  al segundo. Aplicando este criterio Varimax, seleccionaríamos un ángulo de rotación,  $\varphi$ , de  $-24,12$  grados. En ese caso, la matriz de transformación si incluyéramos todos los componentes tendría la forma siguiente:

$$\Lambda = \begin{pmatrix} \text{coseno}\varphi & \text{seno}\varphi \\ -\text{seno}\varphi & \text{coseno}\varphi \end{pmatrix}$$

Cuadro 4.7. Rotación Varimax de los dos componentes principales

#### 1) Cómputo

	F		$\Lambda$	
	C1	C2	0,9127	-0,409
DA	0,8723	0,409		
DB	-0,374	0,7784	0,4086	0,9127
CM	-0,419	0,7421		
CB	0,9228	0,2655		

Pesos factoriales tras la rotación  
 $V = FA$

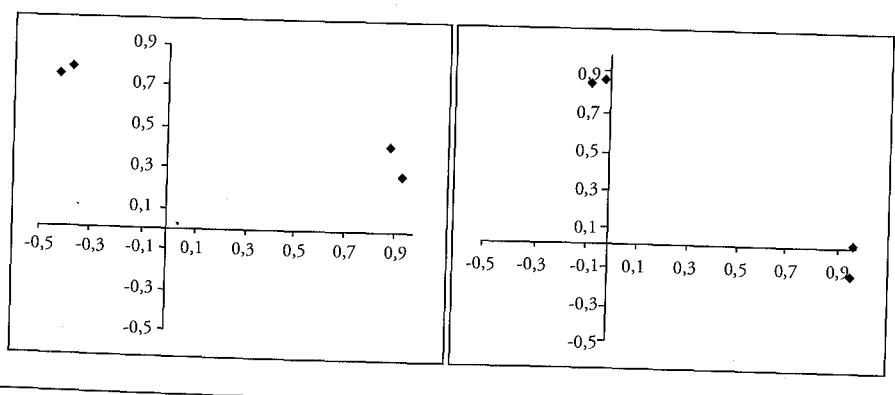
	DA	DB	CM	CB	$\Sigma_k$	Varimax
	0,9633	0,0168				
	-0,023	0,08631				
	-0,079	0,8486				
	0,9507	-0,135				

$$V = (Q-C)/4$$

$$S_k = \sum_{j=1}^p g_{kj}^4 = 1,6780 + 1,0737 \quad Q = s^2 \cdot 2 = 5,5034 \quad -0,019$$

$$r_k = \left( \sum_j g_{kj}^2 \right)^2 = 3,3803 + 2,2006 \quad C = 5,5809$$

#### 2) Pesos factoriales antes (izquierda) y después de la rotación (derecha).



y aplicando para los dos componentes seleccionados tendremos solamente la submatriz marcada en gris. Los resultados de esta rotación aparecen en el Cuadro 4.7., junto con las operaciones matriciales necesarias y el efecto gráfico de la rotación. Obsérvese cómo los pesos factoriales de DB y CM son prácticamente cero en el primer componente y los de DA y CB están también muy próximos a cero en el segundo componente.

Hasta aquí hemos presentado el caso más simple de dos factores, que sólo requiere una rotación. Frecuentemente, sin embargo, se retienen más de dos factores y el problema entonces es cómo se realiza la rotación en ese caso. La cuestión es simple, aunque tediosa si se tuviese que realizar manualmente. Como hemos mencionado más arriba, las rotaciones se realizan por pares de factores, de modo que si retuviésemos tres, tendríamos que realizar tres rotaciones (del primero con el segundo, del primero con el tercero y del segundo con el tercero). En caso de tener cuatro, las rotaciones se elevarían a seis, y así sucesivamente. En cada paso tendríamos una matriz de transformación,  $\Lambda_1$ ,  $\Lambda_2$ ,  $\Lambda_3$ , etcétera, que pueden multiplicarse para obtener una matriz de transformación única que podría emplearse para realizar todas las rotaciones simultáneamente.

#### 3.2.4.1.2. Rotación Cuartimax

La rotación cuartimax se caracteriza por que en lugar de minimizar  $V$ , se trata de encontrar el ángulo que haga máximo el valor de  $Q$ . El efecto más importante de esta estrategia es que reduce el número de factores que son necesarios para explicar una variable superficial. Sus resultados son, en general semejantes a los obtenidos con el procedimiento Varimax.

#### 3.2.4.1.3. Rotación Ecuamax

La rotación ecuamax es una mezcla de las dos anteriores, por tanto minimiza la función Varimax y maximiza  $Q$ , lo que se traduce en que simplifica la interpretación debido a que reduce el número de variables que saturan alto en un factor y reduce también el número de factores necesarios para explicar una variable.

#### 3.2.4.2. «Rotación oblicua»

Las rotaciones ortogonales se caracterizan porque las correlaciones entre los factores son nulas. En ocasiones, sin embargo, se espera que los factores estén correlacionados. Por ejemplo, puesto que las cuatro medidas que hemos empleado en el ejemplo se emplean para evaluar memoria a corto plazo, podríamos pensar que los dos factores obtenidos son no ortogonales, están correlaciona-

dos entre sí. La rotación que podríamos intentar en ese caso sería una que permitiese que los factores correlacionen, lo que gráficamente se traduciría en que los ejes del espacio euclídeo tengan entre sí ángulos no rectos. Para conseguir esto parece lógico pensar que los ejes no deben rotarse por parejas de modo simultáneo y en la misma dirección, sino más bien de forma individual, con ángulos y direcciones específicas para cada eje.

Una cuestión fundamental cuando se realizan rotaciones ortogonales es que la proyección de cada punto sobre los ejes (los factores) puede realizarse de una única forma, de manera que sólo hay un conjunto de pesos posible para cada variable. Sin embargo, cuando la rotación es oblicua, hay dos formas diferentes de realizar la proyección, una trazando paralelas a los ejes de coordenadas (los factores) y otra trazando perpendiculares a esos ejes. Lo importante es que ahora tenemos dos conjuntos de pesos para cada variable. Si la proyección es paralela a los ejes obtendremos los pesos que conforman lo que se llama matriz de configuración (o matriz patrón), mientras que la proyección perpendicular produce la matriz de pesos de estructura. El efecto inmediato de esta duplicidad es confundir la interpretación de los factores. La matriz de estructura contiene las correlaciones entre los factores y las variables superficiales. El problema con estos pesos es que sus cuadrados indican la varianza de una variable explicada por un factor y por su interacción con los otros factores. De ahí que sean poco recomendables para interpretar la estructura factorial, aunque puedan usarse para interpretar los factores. Por otro lado los cuadrados de los pesos de configuración indican la aportación de cada factor a la varianza de cada variable superficial. Pueden emplearse unos u otros para hacer la interpretación, pero la indecisión sobre cuál es más conveniente lleva a muchos investigadores a preferir rotaciones no oblicuas.

Técnicamente la rotación oblicua puede realizarse de una manera similar a las rotaciones ortogonales, cuando nuestro objetivo es obtener la matriz de estructura. Se trata de seleccionar el ángulo y la dirección en que hay que rotar cada factor (ahora por separado). Para ello definiremos de nuevo una matriz de transformación, que en el caso de dos factores tendría la forma siguiente

$$\begin{pmatrix} \text{Coseno}\varphi_1 & -\text{Seno}\varphi_2 \\ -\text{Seno}\varphi_1 & \text{Coseno}\varphi_2 \end{pmatrix}$$

Siendo  $\varphi_1$  y  $\varphi_2$ , los ángulos de rotación para el factor 1 y el factor 2, respectivamente.

Las diferencias entre los métodos, esto es, el ángulo de rotación seleccionado para cada eje, dependen del criterio que se decida utilizar. Hay dos criterios fundamentales, Oblimin directo y Promax. El criterio Oblimin directo (Jennrich y Sampson, 1966) trata de hacer mínima la suma de los productos de los cuadrados de los coeficientes de la matriz de configuración. El criterio Promax (Hendrickson y White, 1964) trata de obtener una transformación que haga

que la razón entre las cargas altas y las bajas sea lo más grande posible, esto implica, lógicamente que las cargas altas sean lo más altas posible y las bajas, lo más próximas a cero posible. En ambos métodos se parte de una rotación ortogonal previa.

### 3.2.4.3. «¿Qué tipo de rotación aplicar?»

Aunque los paquetes estadísticos proporcionan los tipos de rotaciones expuestas más arriba, muchos investigadores prefieren aplicar rotaciones ortogonales, particularmente la rotación Varimax. Las razones de esta elección son variadas. En primer lugar, las rotaciones oblicuas, aunque más versátiles, son más difíciles de interpretar que las ortogonales. Esto se debe, al menos en parte, a que las rotaciones oblicuas proporcionan dos matrices de pesos que tienen uso diferente. En efecto, mientras que la matriz de estructura nos indica la correlación entre factores y variables, la matriz de configuración es empleada para calcular las puntuaciones factoriales de los sujetos. Sin embargo, su semejanza con la anterior hace que pueda confundirse en su significado, e interpretarse como si indicase correlaciones de factores con variables. Esta confusión no se produce cuando se emplean rotaciones ortogonales porque ambas matrices son la misma. Además, es importante caer en la cuenta que en las rotaciones oblicuas, la communalidad no suele obtenerse sumando las cargas de una variable en los diferentes factores, y que, por tanto, estas rotaciones no clarifican bien la proporción de varianza de la variable que es explicada por los factores. Por consiguiente, a no ser que se esperen correlaciones entre los factores, lo cual es dudoso en estudios exploratorios que utilizan EFA como técnica de análisis, lo más recomendable es emplear una rotación ortogonal, especialmente Varimax. Sin embargo, el investigador puede probar primero una solución oblicua y decidir posteriormente en función de las correlaciones observadas entre los factores, de manera que si son menores de 0,35 puede considerar los factores ortogonales e intentar la solución ortogonal. En caso de que sea necesario aplicar una rotación oblicua, quizás la mejor elección sea utilizar la Oblimin directa.

### 3.2.5. *El análisis factorial mediante SPSS*

El conjunto de cómputos que hemos presentado más arriba puede realizarse muy fácilmente mediante diferentes paquetes estadísticos. Como hemos señalado ya en otros capítulos, emplearemos SPSS con fines ilustrativos. Antes de utilizar el paquete es preciso tener presente qué tipo de extracción de factores se pretende realizar y qué rotación pueda ser necesaria para clarificar la interpretación. A partir de ahí, y teniendo presente que hemos comprobado que la matriz de correlaciones cumple las condiciones necesarias para el análisis, las cosas son bien sencillas. La secuencia de comandos necesaria para reali-

zarlo, asumiendo que se va a aplicar una extracción de componentes principales y una rotación ortogonal Varimax para los datos del ejemplo sería:

Analizar → Reducción de datos → Análisis factorial → Variables: DA, DB, CM, CB → Extracción → Método: Componentes principales (es la opción por defecto) → Analizar: Matriz de correlaciones → Extraer: Autovalores mayores que 1 (es la opción más recomendable, como hemos señalado más arriba) → Mostrar: Solución factorial sin rotar, Gráfico de sedimentación → Continuar → Rotación → Método: Varimax → Mostrar: Solución factorial rotada; Gráficos de saturaciones → Continuar → Aceptar (análisis).

SPSS nos presenta en primer lugar la solución factorial inicial, pero debemos tener presente que en el caso de la extracción de componentes principales ésta va a coincidir exactamente con la solución final. A continuación nos presenta los valores propios (no los vectores propios) y la varianza explicada, principalmente por los factores que han sido seleccionados (los que cumplen el criterio de tener un autovalor mayor que 1). Nótese que la salida incluye la varianza explicada antes y después de la rotación.

Varianza total explicada

Componente	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	1,928	48,190	48,190	1,928	48,190	48,190	1,839	45,964	45,964
2	1,934	34,860	83,050	1,934	34,860	83,050	1,483	37,086	83,050
3	,533	13,321	96,371						
4	,145	3,629	100,000						

A continuación, aparece el gráfico de sedimentación (*scree plot*), que hemos presentado anteriormente (Figura 4.4., pág. 139) y que ayuda a tomar la decisión sobre el número de factores a retener, algo que ya hemos decidido al fijar el criterio de autovalor en la unidad. Por tanto, retendríamos los dos primeros componentes principales. Los pesos de cada variable en cada componente antes de la rotación aparecen bajo el epígrafe de matriz de componentes.

Matriz de componentes<sup>a</sup>

	Componente	
	1	2
DA	,872	,409
DB	-,374	,778
CM	-,419	,742
CB	,923	,265

Método de extracción: Análisis de componentes principales

<sup>a</sup> 2 componentes extraídos

Finalmente nos presenta la matriz de componentes rotados, que contiene los pesos después de la rotación Varimax, y la matriz de transformación empleada para realizar la rotación.

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

	Componente	
	1	2
DA	,963	1,68E-02
DB	-2,28E-02	,863
CM	-7,92E-02	,849
CB	,951	-,135

Método de extracción: Análisis de componentes principales

Método de rotación: Normalización Varimax con Kaiser.

<sup>a</sup> La rotación ha convergido en 3 iteraciones.

Matriz de transformación de los componentes

Componente	1	2
1	,913	-,409
2	,409	,913

Método de extracción: Análisis de componentes principales

Método de rotación: Normalización Varimax con Kaiser.

La representación gráfica de los componentes rotados coincide con la que hemos presentado más arriba.

Solamente con fines de exemplificación podemos comparar la solución obtenida con la que obtendríamos si decidiésemos emplear una rotación Oblimin directa. En este caso, la secuencia relativa a la rotación sería:

... → Rotación → Método: Oblimin directo → Delta: 0 → ...

La elección de un delta de cero implica que la rotación será lo más oblicua posible. La salida de SPSS nos presenta la matriz de estructura y la matriz de

configuración, junto con la matriz de correlaciones entre los factores, cuya inspección nos indicará el solapamiento entre componentes. Nótese que en nuestro caso, a pesar de utilizar un delta de cero, la correlación es muy baja, lo que indicaría que los factores pueden considerarse ortogonales.

Matriz de configuración<sup>a</sup>

	Componente	
	1	2
DA	,969	7,368E-02
DB	2,792E-02	,866
CM	-2,96E-02	,848
CB	,948	-7,95E-02

Método de extracción: Análisis de componentes principales

Método de rotación: Normalización Oblimin con Kaiser

<sup>a</sup> La rotación ha convergido en 3 iteraciones.

Matriz de estructura

	Componente	
	1	2
DA	,961	-3,96E-02
DB	-7,33E-02	,863
CM	-1,29	,852
CB	,957	,190

Método de extracción: Análisis de componentes principales

Método de rotación: Normalización Oblimin con Kaiser

Matriz de correlaciones de componentes

Componentes	1	2
1	1,000	-,117
2	-,117	1,000

Método de extracción: Análisis de componentes principales

Método de rotación: Normalización Oblimin con Kaiser

Los resultados del análisis parecen indicar, por tanto, que debemos retener dos componentes, y que las variables DA y CB tienen una alta carga en el primero, mientras que las variables CM y DB la tienen alta en el segundo. Aunque el análisis de componentes principales no está orientado en el sentido de suponer relaciones causa efecto entre componentes y variables superficiales, podríamos estar tentados de concluir que desde un punto de vista teórico este patrón de resultados implicaría que DA y CB dependen de un mecanismo diferente que DB y CM. Examinando las tareas con detalle quizás concluyéramos que las primeras implican más un almacén a corto plazo mientras que las dos últimas implican más a la función ejecutiva de la memoria operativa.

### 3.3. PUNTUACIONES FACTORIALES

A veces interesa conocer las puntuaciones que obtendrían los sujetos si pudiesen ser medidos directamente en cada factor. Sólo podemos conocerlas, sin embargo, de modo indirecto, computando las puntuaciones factoriales mediante una ecuación de predicción del tipo siguiente:

$$F_{ij} = B_{i1}Z_1 + B_{i2}Z_2 + \dots + B_{ip}Z_p$$

Donde  $F_{ij}$  es la puntuación factorial del sujeto  $i$  en el factor  $j$ , los  $B$  dependen de los pesos factoriales de cada variable en cada factor y los  $Z$  son las puntuaciones tipificadas del sujeto en cada variable superficial. Lógicamente el resultado,  $F_{ij}$ , es una puntuación típica, que indica en qué grado posee el sujeto el factor, y que puede ser empleada con diferentes fines analíticos, como extraer factores de segundo orden o agrupar los sujetos en categorías mediante procedimientos algorítmicos como el análisis de cluster. Para hacer los cálculos es preciso calcular primero los valores de  $B$ , que se obtienen a partir de las cargas factoriales y de la inversa de la matriz de correlaciones. SPSS los ofrece cuando en las instrucciones se añade lo siguiente:

... → Puntuaciones → Guardar como variables → ...

y cuyo resultado sería el siguiente:

Matriz de coeficientes para el cálculo de las puntuaciones en los componentes

	Componente	
	1	2
DA	,533	,083
DB	,051	,589
CM	,019	,575
CB	,515	-,022

Método de extracción: Análisis de componentes principales

Método de rotación: Normalización Varimax con Kaiser.

Puntuaciones de componentes.

El análisis de las puntuaciones factoriales de los sujetos, que aparecen en la tabla que se presenta más abajo, nos proporciona una información que depende de la interpretación de factores que hayamos realizado. Por ejemplo, podríamos tentativamente asociar el primer componente a almacén a corto plazo y en ese caso tendríamos una medida de las diferencias entre los distintos sujetos en capacidad del almacén. Es interesante caer en la cuenta de que estas medidas son ortogonales, y que por tanto, pueden emplearse como variables predictoras en una ecuación de regresión múltiple.

Sujetos	C1	C2
1	-0,27	0,38
2	-1,33	0,27
3	0,83	2,18
4	-0,72	0,45
5	0,74	-1,22
6	0,79	-0,34
7	1,85	-0,23
8	-0,31	-1,32
9	-0,83	-0,54
10	-0,77	0,38

### 3.4. OTROS MÉTODOS DE EXTRACCIÓN DE FACTORES

El método de extracción de componentes principales es uno de los métodos posibles para identificar las variables latentes. La principal habilidad de este método es también quizás su principal desventaja, a saber: es capaz de reproducir la matriz de correlaciones, pero puesto que en la matriz hay varianza error, eso indica que es también capaz de explicar el error, lo cual no deja de ser un poco extraño. Sin embargo, cuando no se retienen todos los componentes, el error puede calcularse fácilmente a partir de las communalidades,  $h^2$ , puesto que representan la varianza de cada variable superficial explicada por el conjunto de factores retenidos. El error será, pues,  $e = 1 - h^2$ . No retener todos los componentes reduce el problema, pero no lo elimina de modo definitivo, puesto que en la matriz se siguen introduciendo unos en la diagonal positiva, lo que implica asumir que las pruebas empleadas carecen de error. Esto hace que muchos autores (por ejemplo, Cattell, 1978) prefieran otros métodos que se concentran más en la extracción de factores comunes y separan su contribución de los factores únicos (recordemos que estos incluyen factores específicos y factores de error). Las ventajas de esta aproximación están claras. Primero, los factores comunes son entidades hipotéticas que se supone tienen una influencia causal sobre las variables superficiales, mientras que los componentes son entidades reales resultado de la mera agrupación matemática de variables superficiales. En segundo lugar, como es de esperar, dados los errores de medida y la influencia de factores específicos de cada prueba, las communalidades no son la unidad. El número de métodos de extracción disponibles es elevado, por lo que expondremos con cierto detalle sólo el métodos de factores principales y de forma intuitiva el resto.

#### 3.4.1. Método de los factores principales

El método de los factores principales es un método recursivo cuyas características de cómputo son muy semejantes a las de los componentes principales. La diferencia más importante es que en lugar de utilizar unos en la diag-

onal positiva (recordemos que en los componentes principales esto se debía a que estábamos considerando que la correlación de una variable consigo misma debe ser perfecta), utiliza las communalidades. Lógicamente, para emplearlas primero deben ser estimadas. El problema se reduce, por tanto, al cómputo de las communalidades y al procedimiento iterativo para obtener los factores.

Las communalidades se obtienen inicialmente mediante un procedimiento de regresión múltiple (véase Capítulo VIII), en el que se intenta predecir cada una de las variables a partir de las demás. El cuadrado del coeficiente de correlación es la communalidad inicial de la variable, que se introduce en el lugar correspondiente de la diagonal positiva de la matriz de correlaciones. SPSS proporciona al principio de su salida las communalidades iniciales y las obtenidas después de la extracción. En nuestro caso, la secuencia de comandos de extracción sería:

... → Extracción → Método: Ejes Principales → ...

cuyo resultado inicial aparece a continuación.

Comunalidades		
	Inicial	Extracción
DA	,723	,876
DB	,275	,607
CM	,229	,367
CB	,730	,834

El resto del procedimiento de extracción se reduce a lo siguiente: se obtienen los factores mediante el mismo procedimiento que hemos examinado para los componentes principales. Se computan las nuevas communalidades y se comparan con las anteriores. Si hay diferencias por encima de un criterio, las communalidades actuales reemplazan a las anteriores y se vuelve a repetir la extracción, hasta que se alcance el criterio de convergencia, esto es hasta que la diferencia entre communalidades actual y anterior sea lo suficientemente pequeña. Las cargas factoriales que se consideran correctas son las correspondientes a la última iteración. Una alternativa a este procedimiento, que SPSS emplea para calcular los autovalores de la matriz de correlaciones, consiste en realizar la primera extracción con unos en la diagonal positiva, y continuar después el proceso iterativo mencionado.

Aplicando este procedimiento mediante SPSS obtendríamos:

Varianza total explicada

Factor	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	1,928	48,190	48,190	1,744	43,607	43,607	1,690	42,255	42,255
2	1,394	34,860	83,050	,940	23,498	67,105	,994	24,850	67,105
3	,533	13,321	96,371						
4	,145	3,629	100,000						

Método de extracción: Factorización de Ejes principales.

Esto es, la varianza explicada por los dos factores extraídos es considerablemente menor que en el método de los componentes principales. Nótese cómo la varianza explicada por el primer factor es alrededor de 4,5 menor y alrededor de 11 puntos en el caso del factor 2, que con componentes. El resto de los resultados, matriz factorial, matriz de factores rotados, etc., tienen la misma interpretación que hemos visto más arriba.

#### 3.4.2. *Métodos de mínimos cuadrados*

Por lo que llevamos visto hasta aquí, parece que la diagonal positiva de la matriz es un problema, por tanto, no es de extrañar que algunos autores (Comrey, 1962, Harman, 1976) propusieran procedimientos que evitan su uso, y que son conocidos como métodos de residuo mínimo o de mínimos cuadrados. El objetivo puede establecerse de una forma muy simple, se trata de extraer factores de forma que la suma de los cuadrados de las diferencias entre la matriz predicha (tras la extracción) y la matriz original (sin contar en ningún caso la diagonal) sea lo menor posible. El procedimiento puede aplicarse de dos formas diferentes: no ponderada y generalizada. La diferencia entre ambos reside en que las correlaciones son ponderadas en el caso generalizado. Los mayores pesos corresponden a las variables que tienen mayor varianza compartida con las demás.

#### 3.4.3. *Método de máxima verosimilitud*

Una de las diferencias más importantes del método de máxima verosimilitud, desarrollado por Joreskorg es que comienza con una matriz de covarianzas (llamémosla  $\Sigma$ ) poblacional, en lugar de muestral (llamémosla  $S$ ) como los demás métodos, y asume que las puntuaciones de los sujetos provienen de distribuciones normales multivariadas. El objetivo del análisis es maximizar la función

$$L = \frac{n}{2} \ln |\Sigma| + \text{traza}(S\Sigma^{-1}) \quad (\text{Ec. 4.9})$$

donde la traza es la suma de los elementos diagonales del producto de matrices.

El procedimiento trata de extraer factores de forma sucesiva de manera que cada uno explique tanta varianza como sea posible de la matriz poblacional, no de la muestral. Sus resultados son muy semejantes a los de otros métodos cuando las fiabilidades de las pruebas y las comunidades de las variables son altas. Sin embargo, su eficiencia disminuye notablemente cuando el número de sujetos es bajo, por lo que en este caso es recomendable usar otros métodos.

#### 3.4.4. *Método alfa*

El método fue desarrollado por Kaiser y Coffrey (1965) con el objetivo de enfatizar la generalización de los resultados a un universo de contenido en una batería que consideran una muestra no aleatoria de ese universo. Como el lector sabe, el coeficiente alfa es una medida de la consistencia interna de un test, de su fiabilidad. Esta aproximación no produce resultados realmente diferentes de los que se obtienen con otros métodos más convencionales, por lo que no está claro si la ganancia en fiabilidad resulta relevante respecto de la extracción de factores.

#### 3.4.5. *Método de Imagen*

La idea central de este método es que la parte común (la imagen parcial) de una variable (con otras) es la que puede ser predicha desde las demás variables mediante procedimientos de regresión múltiple, en la que esa variable sería la dependiente y el resto las independientes. A partir de aquí se construye una matriz de covarianzas (no de correlaciones) imagen. Esta matriz puede ser analizada mediante cualquiera de los métodos de extracción mencionados más arriba. El problema más importante de esta aproximación reside en cómo interpretar las cargas factoriales imagen, puesto que no son correlaciones de las variables con los factores.

### 4. LIMITACIONES Y SUPUESTOS DEL ANÁLISIS

El análisis factorial es una técnica de reducción de datos que depende especialmente de las correlaciones (o de las covarianzas) y sus limitaciones y supuestos están, por tanto, íntimamente ligados a ellas. A continuación mencionaremos de forma sucinta las más relevantes.

#### 4.1. TAMAÑO DE LA MUESTRA Y DATOS PERDIDOS

Los coeficientes de correlación son muy sensibles al tamaño de las muestras. Son, de hecho, poco fiables cuando las muestras son pequeñas, lo que hace recomendable que el número de sujetos sea lo suficientemente grande para que el análisis produzca resultados estables. Aunque el número total de sujetos puede depender del grado de relación real entre las variables, Comrey y Lee (1992) sugieren que una buena muestra estaría compuesta por unos 300 sujetos.

La pérdida de sujetos también puede afectar a los coeficientes de correlación, particularmente cuando la atrición se produce de manera no aleatoria. En cualquier caso es recomendable eliminar las puntuaciones parejas con las perdidas, y examinar si hay alguna circunstancia en el estudio que pueda estar produciendo las pérdidas observadas.

#### 4.2. NORMALIDAD

Es un supuesto fundamental cuando el método de extracción es de máxima verosimilitud. En los demás casos su no cumplimiento puede ser importante sólo en lo relativo a la determinación del número de factores, particularmente si se emplean criterios estadísticos. Una solución aceptable en caso de incumplimiento consiste en aplicar transformaciones que puedan normalizar la distribución.

#### 4.3. LINEALIDAD

La linealidad es importante por dos motivos, primero, los coeficientes de correlación miden la relación lineal entre variables, y, segundo, el supuesto de normalidad multivariada la exige. Lógicamente si las correlaciones son no lineales, el análisis no producirá resultados adecuados.

#### 4.4. AUSENCIA DE PUNTOS EXTREMOS

Los coeficientes de correlación son medidas no robustas de relación lineal, lo que implica que son afectados de manera notable por la existencia de puntuaciones extremas. Su existencia compromete la solución factorial, por lo que es importante comprobar si existen, y eliminarlos o realizar transformaciones que limiten su influencia.

#### 4.5. MULTICOLINEALIDAD

La multicolinealidad hace referencia a la existencia de muy altas correlaciones entre las variables superficiales. Su existencia no representa un problema importante debido a que la matriz de correlaciones sólo tiene que ser invertida para computar las puntuaciones factoriales.

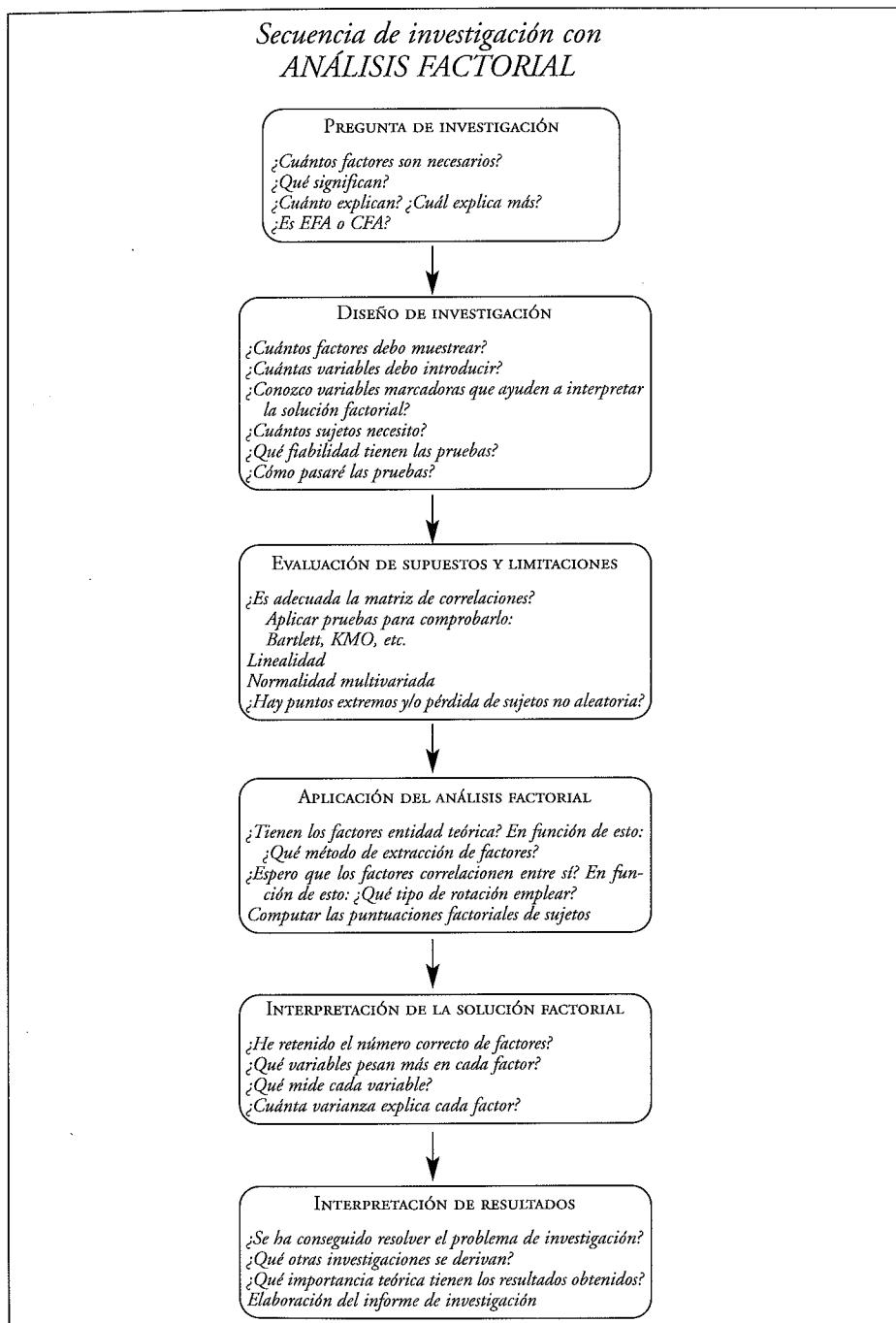
#### 4.6. FACTORIZABILIDAD DE LA MATRIZ DE CORRELACIONES

El análisis factorial produce resultados adecuados cuando la matriz de correlaciones contienen varias correlaciones altas entre pares de variables. Eso hace, como hemos señalado anteriormente, que sea importante examinar la matriz en busca de indicios de que sea factorizable. En general, si ninguna correlación excede de 0,30 es bastante probable que no pueda llegarse a una solución aceptable, y deba pensarse en una técnica de análisis alternativa, como el análisis de *cluster* de variables.

### 5. LA SECUENCIA DEL ANÁLISIS FACTORIAL

El diseño de estudios que pretenden analizarse mediante análisis factorial difiere en algunos aspectos importantes del que se emplea cuando se pretende utilizar otras técnicas (véase Cuadro 4.8. en página siguiente). Para comenzar debemos tener presente que el análisis factorial es una técnica que corre un riesgo importante, su mal uso. No es infrecuente que investigaciones mal planificadas acaben teniendo como último recurso el análisis factorial, porque prácticamente siempre permite extraer algo de los datos. En este manual hemos defendido la idea central de que una buena investigación sólo puede realizarse cuando hay buenas preguntas que resolver, y de que la calidad del diseño es una función del interés teórico que la pregunta tenga. Puesto que el análisis factorial es, en buena medida, una técnica exploratoria, se puede estar tentado de proceder sin más a la aplicación de pruebas y esperar que luego salga algo. Es posible que esta aproximación tenga valor heurístico, y produzca como consecuencia importantes preguntas después, pero los datos obtenidos de esta forma corren el riesgo de ser mal analizados y mal interpretados. A continuación presentamos algunos aspectos fundamentales de las cuestiones que deben tenerse presentes en el momento de diseñar un estudio de tipo exploratorio.

Cuadro 4.8. La secuencia de investigación orientada al análisis factorial



En primer lugar, es importante situarse en el dominio de investigación con un conocimiento suficiente sobre cuáles podrían ser los factores que convenía estudiar. El examen detenido de la literatura, la consideración detallada de las tareas y pruebas que en ella se describen puede indicar qué tipos de constructos podrían explicar la variabilidad del comportamiento en ellas. En la medida de lo posible deben diseñarse estudios que incluyan un buen número de factores, tal vez cinco o seis, cuando pueda haberlos, con el objetivo de abarcar de forma relativamente completa el dominio teórico. En cualquier caso, es importante que los factores incluidos puedan ser realmente relevantes, lo que hará que la solución factorial sea más aceptable, menos distorsionada.

En segundo lugar, la selección de variables debe realizarse teniendo en cuenta que debe haber más de una por cada factor hipotético. La idea es simple, si queremos caracterizar un factor de forma adecuada debemos tener información también adecuada sobre cuáles son las variables que dependen de él. En este sentido puede ser relevante la inclusión de variables sobre las cuales sepamos con cierto grado de seguridad que dependen del factor. Estas variables marcadoras pueden ayudar de forma importante en la interpretación de la solución factorial. Por el contrario, es también relevante considerar la posibilidad de que algunas de las variables seleccionadas puedan correlacionar con más de un factor. Este tipo de variables complejas son problemáticas, puesto que dificultan la interpretación de la solución factorial alcanzada, lo que hace aconsejable su no inclusión.

Como ya hemos mencionado es importante contar con un número suficiente de sujetos. Sin embargo, el número no es una garantía de que se alcancará una solución aceptable. La homogeneidad de la muestra puede alterar de forma artificial los coeficientes de correlación entre las variables superficiales, lo que puede llevar a soluciones consistentes con los datos, pero inadecuadas en cuanto a su generalización. La elección de una muestra heterogénea de sujetos puede llevar a que las variables correlacionen menos, pero asegura en mayor medida la validez de la investigación.

Otro aspecto importante de la selección de sujetos hace referencia a la utilización de muestras diferentes de sujetos, o a la «pasación» en momentos temporales diferentes de las pruebas a una misma muestra. En estas circunstancias se puede estar tentado de reunir todos los datos en un mismo análisis. Sin embargo, es preciso destacar que las soluciones factoriales en muestras de sujetos diferentes no tienen por qué coincidir, aunque naturalmente pueden hacerlo. Además, los factores no tienen por qué ser estables en el tiempo, aunque pueden serlo. Cuando se tienen dos muestras diferentes es preferible realizar el análisis por separado y comparar la solución factorial alcanzada en cada una de ellas.

Una vez diseñado el estudio y aplicadas las pruebas hay unas cuantas decisiones importantes que el investigador debe adoptar. En primer lugar qué método de extracción debería emplear. La cuestión puede resolverse de manera sencilla, puesto que, como hemos visto, los componentes principales son

sólo combinaciones lineales de variables de las que no se presume que tengan entidad psicológica alguna, mientras que en el resto de métodos se asume que los factores son entidades subyacentes a las variables superficiales. Por tanto, si se trata de descubrir constructos es preferible usar un método como el de factores principales, máxima verosimilitud, etc., mientras que si sólo se pretende combinar variables, la elección más adecuada son los componentes principales. Por lo demás, debe recordarse que las soluciones alcanzadas por diferentes métodos son muy semejantes cuando la muestra es lo suficientemente grande. La segunda decisión importante concierne al número de factores que se van a retener. Es posible retener tantos factores como variables superficiales, pero eso solamente desvirtuaría la esencia del análisis factorial, que es reducir el número de variables. Por tanto, el investigador debe tender a retener factores que expliquen una cantidad significativa de varianza, lo que suele conseguirse seleccionando factores cuyo autovalor sea mayor que uno. Esta cuestión es menos relevante cuando se parte de hipótesis sobre la solución, pero esto lleva al análisis factorial confirmatorio que presentaremos junto con los modelos de ecuaciones estructurales.

La tercera decisión importante concierne al tipo de rotación que debe emplearse. Una estrategia razonable consiste en comenzar por una rotación oblicua. El examen de la matriz de correlaciones nos pondrá sobre la pista de si una solución oblicua es adecuada. Concretamente, correlaciones inferiores a 0,30 son lo suficientemente bajas como para pensar que la opción más adecuada, por simplicidad de interpretación, es la ortogonal.

La obtención de las matrices de pesos constituye el hito fundamental del análisis. En este punto el investigador debe meditar si se requiere una interpretación teórica de los factores o no, lo que depende de si ha empleado componentes principales como método de extracción. Como hemos visto, sólo es razonable hacer referencia a constructos teóricos cuando se han empleado otros métodos de extracción. En cualquier caso, las cargas factoriales de las variables son el indicador más adecuado para realizar la interpretación. Recuérdese, no obstante, que cuando se ha realizado una rotación oblicua, es mejor interpretar los pesos de configuración, mejor que los pesos de estructura, puesto que éstos incluyen interacción entre factores.

## 6. LECTURAS RECOMENDADAS

- BISQUERRA, R., *Introducción conceptual al análisis multivariante. Un enfoque informático con los paquetes estadísticos SPSS-X, LISREL y SPAD*, Barcelona, PPU, 1989.
- FERRANDO, P. J., *Introducción al Análisis Factorial*, Barcelona, PPU, 1994.
- KIM, J. y MUELLER, C. W., *An Introduction to Factor Analysis: What it is and how to do it*, Beverly Hills, CA, Sage, 1978.
- FABRIGAR, L. R. y cols., «Evaluating the use of Exploratory Factor Analysis in Psychological Research», *Psychological Methods*, 4, 1999, págs. 272-299.

MARTÍNEZ-ARIAS, R., *Psicometría. Teoría de los tests psicológicos y educativos*, Madrid, Síntesis, 1995.

## LECTURAS COMPLEMENTARIAS

- COMREY, A. L., *Manual de Análisis Factorial*, Madrid, Cátedra, 1985.
- RUMMEL, R. J., *Applied Factor Analysis*, Evanston, Northwistern University Press, 1970.
- SHARMA, S., *Applied Multivariate Techniques*, Nueva Cork, Wiley, 1996.
- VALLEJO, G., *Técnicas Multivariadas Aplicadas a las Ciencias Comportamentales*, Oviedo, Universidad de Oviedo, 1992.

## 7. EJERCICIOS

1. Dada la matriz de covarianzas siguiente, obtenida a partir de los datos de 8 sujetos (sugerencia, use Poptools en Excel)

$$S = \begin{vmatrix} 8 & 0 & 1 \\ 0 & 8 & 3 \\ 1 & 3 & 5 \end{vmatrix}$$

- a) Computar los valores propios.
  - b) Computar los vectores propios asociados a cada valor propio.
2. Compute la matriz de correlaciones y justifique que la matriz es apropiada para realizar un análisis factorial.
3. En un análisis factorial el programa SPSS le presenta una matriz de estructura y otra de configuración.
- a) ¿Qué tipo de rotación se ha realizado?
  - b) ¿Qué matriz usaría para interpretar los factores encontrados?
4. El departamento de recursos humanos de una empresa quiere realizar un estudio sobre la motivación de sus ejecutivos. Los veinte ejecutivos responden a doce cuestiones en las que tienen que indicar en una escala de 1 a 10 su grado de acuerdo (1: desacuerdo total, 10: acuerdo total). Los datos aparecen en la tabla de abajo.
- a) Utilice el análisis de componentes principales para detectar los factores que pueden explicar la motivación de los ejecutivos. Intente primero una rotación ortogonal (Varimax) y después una rotación oblicua (Oblimin directo)
  - b) Realice el mismo análisis mediante el método de máxima verosimilitud.

SUJ.	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12
1	7	4	3	7	4	6	4	5	7	7	6	4
2	5	3	3	4	6	4	3	4	4	6	7	6
3	4	4	4	5	5	5	3	3	4	6	7	4
4	7	6	6	6	3	7	7	7	6	3	3	3
5	3	3	4	7	3	7	4	3	7	7	6	3
6	4	7	7	3	7	3	6	7	4	4	4	7
7	6	7	7	7	3	6	6	7	7	3	3	3
8	5	3	3	5	6	5	3	4	5	6	7	5
9	3	5	5	6	4	7	5	5	7	6	5	4
10	7	7	6	3	6	4	7	7	3	3	4	7
11	3	6	7	6	3	7	6	6	7	4	3	3
12	3	4	4	7	3	6	6	6	7	5	5	3
13	6	5	6	7	4	7	5	5	6	5	4	4
14	7	4	4	4	7	4	3	3	4	7	6	7
15	7	6	5	6	4	6	6	6	7	6	3	3
16	5	7	6	3	6	3	7	7	3	3	7	7
17	4	7	7	5	4	4	6	6	5	3	5	5
18	3	6	7	4	7	3	7	7	3	4	6	7
19	5	5	5	4	7	4	5	5	4	5	7	6
20	6	7	6	5	8	6	6	4	3	6	8	7

### RESPUESTAS A LOS EJERCICIOS

- Los valores propios de la matriz son: 10, 8 y 3. Los vectores propios son  $(0,33 \ 1 \ -0,2)$   $(1 \ -0,33 \ -0,60)$  y  $(0,67 \ 1,08^{-19} \ 1)$ , respectivamente para cada valor propio.
- La matriz de correlaciones es la siguiente:

$$\begin{vmatrix} 1 & 0 & 0,158 \\ 0 & 1 & 0,474 \\ 0,158 & 0,474 & 1 \end{vmatrix}$$

Hay dos índices computables con los datos disponibles, el determinante de la matriz ( $\lambda=0,75$ ) y el test de esfericidad de Bartlett ( $c^2_{(3)} = 1,486$ , N.S.). Por tanto, parece que la matriz no es adecuada para el análisis factorial.

- Se ha utilizado una rotación oblicua. Debería emplear la matriz de configuración.
- Sólo los tres primeros autovalores son mayores que 1.
  - La matriz de componentes rotados (Varimax) es:

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

	Componente		
	1	2	3
I1	-5,66E-02	6,134E-03	,992
I2	-,199	,898	,153
I3	-7,33E-02	,904	-,117
I4	,924	-,178	-3,03E-03
I5	-,936	-,128	,111
I6	,878	-,148	9,007E-02
I7	-6,27E-03	,925	-4,085E,03
I8	1,974E-02	,923	3,717E-02
I9	,960	-7,01E-02	-3,97E-02
I10	,103	-,911	6,780E-02
I11	,599	-,625	-,206
I12	-,966	6,033E-02	7,832E-02

Método de extracción: Análisis de componentes principales

Método de rotación: Normalización Varimax con Kaiser.

<sup>a</sup> La rotación ha convergido en 4 iteraciones.

La rotación oblicua (Oblimin directo) produce las siguientes matrices:

Matriz de configuración<sup>a</sup>

	Componente		
	1	2	3
I1	6,908E-02	-4,58E-02	,994
I2	,184	,888	,139
I3	5,465E-02	,910	-,131
I4	-,921	-,163	1,08E-03
I5	,940	-,149	,112
I6	-,875	-,139	9,386E-02
I7	-1,14E-02	,927	-1,84E-02
I8	-3,68E-02	,922	2,298E-02
I9	-,960	-5,29E-02	-3,73E-02
I10	-8,46E-02	-,914	8,221E-02
I11	,608	-,625	-,197
I12	,967	4,101E-02	7,611E-02

Método de extracción: Análisis de componentes principales

Método de rotación: Normalización Varimax con Kaiser.

<sup>a</sup> La rotación ha convergido en 6 iteraciones.

Matriz de estructura

	Componente		
	1	2	3
I1	5,844E-02	2,256E-02	,990
I2	,213	,904	,196
I3	8,749E-02	,904	-7,10E-02
I4	-,927	-,195	-1,36E-03
I5	,934	-,109	9,378E-02
I6	-,880	-,163	9,259E-02
I7	2,104E-02	,925	4,323E-02
I8	-4,94E-03	,923	8,456E-02
I9	-,962	-8,88E-02	-3,20E-02
I10	-,117	-,912	2,229E-02
I11	,588	-,617	-,244
I12	,967	7,968E-02	7,004E-02

Método de extracción: Análisis de componentes principales

Método de rotación: Normalización Varimax con Kaiser.

- b) La extracción por máxima verosimilitud produce la siguiente solución con rotación Varimax:

Matriz de factores rotados<sup>a</sup>

	Factor		
	1	2	3
I1	-5,43E-02	2,303E-02	,228
I2	-,160	,898	,307
I3	-4,94E-02	,940	,222
I4	,936	-,160	9,248E-03
I5	-,889	-,148	,389
I6	,900	-,130	,154
I7	3,766E-04	,904	,114
I8	-2,81E-02	,892	-,151
I9	,932	-6,86E-02	-,158
I10	,146	-,911	,243
I11	,535	-,597	,139
I12	-,940	4,316E-02	,252

Método de extracción: Análisis de componentes principales

Método de rotación: Normalización Varimax con Kaiser.

<sup>a</sup> La rotación ha convergido en 5 iteraciones.

## CAPÍTULO V

Análisis de *cluster*

## OBJETIVOS DE APRENDIZAJE

- 1) Distinguir el análisis de *cluster* de otras técnicas de reducción de datos que permiten agrupar objetos o variables.
- 2) Conocer en qué contextos de investigación es apropiado el análisis de *cluster*.
- 3) Conocer qué tipo de cuestiones de investigación puede ayudar a resolver el análisis de *cluster*.
- 4) Conocer las medidas de semejanza, distancias y asociación.
- 5) Conocer los tipos básicos de análisis de *cluster*.
- 6) Conocer los algoritmos fundamentales del agrupamiento jerárquico y del no jerárquico.
- 7) Conocer los procedimientos de valoración de las soluciones obtenidas en el análisis.
- 8) Saber evaluar e interpretar los resultados del análisis de *cluster*.

## 1. OBJETIVOS DEL ANÁLISIS DE «CLUSTER»

El análisis de *cluster* es una técnica multivariada no explicativa, en la que se intentan reducir las dimensiones necesarias para interpretar un conjunto de datos. Típicamente el objetivo fundamental consiste en identificar grupos de individuos sin que *a priori* se conozca criterio de agrupamiento alguno. Por ejemplo, pasamos una encuesta a una muestra amplia de personas sobre cuestiones ideológicas y queremos identificar a las personas que comparten características. Lógicamente lo que haríamos sería comparar a unos encuestados con otros y pondríamos juntos, en el mismo grupo, *cluster* o conglomerado, a los

que fuesen similares en su patrón de respuestas. El objetivo fundamental es, pues, asignar individuos a grupos de manera que se cumplan dos criterios importantes. Por una parte, los individuos asignados a un mismo grupo deben ser lo más semejantes posible entre sí, y por otra parte, cada grupo debe ser lo más diferente posible de los demás grupos. En definitiva se trata de que cada grupo sea lo más homogéneo posible, los individuos dentro de un grupo se parezcan mucho entre sí, y los grupos sean heterogéneos unos respecto de otros, es decir los individuos de un grupo se parezcan poco a los de los demás grupos. Aunque el objetivo del análisis de *cluster* es similar al del análisis discriminante, estos dos análisis no deben ser confundidos entre sí, puesto que en el primer caso la clasificación se produce sin conocer criterio *a priori*, mientras que en el segundo, el criterio de pertenencia al grupo está especificado claramente desde el principio.

Típicamente el análisis de *cluster* consiste en medir a un conjunto (*n*) de individuos en una serie de variables (*p*) métricas. A continuación es necesario establecer una medida de semejanza. Como veremos pueden emplearse un buen número de medidas de semejanza. La decisión sobre cuál utilizar es importante, porque puede determinar los agrupamientos que se realicen. Después hay que adoptar una decisión sobre qué técnica de agrupamiento se empleará. De nuevo hay una gran variedad de técnicas, pero pueden clasificarse en dos tipos: jerárquicas y no jerárquicas. La elección es, otra vez, importante porque puede determinar el resultado que se obtenga. Dado que el número de grupos no está determinado a priori porque no hay ningún criterio objetivo de clasificación, es preciso decidir el número de grupos que se van a formar. En las técnicas jerárquicas el número de grupos puede definirse *a posteriori*, una vez que han sido aplicadas. Sin embargo, en las no jerárquicas este número debe ser especificado antes de aplicar la técnica. Las técnicas de agrupamiento no proporcionan ningún criterio objetivo que permita al investigador tomar una decisión automática sobre el número de grupos que formará, aunque algunas de ellas, las jerárquicas como veremos más adelante, sí la facilitan. Finalmente la solución alcanzada tiene que ser evaluada e interpretada.

## 2. CLASES DE CUESTIONES QUE EL ANÁLISIS DE «CLUSTER» AYUDA A RESOLVER

El análisis de *cluster* es una técnica orientada a la clasificación cuando no existe un criterio previo establecido. Como hemos dicho ya, en principio puede emplearse para asignar individuos a grupos en función de su semejanza, pero también puede emplearse para formar grupos de variables. En el primer uso es semejante al análisis discriminante, y de hecho puede utilizarse en combinación con él, puesto que el objetivo es el mismo: asignar individuos a grupos. La diferencia fundamental reside en que en el análisis discriminante el criterio de formación de los grupos es conocido. El resultado de ambas técnicas es

similar, en cuanto que al final se obtendrán grupos de individuos que son semejantes en el patrón de datos observado en las variables medidas (dependientes en el caso del análisis discriminante). La segunda diferencia importante entre ambas técnicas reside en que el análisis discriminante permite resolver cuestiones acerca de la importancia que cada variable medida tiene para discriminar a los grupos entre sí, mientras que el análisis de *cluster* concede a todas las variables el mismo peso. Esto puede tener como consecuencia la inclusión de variables cuya contribución a la distinción entre grupos de individuos sea escasa o nula.

En el segundo sentido, el agrupamiento de variables, el análisis de *cluster* es similar al análisis factorial exploratorio o al de componentes principales, en cuanto que se trata de formar grupos de variables que sean semejantes entre sí. En ambas técnicas el investigador suele desconocer el número de grupos de variables que pueden formarse, y en ambas tiene que proceder de forma tentativa. Sin embargo, el análisis factorial exploratorio proporciona estadísticos relativamente claros acerca de si la solución final es satisfactoria, lo que no ocurre en el análisis de *cluster*. Por tanto, si el objetivo es agrupar variables, es más recomendable optar por un análisis factorial exploratorio que por el análisis de *cluster*.

El análisis de *cluster* ayuda, pues, a resolver preguntas que otras técnicas también permiten responder, pero puede ser una técnica inestimable en contextos de investigación en los que el investigador trata de identificar sujetos o variables similares entre sí.

## 3. MEDIDAS DE SEMEJANZA

El primer paso de la investigación analizable mediante análisis de *cluster* consiste en seleccionar un conjunto de *p* variables que van a ser medidas a un grupo *n* de sujetos. Tanto si el objetivo es agrupar variables como sujetos, la técnica es la misma, por tanto, de aquí en adelante hablaremos sólo del agrupamiento de sujetos. La importancia que las variables tengan para separar después grupos de sujetos es algo que sólo tiene una solución teórica y que depende, por tanto, del conocimiento que el investigador tenga del campo en el que trabaja. Baste, en principio, con indicar que las variables medidas pueden determinar el alcance de la solución obtenida.

El segundo paso importante consiste en la selección de una medida de semejanza entre los individuos. Debe quedar claro desde el principio que la semejanza sólo puede calcularse entre pares de objetos, esto es entre dos sujetos, entre un sujeto y un grupo, etc. El número de medidas de esta clase es enorme, pero pueden agruparse en dos tipos generales: medidas de distancia y de igualación. A continuación veremos algunas de las más importantes de cada tipo, teniendo en cuenta que se han incluido las medidas que SPSS introduce en el *cluster* jerárquico.

### 3.1. MEDIDAS DE DISTANCIA

Consideremos que hemos medido a cinco sujetos en cuatro pruebas, ansiedad estado (AE), ansiedad rasgo (AR), depresión (D) y número de errores cometidos realizando una tarea de detección de objetivos (E). Supongamos los resultados (ficticios) que aparecen en la tabla siguiente:

Sujetos	AE	AR	D	E
1	14	7	6	14
2	17	8	9	16
3	14	6	6	13
4	14	13	12	14
5	15	16	15	15

Cada uno de estos individuos está representado por un conjunto de cuatro números, un vector. Por ejemplo, el vector del sujeto 1 es (14, 7, 6, 14). Si denominamos  $x_{ik}$  a la puntuación del sujeto  $i$  en la prueba  $k$ , tendremos que el vector del sujeto 1 será  $X_1 = (x_{11}, x_{12}, x_{13}, x_{14})$ . En general, cualquier sujeto del estudio estará representado por su vector,  $X_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$ .

¿Cómo podemos computar la distancia entre los individuos de este estudio? Una buena cantidad de medidas se basan en la métrica de Minkowski, definida como:

$$d_{ij} = \sqrt[r]{\sum |x_{ij} - x_{jk}|^r} \quad (\text{Ec. 5.1})$$

donde  $d_{ij}$  es la distancia entre dos individuos  $i$  y  $j$ , y  $r$  es un valor entero que da lugar a una variedad de medidas de distancia, entre las que destacan especialmente dos: la distancia de bloques (*city-block*) y la distancia euclídea.

La distancia de bloques, también conocida como distancia de Manhattan, recibe su nombre de un hecho simple. Imagínese que debe llegar de una esquina de su salón a la esquina opuesta. Una de las formas de hacerlo consiste en recorrer la distancia pegado a las paredes, de modo que primero recorre la pared de su izquierda y luego la pared perpendicular a ella (véase Figura 5.1. en página siguiente). Así, la distancia recorrida es la suma de las paredes que ha recorrido. Matemáticamente esto implica que  $r = 1$ , lo que puede expresarse de la siguiente forma:

$$d_{ij} = \sum |x_{ij} - x_{jk}| \quad (\text{Ec. 5.2})$$

La distancia entre los sujetos 1 y 2 de nuestro estudio sería:

$$d_{12} = |14-17| + |7-8| + |6-9| + |14-16| = 3 + 1 + 3 + 2 = 9$$

### Análisis de *cluster*

A continuación podríamos calcular las distancias entre todos los pares de individuos posibles mediante este mismo procedimiento, con lo que obtendríamos una matriz cuadrada de distancias como la siguiente:

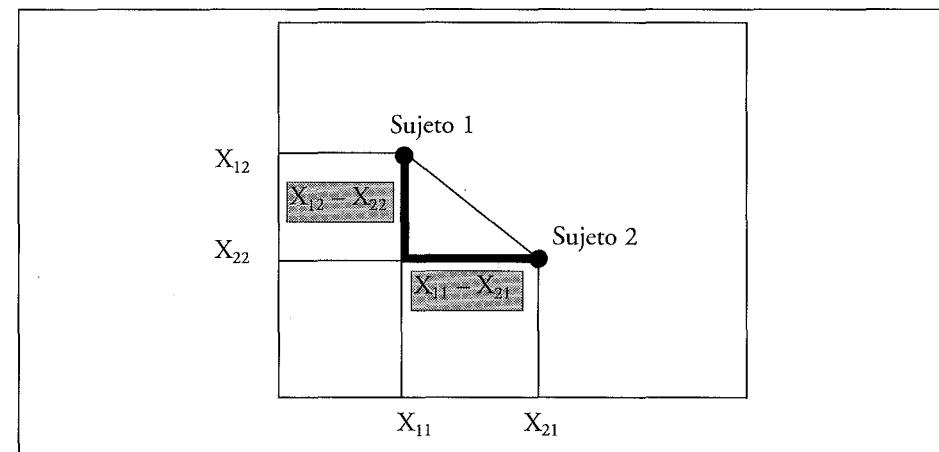
Caso	distancia de bloques de ciudad				
	1	2	3	4	5
1		9,000	2,000	12,000	20,000
2	9,000		11,000	13,000	17,000
3	2,000	11,000		14,000	22,000
4	12,000	13,000	14,000		8,000
5	20,000	17,000	22,000	8,000	

Ésta es una matriz de disimilaridades

Usted se habrá percatado de que ir de una esquina a la esquina de enfrente de su salón resulta más fácil si simplemente se camina en línea recta. Naturalmente el problema es simple, qué longitud tiene esa línea. La respuesta es fácil, usted está recorriendo la hipotenusa de un triángulo rectángulo, cuyos catetos son las paredes que recorrió antes cuando hablábamos de la medida de bloques. Por tanto, basta con aplicar el teorema de Pitágoras para obtener la distancia que se conoce como distancia euclídea. Cuando tenemos más de dos variables, la ecuación quedaría como:

$$d_{ij} = \sqrt{\sum |x_{ij} - x_{jk}|^2} \quad (\text{Ec. 5.3})$$

Figura 5.1. Distancias entre dos puntos en la métrica de Minkowski. La distancia de bloques implica recorrer las líneas remarcadas en negro. La distancia euclídea, recorrer la línea oblicua



lo que implica que cuando el valor de  $r$  es 2, tenemos la distancia euclídea. Ahora aplicaríamos esta ecuación para obtener las distancias euclídeas entre todos los pares de sujetos. Por ejemplo, la distancia euclídea entre el sujeto 1 y el sujeto 2 sería:

$$d_{ij} = \sqrt{|14 - 17|^2 + |7 - 8|^2 + |6 - 9|^2 + |14 - 16|^2} = \sqrt{9 + 1 + 9 + 4} = 4,796$$

La matriz de distancias euclídeas resultante tendría el siguiente aspecto:

Caso	distancia euclídea				
	1	2	3	4	5
1		4,796	1,414	8,485	12,806
2	4,796		5,568	6,856	10,247
3	1,414	5,568		9,274	13,638
4	8,485	6,856	9,274		4,472
5	12,806	10,247	13,638	4,472	

Ésta es una matriz de disimilaridades

El valor de  $r$  puede cambiarse de forma arbitraria. Los paquetes estadísticos como SPSS permiten una gama de valores relativamente amplia a través del parámetro etiquetado como *potencia*. Sin embargo, los dos valores de  $r$  más empleados son los que hemos presentado más arriba. Las distancias basadas en la métrica de Minkowski tienen un problema fundamental, su dependencia de la escala. Si se cambia la escala de una variable, la distancia entre los sujetos también cambia, pero también implica que si una de las variables está medida en una escala mucho más amplia que el resto, el peso de las que tienen la escala menor en la medición de la distancia puede ser irrelevante. Por ejemplo, imaginemos una variable que oscila entre 0-100 y otra que oscila entre 0-10. La distancia de bloques entre dos sujetos (sean sus valores (90, 8) y (70, 2)) dependerá más de la diferencia en la primera (90-70) que en la segunda (8-2) variable, a pesar de que en términos relativos, la diferencia entre ambos sujetos es considerablemente mayor en la segunda que en la primera, lo que puede apreciarse claramente si dividimos la primera variable por 10.

Una solución razonable al problema de la escala consiste en estandarizar las variables, de manera que su escala sea común. Para estandarizar las variables hay que sustraer la media y dividir por la desviación típica respectiva. Otra alternativa consiste en utilizar la distancia generalizada de Mahalanobis, que es invariante a la escala. Su cálculo requiere determinar previamente la matriz de varianzas-covarianzas entre las variables (sea  $S$ ), de modo que tendremos:

$$D_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)' S^{-1} (\mathbf{x}_i - \mathbf{x}_j)$$

donde  $D^2$  es la distancia de Mahalanobis al cuadrado y los  $\mathbf{x}$  son los vectores de los sujetos cuya distancia computamos. Desafortunadamente SPSS no permite el uso de esta medida en el contexto del análisis de *cluster*, aunque otros paquetes estadísticos como STATISTICA sí.

A continuación haremos referencia a un conjunto de medidas de semejanza ajenas a la métrica de Minkowski, pero que pueden usarse también en el caso de que las variables estén medidas en una escala de intervalo o de razón. Se trata de la medida coseno y de la distancia de Chebychev. La medida coseno se obtiene como un cociente entre la suma del producto puntual de los vectores de los dos sujetos y la raíz cuadrada del producto de las sumas de cuadrados de los vectores. Esto es:

$$\text{Coseno}_{\mathbf{x}_i, \mathbf{x}_j} = \frac{\sum x_i x_j}{\sqrt{\sum x_i^2 \sum x_j^2}}$$

que mide el coseno del ángulo entre los vectores de los sujetos  $i$  y  $j$ , lo que equivale a su coeficiente de correlación. En el caso de los sujetos 1 y 2, tendremos:

$$\text{Coseno}_{\mathbf{x}_1, \mathbf{x}_2} = \frac{14 \cdot 17 + 7 \cdot 8 + 6 \cdot 9 + 14 \cdot 16}{\sqrt{(14^2 + 7^2 + 6^2 + 14^2)(17^2 + 8^2 + 9^2 + 16^2)}} = 0,997$$

En la medida de Chebychev la distancia entre dos sujetos se computa determinando primero la diferencia absoluta entre los sujetos en cada variable, y seleccionando la mayor de las diferencias como medida de distancia. Esto es:

$$\text{Chebychev} = \text{Máximo } |\mathbf{x}_i - \mathbf{x}_j|$$

En nuestro ejemplo, la distancia de Chebychev entre el sujeto 1 y el 2 será el máximo de entre las diferencias siguientes: ( $|14-17|, |7-8|, |6-9|, |14-16|$ ), es decir, 3.

### 3.2. MEDIDAS DE IGUALACIÓN

Distinguiremos dos subtipos en función de la clase de datos: para variables dicotómicas o binarias (que pueden adoptar dos valores), y para frecuencias.

### 3.2.1. Medidas de igualación para variables dicotómicas

Las medidas de igualación son conocidas también como medidas de asociación. Son las medidas de elección cuando las variables están medidas en escala nominal y puede organizarse de manera simple en tablas de confusión de dos vías. Por ejemplo, supongamos que se midieron 10 variables a dos sujetos (1 y 2), que obtuvieron:

Sujeto	Variables									
	1	2	3	4	5	6	7	8	9	10
1	1	0	1	1	1	0	1	1	0	1
2	0	0	0	0	1	0	1	0	1	1

Donde 1 significa que el sujeto poseía una característica de la variable y 0 que poseía otra característica. Estos datos pueden organizarse en una tabla de doble entrada, con un sujeto en las filas, y el otro en las columnas, de modo que en las casillas tendremos el número de veces que los sujetos coinciden en la característica 1 (casilla a), en la característica 0 (casilla d), el sujeto 2 tiene la característica 1 pero el sujeto 1 tiene la 0 (casilla b), y el sujeto 2 tiene la característica 0, pero no tiene la 1 (casilla c). Esto es:

		Sujeto 1	
		1	0
Sujeto 2	1	3 (a)	1 (b)
	0	4 (c)	2 (d)

Ahora es posible calcular un buen número de medidas de asociación entre estos dos sujetos. Las más importantes son las siguientes:

*Distancia euclídea.* Se computa como  $DE = \sqrt{b + c}$ .

*La diferencia de Tamaño*, un índice de asimetría que oscila entre 0 y 1. Se obtiene como,  $\frac{(b - c)^2}{n^2}$  donde n es la suma de todas las casillas.

*La diferencia de Configuración*, que también oscila entre 0 y 1, y se obtiene como  $\frac{bc}{n^2}$ .

*Varianza.* Oscila entre 0 y 1, y se computa como  $(b+c)/4n$ .

*Concordancia simple.* Es la razón entre las coincidencias y el total. Se obtiene como  $(a+d)/n$ .

*Índice de Dice.* En esta medida se excluyen los casos en que los sujetos concuerdan en la ausencia de la característica (concuerdan en 0), y se ponderan el doble las concordancias en la presencia (concordancias en 1). Se computa como  $2a/(2a+b+c)$ .

*Índice de Lance y Williams.* Oscila entre 0 y 1. Se obtiene como  $(b+c)/(2a+b+c)$ .

*Coeficiente Q de Yule.* Conocido también como coeficiente de coligación. Oscila entre -1 y 1 y es independiente de las frecuencias marginales. Se obtiene como:  $(ad-bc)/(ad+bc)$ . (En el coeficiente Y de Yule a cada producto se le hace la raíz cuadrada.)

*Russel y Rao.* Tiene en cuenta las concordancias en la presencia (en 1). Se obtiene como  $a/n$ .

*Jaccard.* Semejante al anterior, pero excluye d. Se obtiene como  $a/(a+b+c)$ .

*Sokal y Sneath.* Se obtiene como  $2(a+d)/(2(a+d)+b+c)$ .

*Rogers y Tanimoto.* Se computa como  $(a+d)/(a+d+2(b+c))$ .

*Kulczynski.* Se computa como  $a/(b+c)$ .

*Coeficiente de correlación para variables dicotómicas* ( $\phi$ ). Se obtiene como

$$\phi = \frac{ad - bc}{(a + b)(a + c)(b + d)(c + d)}$$

### 3.2.2. Medidas de igualación para frecuencias

Las medidas más comunes cuando los datos corresponden a frecuencias son la chi-cuadrado y el coeficiente phi-cuadrado. La primera se basa en la prueba de chi-cuadrado para la igualdad de dos conjuntos de frecuencias, y la segunda es la chi-cuadrado normalizada por la raíz cuadrada de las frecuencias combinadas.

Como ha podido comprobarse la cantidad de medidas que pueden ser empleadas para agrupar a los sujetos es enorme. Sólo para el caso de variables dicotómicas SPSS introduce 27, 2 para frecuencias y 8 para variables de intervalo. Lo que indica que el investigador puede en la práctica encontrar una medida satisfactoria en función de sus objetivos de investigación. Sin embargo, el análisis de cluster no excluye que puedan emplearse medidas alternativas que sean más interesantes para el investigador. Lo importante es tener siempre presente que la medida de semejanza empleada puede determinar el agrupamiento que se va a obtener. Por supuesto, el investigador puede probar con varias medidas antes de seleccionar la que pueda ser más adecuada a sus objetivos.

## 4. MÉTODOS DE AGRUPAMIENTO

Hemos mencionado más arriba que hay métodos muy diferentes para conseguir el objetivo de agrupar a los sujetos más semejantes entre sí y que los grupos sean lo más diferentes posible entre sí. Esta variedad de métodos puede organizarse en dos tipos generales: métodos jerárquicos y métodos de partición. A continuación presentamos cada uno de estos tipos, incluyendo los métodos concretos más característicos en cada caso.

### 4.1. MÉTODOS JERÁRQUICOS

La característica general más importante de los métodos jerárquicos es que una vez que un individuo ha sido asignado a un *cluster* no puede ser reasignando a otro diferente. Hay dos subtipos fundamentales: de aglomeración o ascendentes y de disociación o descendentes. Los más utilizados en la actualidad, fundamentalmente porque son los que incorporan los paquetes estadísticos, son los de aglomeración, lo que ha hecho que algunos autores hablen de métodos jerárquicos para referirse exclusivamente a los métodos de aglomeración (por ejemplo, Sharma, 1996). Siguiendo esta tendencia a continuación expandiremos con detalle los de aglomeración, y haremos sólo una breve referencia a los métodos de disociación.

#### 4.1.1. Métodos jerárquicos de aglomeración

En estos métodos el agrupamiento se produce por un proceso iterativo que comienza siempre asumiendo que hay tantos *clusters* como sujetos y se finaliza con un solo *cluster* en el que están incluidos todos los sujetos. En la mayoría de los procedimientos (excluyendo el método de Ward) se utiliza la matriz de semejanzas para realizar el agrupamiento. La secuencia sería la siguiente. Se examina la matriz original de semejanzas y se agrupan los dos sujetos más semejantes entre sí, de manera que el número de *clusters* se reduce de  $n$  a  $n-1$ . A continuación se obtiene una nueva matriz de distancias entre los  $n-1$  *cluster*. El problema ahora reside en que en todos los *clusters* hay un sujeto, excepto en uno de ellos que contiene los dos sujetos que se agruparon en la primera etapa. Hay, por consiguiente, que adoptar una regla para calcular la nueva matriz de distancias. Justamente en cuál sea esa regla es donde difieren de manera sustancial los diferentes procedimientos de aglomeración. Una vez establecida la regla, el proceso de agrupamiento continua asignando los sujetos al *cluster* más cercano, hasta que sólo queda un *cluster* que los incluye a todos.

#### 4.1.1.1. «Método del centroide»

En este método cada *cluster* es reducido a un sujeto virtual o promedio, que es el centroide del grupo. Recordemos que el centroide es un vector que contiene las medias del grupo en cada una de las variables medidas. La semejanza entre los *cluster* es computada según el procedimiento que se haya decidido adoptar. Veamos cómo funciona el método en nuestro ejemplo, suponiendo que hemos decidido emplear la distancia euclídea al cuadrado. El Cuadro 5.1. presenta la secuencia de pasos del análisis. Nótese que necesitamos cuatro etapas (número de sujetos menos 1) para agrupar todos los sujetos en un único *cluster*.

Cuadro 5.1. Agrupamiento de aglomeración por el método del centroide

	Centroides de <i>cluster</i>				Caso	Distancia euclídea al cuadrado				
	a	b	c	d		1	2	3	4	5
1	14	7	6	14	1	0	23	2	72	164
2	17	8	9	16	2	23	0	31	47	105
3	14	6	6	13	3	2	31	0	86	186
4	14	13	12	14	4	72	47	86	0	20
5	15	16	15	15	5	164	105	186	20	0

La menor distancia corresponde a los sujetos 1 y 2, por tanto se agrupan para formar un nuevo *cluster*, representado por la media de los sujetos en cada variable (el centroide del *cluster*)

	Centroides de <i>cluster</i>				Caso	Distancia euclídea al cuadrado				
	M(1,3)	2	4	5		M(1,3)	0	3,78	27,1	65,4
2	15	7	7	14,3	2	3,78	0	47	105	
4	17	8	9	16	4	27,1	47	0	17	
5	14	13	12	14	5	65,4	105	17	0	

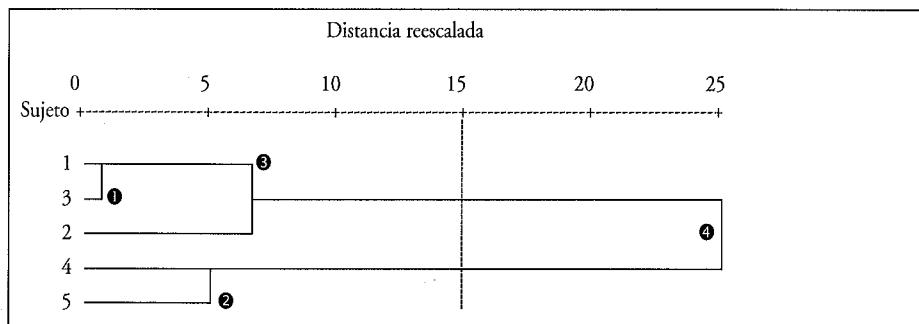
La distancia entre los *cluster* se computa entre los centroides. Los más cercanos ahora son 4 y 5, que se agrupan y son representados por la media del grupo en cada variable.

	Centroides de <i>cluster</i>				Caso	Distancia euclídea al cuadrado				
	M(1,3)	2	4,5	M(1,3)		0	3,78	43,5		
2	15	7	7	14,3	2	3,78	0	71,8		
Media(4,5)	17	8	9	16	M(4,5)	43,5	71,8	0		
Media(1,2,3)	14,5	14,5	13,5	14,5		0	43,5			
Media(4,5)	15	16	15	15		43,5	0			
Media(1,2,3,4,5)	14,8	10	9,6	14,4						

Los resultados obtenidos suelen representarse en un gráfico conocido como dendrograma, en el que la distancia de agrupamiento aparece en un eje y los sujetos en otro. El dendrograma es un árbol lógico que indica visualmente la

secuencia en que se han ido formando los *clusters*. El correspondiente al análisis que acabamos de realizar aparece en la Figura 5.2. Los números indican el momento en que los sujetos fueron agrupados. Así, en la primera etapa se agruparon los sujetos 1 y 3, después los 4 y 5, después el 2 se agrupó con el 1 y 3. Una de las ventajas más importantes que presenta esta Figura es que permite tomar una decisión clara respecto del número de *cluster* que es posible formar. Por ejemplo, si esperamos que haya dos tipos diferentes de sujetos, podríamos cortarlo donde aparece la línea punteada. Evidentemente el número de grupos diferentes de sujetos que finalmente se decide asumir puede ser una función de lo que teóricamente se espere obtener, sin embargo, como veremos más adelante hay algunos indicadores que pueden proporcionar ayuda útil para tomar una decisión fundamentada estadísticamente.

Figura 5.2. Dendrograma del agrupamiento realizado con el método del centroide



#### 4.1.1.2. «Método de la vinculación promedio»

Es uno de los métodos preferidos para realizar el agrupamiento, debido a que emplea toda la información disponible sobre las distancias entre los individuos de diferentes *clusters*. Concretamente, la distancia entre dos *clusters* es computada como el promedio de las distancias entre todos los individuos de un *cluster* respecto de todos los individuos del otro. En la etapa inicial, como en el método centroide cada individuo es un *cluster*, de modo que la primera agrupación es idéntica en ambos métodos. En nuestro ejemplo esto implica que primero se agrupan los sujetos 1 y 3. Para calcular la distancia de este grupo respecto de los demás es necesario obtener la distancia del sujeto 1 a los demás y del 3 a los demás. La distancia con respecto a los sujetos 2, 4 y 5 será el promedio de esas dos. Por ejemplo, la distancia respecto del sujeto 2,  $d_{(1,3),2}$ , sería (los datos necesarios aparecen en la matriz de distancias euclídeas al cuadrado del Cuadro 5.2., arriba a la derecha):

Distancia 1-2: 23

Distancia 3-2: 31

Por tanto, tendríamos que  $d_{(1,3),2} = (23+31)/2 = 27$ . Una vez computadas todas las distancias, se agruparían los dos *clusters* más cercanos, y así sucesivamente.

En SPSS este método es conocido como *vinculación inter-grupo*. Un procedimiento semejante es el de *vinculación intragrupo*, en el que se trata de minimizar la distancia entre todos los casos dentro de un *cluster*.

#### 4.1.1.3. «Método del vecino más próximo»

Este método se conoce también como de la distancia mínima (*single linkage*). Como en los casos anteriores, comienza asumiendo que cada sujeto es un *cluster*. A continuación se agrupan los más cercanos según la matriz de semejanzas, de modo que uno de los *cluster* tiene dos sujetos y el resto un solo sujeto. La regla para medir la distancia entre los *clusters* es la que da nombre al método, es la mínima de entre todas las que pueden definirse entre los miembros de uno y otro *cluster*. En nuestro ejemplo, la primera agrupación se realizaría como en los casos anteriores entre los dos sujetos más semejantes, el 1 y el 3. A continuación, reharíamos la matriz de distancias teniendo en cuenta que la distancia entre el *cluster* (1, 3), se obtendría computando la distancia entre el sujeto 1 y los demás, y el sujeto 3 y los demás. Respecto de cada uno de los sujetos 2, 4 y 5, la distancia sería la menor de las dos anteriores. Por ejemplo, respecto al sujeto 2, la distancia sería la menor de las distancias entre 1 y 2 (23) y entre 3 y 2 (31), es decir 23. Así, la matriz obtenida sería la siguiente:

	(1, 3)	2	4	5
(1, 3)	0	23	72	164
2	23	0	47	105
4	72	47	0	20
5	164	105	20	0

Ahora, la distancia menor corresponde a los sujetos 4 y 5, que serían agrupados para formar el *cluster* (4, 5). A continuación computamos las nuevas distancias entre *cluster* como en el caso anterior. Por ejemplo, la distancia entre el *cluster* (4, 5) y el sujeto 2 sería la menor de entre las distancias entre el sujeto 4 y el 2 (47) y entre el sujeto 5 y el 2 (105), es decir 47. La nueva matriz sería, por tanto:

	(1, 3)	2	(4, 5)
(1, 3)	0	23	72
2	23	0	47
(4, 5)	72	47	0

Ahora, el sujeto 2 está más cercano al *cluster* (1, 3) que al (4, 5), de modo que sería asignado a ese *cluster* y finalmente todos los sujetos formarían un solo grupo.

#### 4.1.1.4. «Método del vecino más lejano»

Este método se conoce también como de la distancia máxima (*complete-linkage*). La diferencia respecto del método anterior reside en que la distancia entre dos *clusters* es la máxima de las distancias entre todas las parejas posibles de miembros de un *cluster* con respecto a miembros del otro. En nuestro ejemplo, se agruparían primero los sujetos 1 y 3. La distancia entre el *cluster* (1,3) y el sujeto 2, sería ahora la mayor de las distancias entre 1 y 2 (23) y entre 3 y 2 (31), es decir, 31. Así, la nueva matriz de distancias sería la siguiente:

	(1, 3)	2	4	5
(1, 3)	0	31	186	164
2	31	0	47	105
4	186	47	0	20
5	164	105	20	0

A continuación se agrupan los *clusters* más cercanos, el 4 y el 5. Y así sucesivamente.

#### 4.1.1.5. «Método de Ward»

La característica más distintiva del método de Ward es que no emplea las distancias entre *clusters* para realizar la agrupación, sino que trata de hacer mínima la variabilidad intracluster, esto es de hacer que cada *cluster* sea lo más homogéneo posible. La homogeneidad se mide mediante la suma de los cuadrados de las diferencias entre los sujetos dentro del *cluster*. Por tanto, primero se agruparán los dos sujetos más semejantes, es decir cuya suma de cuadrados error sea menor. Teniendo en cuenta que el menor error posible se produce cuando cada *cluster* está formado por un solo sujeto, la estrategia consiste en asignar a un *cluster* el sujeto que produzca el menor incremento posible en la suma de cuadrados error. En nuestro ejemplo, la  $SC_{error}$ , suponiendo que agrupamos los sujetos 1 y 2, se obtendría computando primero la media de cada variable  $((14+17)/2 = 15,5; (7+8)/2 = 7,5; (6+9)/2 = 7,5 \text{ y } (14+16)/2 = 15)$ , de modo que el error sería:

$$((14-15,5)^2 + (17-15,5)^2 + (7-7,5)^2 + (8-7,5)^2 + \dots + (14-15)^2 + (16-15)^2 = 12),$$

mientras que el error si agrupamos a los sujetos 1 y 3 sería:

$$((14-14)^2 + (14-14)^2 + (7-6,5)^2 + (6-6,5)^2 + \dots + (14-13,5)^2 + (13-13,5)^2 = 1),$$

Si computamos los errores para el resto de parejas posible comprobaremos que los sujetos que primero deben agruparse son el 1 y el 3. A continuación volveríamos a computar las  $MS_{error}$  correspondientes a las soluciones de tres *clusters* y agruparíamos los *clusters* cuya  $MS_{error}$  fuese menor. Es importante caer en la cuenta de que una vez que un sujeto ha sido asignado a un *cluster* ya no puede ser excluido del mismo.

#### 4.1.2. Agrupamiento jerárquico mediante SPSS

En este apartado presentamos los comandos necesarios para realizar el análisis de *cluster* jerárquico aglomerativo y los resultados más importantes que proporciona SPSS, respecto de los datos del ejemplo que venimos considerando hasta aquí. Recordemos que los aspectos fundamentales a tener en cuenta en el análisis son relativos al tipo de variable (frecuencia, de intervalo o dicotómica), al tipo de medida de semejanza que va a emplearse (y que depende en gran parte del tipo de variable) y al método de aglomeración que se va a emplear. Para exemplificar supondremos que se ha decidido emplear la distancia euclídea al cuadrado (puesto que la variable es métrica) y el método del promedio o vinculación intragrupo. La secuencia de comandos sería:

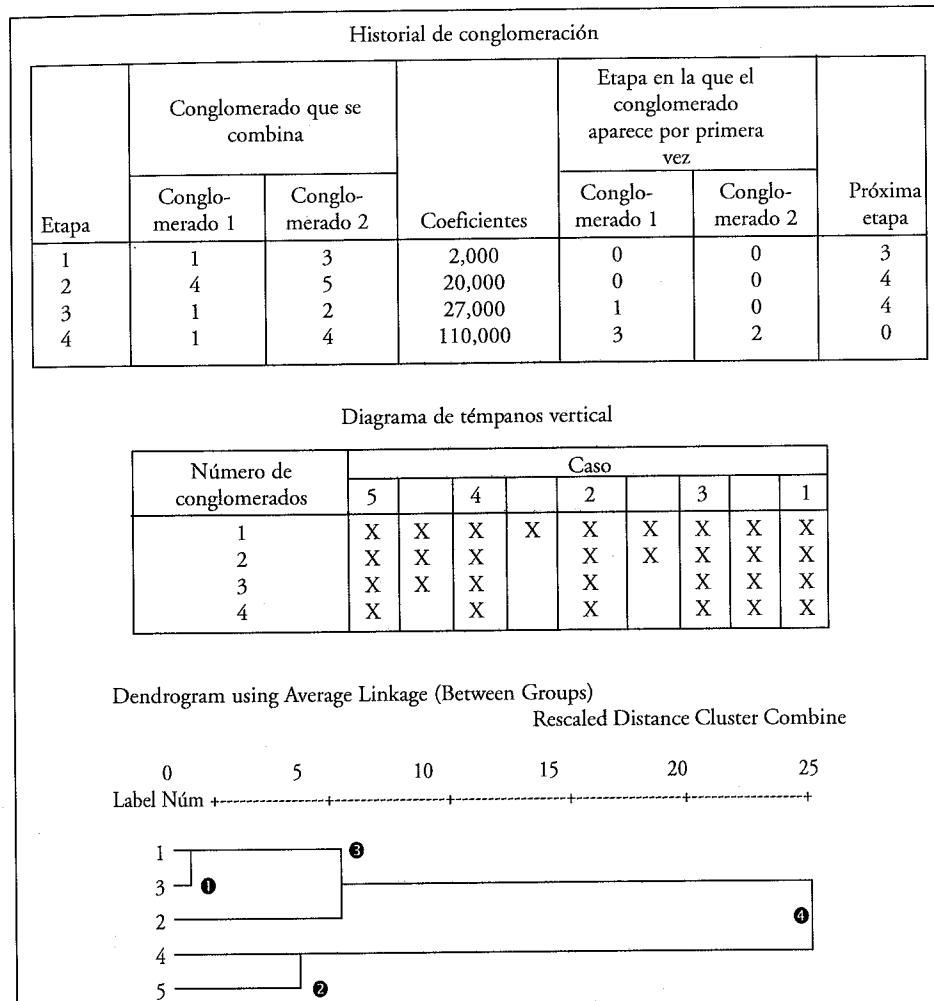
Analizar → Clasificar → Conglomerados jerárquicos → Variables: TipoA, TipoB, TipoC, TipoD → Conglomerar: Casos (sujetos) → Estadísticos → Mostrar: Matriz de distancias → Historial de conglomeración → Continuar → Gráficos → Dendrograma → Témpanos: todos los conglomerados → Continuar → Método → Método de conglomeración: Vinculación Intergrupo → Medida: Intervalo: Distancia euclídea al cuadrado → Continuar → Aceptar (análisis)

El Cuadro 5.2., que aparece en la página siguiente, presenta parte de la salida de SPSS. En primer lugar, el historial de conglomeración nos indica la secuencia en la que se han ido agrupando los sujetos. Primero se formó el *cluster* (1, 3), con una distancia de 2, después se formó el *cluster* (4, 5) con un coeficiente de 20; luego, el sujeto 2 se incluyó en el *cluster* (1, 3), con un coeficiente de 27, para formar el *cluster* (1, 2, 3) y finalmente todos los *clusters* se agruparon en un solo, el (1, 2, 3, 4, 5), con un coeficiente de 100.

En segundo lugar, SPSS presenta el diagrama de témpanos, en el que se representa gráficamente la progresión del agrupamiento. El diagrama se entiende mejor si se analiza de abajo hacia arriba, de más a menos *clusters*. Presenta la misma información que el historial de agrupamiento. Esta información es proporcionada también por el dendrograma.

Una cuestión importante concierne al número de *clusters* que finalmente se van a adoptar. Como hemos mencionado antes, la decisión debe estar guiada teóricamente.

Cuadro 5.2. Agrupamiento jerárquico de aglomeración mediante el método del promedio (vinculación intersujetos) y distancia euclídea al cuadrado



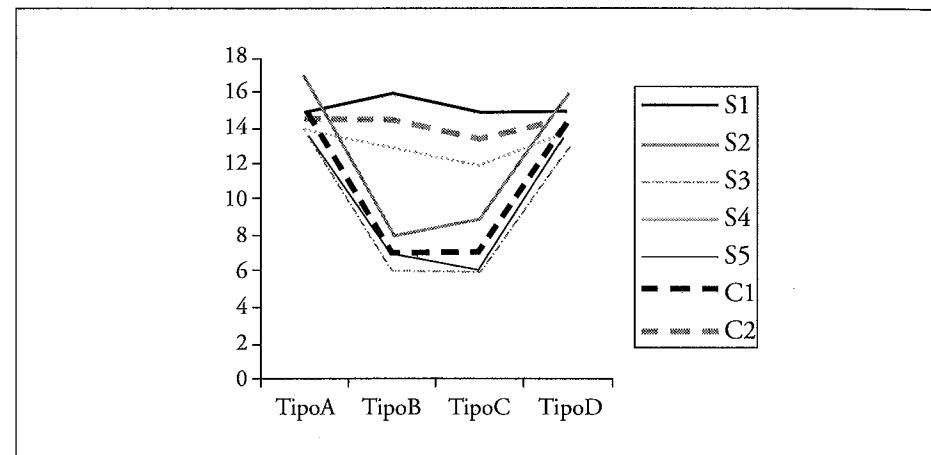
Por ejemplo, podríamos esperar que hubiese dos grupos, uno caracterizado porque sus valores en las variables TipoA y TipoD son mayores que en las variables TipoB y TipoC, y otro en el que sus valores son iguales en todos los tipos. Por tanto, podríamos pedir a SPSS que asignase los sujetos a dos *clusters*. La secuencia de comandos sería:

... → Guardar → Solución única: 2 grupos → Continuar → ...

lo que hará que se cree una nueva variable (CLU2-1), que contiene el número de *cluster* al que ha sido asignado cada sujeto. En nuestro caso, como ya sabemos, los

sujetos 1, 2 y 3 serán asignados al *cluster* 1 y los sujetos 4 y 5 al *cluster* 2. Si representamos gráficamente los resultados de cada sujeto en cada variable nos percataremos de que en efecto parece haber dos patrones distintos, que serán reflejados por las medias de *cluster* correspondiente. En la Figura 5.3. puede apreciarse cómo efectivamente el patrón de respuesta de cada *cluster* es diferente. Las líneas continuas corresponden a los sujetos y las discontinuas a las medias de *cluster*.

Figura 5.3. Patrones de valores de cada sujeto y de cada uno de los dos grupos solicitados a SPSS



#### 4.1.3. Evaluación de la solución de agrupamiento

Aunque no están disponibles en SPSS, hay algunos estadísticos que pueden ayudarnos a evaluar la solución de agrupamiento alcanzada, así como si el número de *clusters* seleccionado es adecuado o no. Son la correlación semi-parcial al cuadrado, la R cuadrado, la distancia entre *clusters* y la raíz cuadrática media de las desviaciones típicas. Típicamente se deberían obtener para cada etapa del análisis, puesto que proporcionan información acerca de los nuevos *clusters* y las consecuencias de añadir nuevos *clusters*. Puesto que SPSS no los proporciona nos limitaremos aquí a examinar su aplicación a la solución final.

##### 4.1.3.1. «Raíz cuadrática media de las desviaciones típicas (RMSSTD)»

Es la desviación típica promedio de todas las variables que forman un *cluster*. Su cálculo se realiza del siguiente modo. Primero se obtiene la suma de cuadrados de cada variable en el *cluster*, se suman las sumas de cuadrados de las variables, se obtienen los grados de libertad del *cluster* (número de sujetos en el

*cluster* menos 1). Se divide la suma de las sumas de cuadrados por la suma de los grados de libertad. Esto es:

$$\text{RMSSTD} = \sqrt{\frac{\sum_{j=1}^p \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2}{\sum_{j=1}^p g_j^l}} \quad (\text{Ec. 5.4})$$

donde  $p$  es, como siempre el número de variables medidas,  $i$  es el sujeto,  $j$  la variable, y  $g_l$  son los grados de libertad de la variable  $j$  en el *cluster*. Así, puesto que las medias del *cluster* (1,2,3) son: (15, 7, 7, 14, 33), las distancias al cuadrado de cada individuo a la media correspondiente en cada variable serían:

TipoA	TipoB	TipoC	TipoD	SC	RMSSTD
1	0	1	0,11		
4	1	4	2,78		
1	1	1	1,78	18,7	2,33

Por tanto, la suma de cuadrados (SS) sería 18,7 y RMSSTD = SS/g<sub>l</sub> = 18,7/(4\*2) = 2,33. Para el otro *cluster*, compuesto por los sujetos 4 y 5, RMSSTD sería 2,5. Desafortunadamente RMSSTD no puede interpretarse más que en sentido relativo, puesto que sólo podemos exigir que sea lo menor posible. Pero el procedimiento de aglomeración asegura que cuantos más *clusters* (menos sujetos en cada *cluster*) menor será RMSSTD, y cuantos menos *clusters* (más sujetos por grupo), mayor RMSSTD, lo que impide que pueda usarse de manera absoluta para tomar una decisión. Por ejemplo, una solución de tres *clusters*, (1,3) (2) y (4,5), tendría asociadas RMSSTD de 0,13, 0 y 2,5, respectivamente. Por el contrario, una solución de un único *cluster* tendría una RMSSTD de 9,2. En términos comparativos, parece claro que RMSSTD es considerablemente alta adoptando una solución de un solo *cluster* en relación a una de dos *clusters*, pero no hay tanta diferencia entre una solución de dos y una de tres. El examen de la Figura 5.3. puede ayudar a convencernos de que el sujeto 2 no es tan diferente de los sujetos 1 y 3 como para considerar que forma un *cluster* diferente.

#### 4.1.3.2. «R cuadrado»

El coeficiente R cuadrado se obtiene como un cociente entre la suma de cuadrados entre *clusters* ( $SC_{EG}$ , entre grupos) y la suma de cuadrados total ( $SC_T$ ). Indica, por tanto, la proporción de variabilidad total explicada por los *clusters*. Para obtener el índice es preciso computar ambas sumas de cuadrados

en cada variable, y sumarlas para obtener sumas de cuadrados del conjunto. Para computar la suma de cuadrados total es preciso obtener primero las sumas de cuadrados intragrupo, y sumarlas con las entre grupos, puesto que  $SC_T = SC_{EG} + SC_{IG}$  (véase Capítulo II). En nuestro caso, primero computamos las medias de *cluster*, y después sus diferencias con respecto a la media total (que incluye a todos los sujetos). Elevando al cuadrado y sumando tendríamos  $SC_{EG}$ . Para obtener la  $SC_{IG}$ , calculamos las diferencias de cada sujeto a su media de *cluster*, elevamos al cuadrado cada una y sumamos todos los cuadrados obtenidos. Estos cálculos aparecen resumidos en la siguiente tabla:

#### 1) Medias de cada *cluster* y media total

	TipoA	TipoB	TipoC	TipoD
Cluster 1	15,00	7,00	7,00	14,33
Cluster 2	14,50	14,50	13,50	14,50
Total	14,80	10,00	9,60	14,40

#### 2) Cuadrados de las diferencias entre medias de *cluster* y media total

0,04	9,00	6,76	0,00	$SC_{EG}$
0,09	20,25	15,21	0,01	51,36

#### 3) Cuadrados de las diferencias entre cada sujeto en cada variable respecto de la media de la variable en su *cluster* correspondiente

	1	0	1	0,11	$SC_{IG}$
Cluster 1	1	0	1	0,11	
	4	1	4	2,78	
	1	1	1	1,78	18,67
Cluster 2	0,25	2,25	2,25	0,25	
	0,25	2,25	2,25	0,25	10,00 $R^2$
					28,67 0,64

Así, pues,  $SC_T = 51,36 + 28,67 = 80,03$ , y por tanto  $R^2 = 51,36/80,67 = 0,64$ . Lo que implica que el 64 por 100 de la variabilidad total es debida a las diferencias entre los *clusters* obtenidos.

#### 4.1.3.3. « $R^2$ Semiparcial»

Ésta es una medida de la pérdida de homogeneidad dentro del grupo que se produce cuando a él se añade un nuevo sujeto en la etapa actual, respecto de la etapa precedente. Por ejemplo, al incluir el sujeto 2 en el *cluster* (1,3), se reduce la homogeneidad del grupo. Habitualmente se divide por la variabilidad total. Un valor pequeño es indicativo de que estamos incluyendo en el *cluster* un sujeto que es bastante semejante a los que ya había incluidos. En nuestro caso, la  $SC_{IG}$  (=18,67) del *cluster* 1 (1,2,3) la hemos obtenido en el proceso

de cálculo de  $R^2$ , que aparece en la tabla precedente. La  $SC_{IG}$  de los clústeres que se agruparon para formar el *cluster* 1 ((1,3) y (2)) eran 1 y 0, respectivamente. Por tanto, el cambio en la suma de cuadrados ha sido de  $18,67 - 1 = 17,67$ . Teniendo en cuenta la variabilidad total, tendremos que  $R^2$  Semiparcial =  $17,67 / 80,03 = 0,22$ . Lo que indica que la pérdida de homogeneidad no ha sido muy grande.

#### 4.1.3.4. «Distancia entre *clusters*»

La distancia entre dos *clusters* que se van a mezclar depende del método de aglomeración que se emplee. Por exemplificar, en el caso del método centroide se emplea la distancia euclídea entre los centroides de los dos *clusters*. Por ejemplo, el centroide del *cluster* (1,3) es (14, 6,5, 6, 13,5) y el centroide del *cluster* (2) es (17, 8, 9, 16). Por tanto, la distancia euclídea al cuadrado entre ambos es:

$$(14-17)^2 + (6,5-8)^2 + (6-9)^2 + (13,5-16)^2 = 26,5,$$

y la distancia euclídea es 5,15. Naturalmente, esta distancia no debe ser grande para que los dos *clusters* puedan mezclarse.

#### 4.1.4. *Métodos de división*

Los métodos de división afrontan la formación de los grupos en un orden inverso del empleado por los métodos de aglomeración. Se comienza asumiendo que todos los individuos son miembros de un mismo *cluster*, para a continuación asignarlos a dos grupos. El problema fundamental es cómo dividir la muestra. Con  $n$  sujetos las posibilidades son  $2^{n-1}$ . Posteriormente hay que seguir subdividiendo los *clusters* formados hasta alcanzar el número de *cluster* deseado (o hasta que cada *cluster* está formado por un solo individuo). Estas técnicas están en desuso, hasta el punto de que paquetes estadísticos como SPSS o STATISTICA no las incluyen como opciones de análisis. Es bien sabido que una técnica que deba ser aplicada a mano tiene pocas posibilidades de ser de amplio uso, por lo que aquí nos limitados a señalar las dos más importantes: la distancia astilla-promedio (*Splinter-Average Distancie*) y la detección automática de interacción (*Automatic Interaction Detection*, AID).

#### 4.1.4.1. «Distancia astilla-promedio»

El proceso comienza separando los objetos que tienen mayor distancia entre sí para formar dos grupos, denominados grupo propio ( $G_p$ ) y grupo otro ( $G_o$ ). Para cada sujeto se calculan dos distancias, una a su propio grupo y otra al grupo otro. Si la distancia a  $G_p$  es mayor que a  $G_o$ , el individuo se segregá de su grupo y se añade a  $G_o$ . La primera etapa del proceso continua hasta que las distancias dentro de cada grupo son menores que las distancias entre grupos. A continuación se procede de la misma manera en cada uno de los dos grupos formados. El procedimiento prosigue hasta que se obtiene el número de *clusters* deseado.

#### 4.1.4.2. «Detección automática de interacción»

El objetivo es dividir el grupo inicial en dos subgrupos utilizando como criterio el cociente entre la  $SC_{EG}$  y la  $SC_T$ , que se aplica a cada una de las variables medidas. Para realizar la partición se selecciona la variable en la que el cociente es máximo. El proceso continua con cada partición por separado de manera idéntica. Puede encontrarse en algunos paquetes estadísticos (STATISTICA 6.0) como una técnica de Minería de Datos (*Data Mining*).

### 4.2. MÉTODOS DE PARTICIÓN

La diferencia fundamental de estos métodos con respecto a los jerárquicos reside en que no suponen que la asignación de un sujeto a un *cluster* sea irreversible, un sujeto puede ser asignado a un *cluster* en una iteración y reasignado a otro *cluster* diferente en otra. Como veremos, los algoritmos difieren en cuál es el criterio de asignación que emplean. La segunda diferencia importante respecto de los métodos jerárquicos reside en que es absolutamente imprescindible especificar de antemano el número de grupos que se pretende obtener. El más importante, y el único incorporado en paquetes como SPSS y STATISTICA es el método K-Medias (*K-means*).

El método K-Medias trata de obtener el conjunto de grupos especificado por el investigador de manera que al final del proceso iterativo de asignación y reasignación de casos, cada sujeto sea miembro del grupo a cuyo centroide es más cercano. Por tanto, en el método hay tres aspectos fundamentales: ¿Cuáles son los centroides iniciales de cada grupo? ¿Cómo se evalúa la distancia de cada sujeto a los centroides de grupo? y ¿Cuántas iteraciones (ciclos de asignación-reasignación de sujetos) son necesarias para dar por concluido el proceso?

La elección de los centroides iniciales de cada uno de los *clusters* que se pretenden obtener no es, en absoluto, algo irrelevante. Aunque los centroides

cambiarán a medida que los sujetos son asignados y reasignados a los grupos, los valores seleccionados para iniciar el proceso pueden determinar el resultado del mismo. SPSS computa las distancias euclídeas entre los sujetos y selecciona como centros iniciales de cada *cluster* los sujetos que son más distantes entre sí. Por ejemplo, si queremos agrupar nuestros 5 sujetos en dos *clusters*, examinando la matriz de distancias euclídeas al cuadrado del Cuadro 5.1., nos percataremos de que los dos sujetos más distantes entre sí son el 3 y el 5, de modo que los centroides iniciales serán (14, 6, 6, 13) para el *cluster* 1 y (15, 16, 15, 15). Es recomendable que los centroides finales sean archivados y se utilicen en un reanálisis como centroides iniciales de los *clusters* que se desean formar.

La distancia de cada sujeto a los centroides de *cluster* es la euclídea. El error dentro de cada grupo es la suma de las distancias euclídeas de los sujetos del grupo al centroide de su *cluster*. Los sujetos que tienen una gran distancia respecto de su centroide de *cluster* pueden ser reasignados a otro *cluster* si con ello conseguimos reducir el error en el *cluster* al que son reasignados.

El número de iteraciones puede ser un determinante de la solución alcanzada. Si es bajo, es posible que la solución no sea óptima, es decir, que los errores dentro de los *clusters* pueda reducirse aún más cambiando algunos sujetos de unos *clusters* a otro. Es recomendable, por tanto, que se utilice un número elevado de iteraciones (normalmente con 20 es suficiente) para conseguir que se alcance la convergencia en la solución.

El análisis mediante SPSS se realiza empleando la siguiente secuencia de comandos:

Anализar → Clasificar → Conglomerados de K-Medias → Variables: TipoA, TipoB, TipoC, TipoD → Iterar y Clasificar → Número máximo de iteraciones: 10 → Criterio de convergencia: 0 → Usar medias actualizadas → Continuar → Centros → Escribir finales en archivo (indicar archivo) → Opciones → Información sobre el conglomerado de cada caso → Continuar → Aceptar (análisis)

El resultado ofrecido por SPSS aparece en el Cuadro 5.3. (véase página siguiente). En primer lugar aparecen los centros iniciales de los conglomerados. Después la historia de iteraciones donde se incluyen los cambios en los centros de los *clusters* en cada iteración. A continuación los sujetos que han sido asignados a cada uno de los *clusters* solicitados. Finalmente, aparecen los centroides obtenidos tras alcanzar la convergencia o realizar el número de iteraciones especificado, y las distancias entre los centroides de los *clusters*.

Una variación del procedimiento, no incluida en SPSS, consiste en utilizar uno de los sujetos del *cluster* como centroide. En cada iteración se decide si el centroide debe cambiarse, es decir, si emplear otro sujeto reduciría el error dentro del *cluster*. Esta técnica es conocida como K-Medoides (Kaufman y Rousseeuw, 1990). Su principal inconveniente, aparte de no formar parte de

los paquetes estadísticos más extendidos, es que no funciona de manera apropiada con grandes muestras de sujetos.

Cuadro 5.3. Salida del programa K-Medias de SPSS

CENTROS INICIALES DE LOS CONGLOMERADOS		
	Conglomerado	
	1	2
TIPOA	14	15
TIPOB	6	16
TIPOC	6	15
TIPOD	13	15

HISTORIAL DE ITERACIONES <sup>a</sup>		
	Cambio en los centros de los conglomerados	
Iteración	1	2
1	2,186	2,236
2	,000	,000

PERTENENCIA A LOS CONGLOMERADOS		
Número de caso	Conglomerado	Distancia
1	1	1,453
2	1	3,432
3	1	2,186
4	2	2,236
5	2	2,236

CENTROS DE LOS CONGLOMERADOS FINALES		
	Conglomerado	
	1	2
TIPOA	15	15
TIPOB	7	15
TIPOC	7	14
TIPOD	14	15

DISTANCIA ENTRE LOS CENTROS DE LOS CONGLOMERADOS FINALES		
Conglomerado	1	2
1		
2		9,939

CENTROS INICIALES DE LOS CONGLOMERADOS

CENTROS INICIALES DE LOS CONGLOMERADOS		
	Conglomerado	
	1	2
TIPOA	14	15
TIPOB	6	16
TIPOC	6	15
TIPOD	13	15

HISTORIAL DE ITERACIONES<sup>a</sup>

HISTORIAL DE ITERACIONES <sup>a</sup>		
	Cambio en los centros de los conglomerados	
Iteración	1	2
1	2,186	2,236
2	,000	,000

PERTENENCIA A LOS CONGLOMERADOS

Número de caso	Conglomerado	Distancia
1	1	1,453
2	1	3,432
3	1	2,186
4	2	2,236
5	2	2,236

CENTROS DE LOS CONGLOMERADOS FINALES

CENTROS DE LOS CONGLOMERADOS FINALES		
	Conglomerado	
	1	2
TIPOA	15	15
TIPOB	7	15
TIPOC	7	14
TIPOD	14	15

DISTANCIA ENTRE LOS CENTROS DE LOS CONGLOMERADOS FINALES

Conglomerado	1	2
1		
2		9,939

La evaluación de la solución alcanzada puede realizarse por los mismos procedimientos que hemos presentado en el apartado 4.1.3. Respecto de la interpretación, es bueno tener en cuenta las recomendaciones presentadas más arriba, especialmente las que hacen referencia a la representación gráfica de los datos y la consideración de los centroides de *cluster*.

#### 4.3. ¿QUÉ MÉTODO DE AGRUPAMIENTO EMPLEAR?

Como hemos mencionado antes, la elección de una técnica de agrupamiento puede determinar la composición final de los *clusters*. La variedad de técnicas disponibles no debe enmascarar, sin embargo, algo obvio aún en el caso de técnicas de carácter descriptivo o exploratorio como el *cluster*. Nos estamos refiriendo a que la técnica está al servicio de unos objetivos de investigación y que son esos objetivos los que deben guiar la elección que se realice. Naturalmente, es importante conocer las características de cada procedimiento, y sus limitaciones, para que el método elegido consiga efectivamente que se alcancen esos objetivos. En las páginas siguientes haremos pues, referencia, a las limitaciones de las dos clases generales de métodos, jerárquicos y de partición.

##### 4.3.1. *Métodos jerárquicos*

Los métodos jerárquicos tienen varias ventajas sobre los no-jerárquicos. En primer lugar, no requieren que se especifique *a priori* el número de grupos que se desea formar. Esta ventaja es importante cuando el investigador no tiene una idea precisa de qué patrones de respuesta puede encontrar en sus sujetos. La técnica puede aplicarse con un conocimiento inicial mínimo. Además, las técnicas jerárquicas son menos sensibles a la presencia de puntos extremos que las técnicas de partición. Sin embargo, ambos tipos de técnicas no son excluyentes. Una práctica frecuente consiste en utilizar alguna de las técnicas jerárquicas para producir una solución exploratoria y las técnicas de partición para confirmar esa solución.

La cuestión de cuál de las diferentes técnicas jerárquicas emplear no puede, igual que la anterior, resolverse de una manera objetiva, puesto que no hay criterio generalmente aceptado. Actualmente los métodos más frecuentes son los del promedio y los del centroide. Sin embargo, en principio el investigador puede legítimamente utilizar cualquiera. Quizá lo relevante sea tener en cuenta los problemas que pueden aparecer en cada caso.

En general, un problema importante es el conocido como de encadenamiento, que consiste en el riesgo de que los sujetos vayan siendo asignados a *clusters* ya formados, en lugar de formar clústeres nuevos. Esto se produce cuando los sujetos no incluidos todavía en ningún *cluster* están más cercanos a los *clusters* ya formados que entre sí. El problema es más grave con el método de

la distancia mínima que con el de la distancia máxima. Cuando los *clusters* son no homogéneos este inconveniente se transforma en una ventaja, lo que hace que el método del vecino más próximo sería el más recomendable en ese caso. Sin embargo, este método es más perjudicado por la presencia de sujetos extremos que el del vecino más lejano. Esto implica que el primero puede agrupar más en un mismo *cluster* a sujetos que pertenecerían a *clusters* diferentes que el segundo. En resumen el método del vecino más lejano puede emplearse cuando se pretende construir *clusters* en los que los sujetos sean muy semejantes entre sí. Por su parte, el método de Ward es útil cuando se pretende que sean homogéneos y de aproximadamente el mismo número de sujetos (Sharma, 1996).

##### 4.3.2. *Métodos de partición*

Estos métodos tienen dos problemas fundamentales. En primer lugar, requieren que se especifique previamente el número de grupos que se desea obtener, o, lo que es lo mismo, requieren que se conozca el posible número de patrones diferentes de valores de las variables medidas. Esto lleva aparejada la necesidad de que se conozcan los centros iniciales de los *clusters*. Y, como hemos mencionado antes, centros diferentes pueden llevar a soluciones diferentes. Una solución razonable consiste en seleccionar centros iniciales de modo que las diferencias entre ellos sean lo más grandes posible. Esto puede lograrse si primero se emplea una técnica jerárquica y después se usa la solución obtenida como punto de inicio para la técnica de partición. En cualquier caso, algunos procedimientos, como el uso de los primeros sujetos de la lista (una de las opciones implementada en STATISTICA) o la selección aleatoria de los sujetos, pueden llevar a soluciones no satisfactorias.

Otro problema importante, especialmente en el caso del método K-Medias, es que no son aplicables de una forma aceptable cuando las variables son no-métricas, esto es, las medias de cada variable no son computables, y por tanto, es inútil cuando las variables son no-métricas, categóricas, ordinales o nominales. Huang (1997) ha propuesto un procedimiento aplicable a estos casos, consistente en utilizar las modas en lugar de las medias y métodos de conteo de frecuencias para actualizar las modas de los *clusters*. Naturalmente, es preciso incluir una nueva medida de distancia entre dos sujetos *i* y *j*. La propuesta por Huang es muy simple, la distancia es cero si los dos tienen el mismo valor en la variable *y* 1 si tiene distintos valores.

Por último, el método de K-Medias es muy sensible a la presencia de sujetos extremos y a datos que presentan ruido, por una razón simple, la media es un estadístico no robusto, que depende en gran medida de estas puntuaciones. En estas situaciones hay varias soluciones posibles, una consiste en transformar los datos para intentar eliminar las puntuaciones extremas. Pero también es posible emplear métodos jerárquicos cuya solución depende menos de estos valores.

## 5. LIMITACIONES

El análisis de *cluster* es una técnica exploratoria que, como ocurre en el análisis factorial exploratorio, puede ser empleada incluso en los casos extremos en los cuales el investigador solamente pretende «obtener algo» de los datos. Esta cuestión es todavía más grave si consideramos que el análisis de *cluster*, en sí mismo carece de supuestos, aunque naturalmente cuando se emplea una medida de semejanza, al menos deben cumplirse los supuestos de la medida. Por ejemplo, si se emplea la correlación de Pearson, es obvio que se asume que la relación es lineal. Por tanto, siempre que se aplique la técnica es posible obtener una solución de agrupamiento de sujetos (o de variables), pero no siempre ese agrupamiento tiene sentido. Dicho en otras palabras, el análisis trata de descubrir en los datos la presencia de grupos de sujetos cuya existencia ignora, pero es posible que tales grupos no existan, a pesar de lo cual, se obtendrá una solución de tantos grupos como se desee. Cuando los grupos no existen de hecho la solución carece de interés y de sentido teórico. El investigador debe examinarla de manera cuidadosa para cerciorarse de si es satisfactoria o simplemente un artefacto debido a las características de la técnica de análisis que se emplea. Dos son los análisis adicionales que pueden ayudar a valorar los resultados obtenidos: la validación cruzada o replicación y la validación externa.

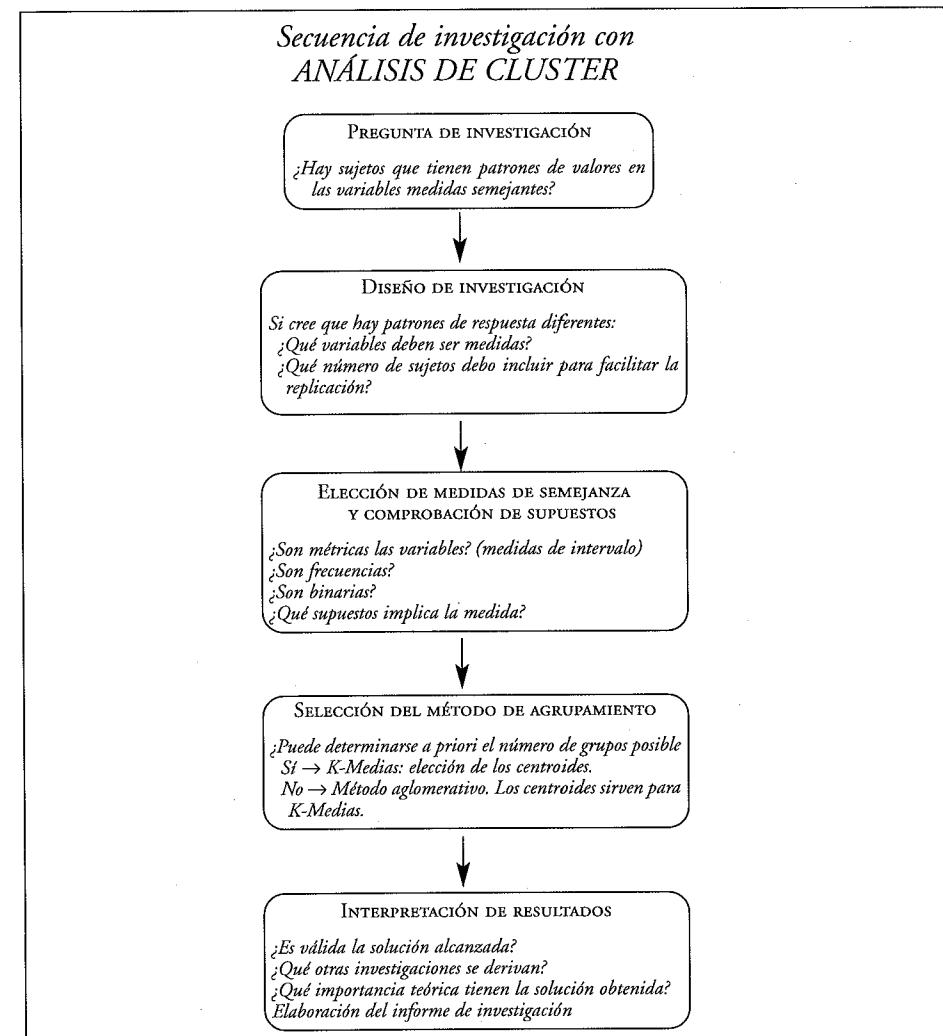
Una técnica efectiva para validar de forma cruzada los resultados del análisis de *cluster* es la replicación (Breckenridge, 1989). Puede realizarse cuando el número de sujetos es lo suficientemente grande y la muestra puede dividirse en dos mitades de forma aleatoria. A continuación la primera mitad es sometida a un análisis de *cluster* completo, y se obtienen los centroides de cada *cluster*. Después se calculan las distancias de cada sujeto de la segunda muestra a cada uno de los centroides de los *clusters* de la primera. Con esto se consigue agrupar los sujetos de la segunda muestra según las características de la primera. Inmediatamente después se procede a realizar un análisis de *cluster* de la segunda muestra, sin tener en cuenta ahora las características de la primera, pero con los mismos parámetros (método, medida de distancia, etc.). Ahora tenemos dos medidas de agrupamiento de la segunda muestra que pueden ser comparadas para evaluar el grado de acuerdo de las dos soluciones. Evidentemente cuanto mayor sea el acuerdo, más evidencia a favor de la estabilidad de la solución, pero también el investigador puede sentirse más confiado en que puede realizar una interpretación teórica. Una forma alternativa de valorar la solución consiste en aplicar un análisis discriminante para tratar de establecer con claridad si los grupos obtenidos son realmente diferentes y cuáles son las variables que contribuyen a esa diferencia.

La validación externa requiere que el investigador disponga de un criterio externo a la investigación para poder valorar la solución alcanzada. Por ejemplo, puede pedirse a jueces que clasifiquen a los sujetos en función de sus propios criterios. La correlación entre la clasificación de los jueces y la obtenida mediante el análisis indicará la validez de la solución obtenida.

## 6. LA SECUENCIA DEL ANÁLISIS DE «CLUSTER»

La secuencia de investigación característica del análisis de *cluster* aparece en el Cuadro 5.4. En primer lugar, debemos recordar que se trata de una técnica exploratoria en la que no se pretende contrastar hipótesis, sino simplemente determinar si hay patrones comunes de valores de las variables medidas. Esto no obsta para que el análisis sirva de complemento en investigaciones en las que el objetivo sí era contrastar hipótesis.

Cuadro 5.4. La secuencia de investigación orientada al análisis de *cluster*



Por ejemplo, Catena, Maldonado y Cádido (1998) emplearon el método K-Medias para comprobar si el patrón de respuestas a cuatro tipos diferentes de ensayos era diferente en dos grupos distintos de sujetos. Los grupos diferían en la frecuencia con la que tenían que emitir un juicio. Sus resultados indicaron que alrededor de un 80 por 100 de los sujetos del grupo que realizaba el juicio con alta frecuencia tenían un patrón muy semejante, mientras que el patrón del 20 por 100 restante era similar al del grupo que emitía el juicio con baja frecuencia. En segundo lugar, como hemos mencionado antes, la selección de las variables que van a ser medidas, la medida de distancia, y el método de agrupamiento pueden determinar la solución alcanzada.

Es importante recordar que las medidas de distancia seleccionadas dependen de la naturaleza de las variables, y que los métodos de partición exigen especificar de antemano el número de grupos posible, por lo que es recomendable emplear primero de forma exploratoria un método de aglomeración y utilizar sus resultados para realizar un segundo análisis de partición.

## 7. LECTURAS RECOMENDADAS

- BLASHFIELD, R. K. y ALDENDERFER, M. S., *Cluster analysis*, Sage University Papers, Quantitative Applications in the Social Sciences, Beverly Hills, CA, Sage, 1987.
- EVERITT, B. S., *Cluster analysis*, Londres, Edward Arnold, 1993.
- HERNÁNDEZ, M., *Técnicas de clasificación: medidas de proximidad, métodos de agrupamiento y validación de una solución Cluster*, La Laguna, Universidad de La Laguna, 1998.
- KAUFMAN, L. y ROUSSEEUW, P. J., *Finding Groups in Data: An Introduction to Cluster Analysis*, NY, Wiley, 1990.

## LECTURAS COMPLEMENTARIAS

- BRECKENRIDGE, J. N., «Replicating cluster analysis: Method, consistency, and validity». *Multivariate Behavioral Research*, 24, 1989, págs. 147-161.
- ESTARELLES, R.; DE LA FUENTE, E. I. y OLMEDO, P., «Aplicación y valoración de diferentes algoritmos no-jerárquicos en el análisis cluster y su representación gráfica», *Anuario de Psicología*, 55, 1992, págs. 63-90.
- HUANG, Z., «A Fast Clustering Algorithm to Cluster Very Large Categorical Data sets in Data Mining», en *Proceedings of SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, 1997.
- LANDAU, S.; LEESE, M. y EVERITT, B. S., *Cluster analysis*, Londres, Arnold, 2001.

## 8. EJERCICIOS

Use los datos del ejercicio 4 del Capítulo IV y:

- Realice el *cluster* de sujetos mediante la vinculación del vecino más próximo. ¿Cuántos *clusters* cree que debería retener?
- Realice el *cluster* mediante el método del centroide. ¿Cuántos *clusters* debería retener ahora?
- Realice un análisis de *cluster* de sujetos mediante el método de k-Medias, obteniendo tantos grupos como crea convenientes en función de sus respuestas a 1 y 2.
- Obtenga las puntuaciones de cada sujeto del ejercicio 4 del Capítulo IV en cada componente factorial retenido (use la rotación Varimax y el método de regresión). Agrupe los sujetos por el procedimiento k-Medias para formar 3 grupos.

## RESPUESTAS A LOS EJERCICIOS

- En función del dendrograma, o del gráfico de témpanos, podrían definirse dos o tres grupos de sujeto. Una solución de dos *clusters* incluiría en un grupo a los sujetos (2, 8, 3, 14, 19, 20, 6, 18, 16, 10, 17) y en otro los sujetos (4, 7, 9, 11, 12, 13, 15, 5, 1).
- En este caso parece más natural una solución de tres grupos, el primero con los sujetos (2, 8, 3, 14, 19, 20), el segundo con (6, 18, 16, 10, 17) y el tercero con (4, 7, 9, 11, 12, 13, 15, 5, 1).
- Con una solución de dos *clusters* tendríamos el primero con sujetos (1, 4, 5, 7, 9, 11, 12, 15, 17) y el segundo con los sujetos (2, 3, 6, 8, 10, 13, 14, 16, 18, 19, 20). La tabla de ANOVA indica que habría diferencias entre los dos *clusters* en los ítem (I4, I5, I6, I9, I11, I12). Una solución de tres *clusters* asignaría los sujetos (1, 2, 3, 5, 8, 9, 12) al *cluster* uno, los sujetos (6, 10, 14, 16, 18, 19, 20) al *cluster* dos y los sujetos (4, 7, 11, 13, 15, 17) al *cluster* 3. Ahora habría diferencias entre los grupos en todos los ítem, excepto I1. Por tanto, parece más razonable una solución de 3 *clusters*.
- Una solución de tres *clusters* agrupando sujetos en función de sus puntuaciones en los componentes principales daría los siguientes agrupa-

miento: primer *cluster*: (1, 2, 3, 8, 14, 19, 20), segundo *cluster*: (4, 5, 7, 9, 11, 12, 13, 15) y tercer *cluster*: (6, 10, 16, 17, 18). Los grupos difieren en los dos primeros componentes pero no en el tercero. Teniendo en cuenta los resultados del ejercicio 4 del Capítulo IV, es posible que el tercer componente no sea relevante y hubiera debido descartarse.

## CAPÍTULO VI

### Análisis de correlación canónica

#### OBJETIVOS DE APRENDIZAJE

- 1) Discriminar los contextos de investigación en los que es apropiado el uso de la correlación canónica.
- 2) Conocer las preguntas de investigación que la correlación canónica puede ayudar a responder.
- 3) Conocer las limitaciones y supuestos del análisis.
- 4) Conocer los aspectos fundamentales del cálculo de la correlación canónica.
- 5) Saber interpretar los resultados obtenidos en relación con las preguntas de investigación.

#### 1. CONCEPTOS FUNDAMENTALES EN EL ANÁLISIS DE CORRELACIÓN CANÓNICA

La correlación canónica es una correlación lineal, de la misma manera en que lo es la correlación de Pearson. Sin embargo, difiere de ésta en una cuestión sustancial. En la correlación de Pearson computamos un coeficiente que mide el grado de relación lineal entre dos y sólo dos variables medidas. En la correlación canónica también computamos coeficientes con los que pretendemos medir el grado de relación lineal entre variables, pero no respecto de dos variables, sino respecto de dos conjuntos de variables. Los dos conjuntos contienen variables medidas, por lo tanto, un conjunto no debe ser entendido como predictor y el otro como predicho, aunque en el diseño de la investigación se haya pretendido explicar uno a partir del otro.

Por ejemplo, supongamos que estamos interesados en medir la relación entre variables de personalidad y variables cognitivas. Imaginemos que medimos como variables de personalidad la ansiedad estado (E) y la depresión (D) y como variables cognitivas la velocidad para resolver problemas (P) y la realización de dos tareas simultáneas (S). Una aproximación no incorrecta, pero poco adecuada para evaluar la relación entre personalidad y cognición consistiría en computar todas las correlaciones posibles entre pares de variables de los dos conjuntos. De este modo tendríamos las correlaciones entre E y P, entre E y S, entre D y P y entre D y S, cuya interpretación podría ser más o menos clara en función del patrón que se obtenga. Nótese que, en cualquier caso no podríamos afirmar que las variables de personalidad son la explicación de las cognitivas ni viceversa, aunque de hecho pudieramos pensar que en efecto así es. La aproximación más razonable consistiría en computar un coeficiente de correlación entre el conjunto de variables cognitivas y el de variables de personalidad. La cuestión está, obviamente en cómo pueden computarse correlaciones entre conjuntos de variables.

La solución es más simple de lo que parece. Imaginemos que sumamos las variables cognitivas por un lado y las de personalidad por otro. De este modo tenemos dos combinaciones lineales de variables en las que todas las variables tienen el mismo peso. Ahora podemos obtener el coeficiente de correlación de Pearson entre las dos sumas. La interpretación del resultado es bien conocida, el coeficiente puede oscilar entre -1 y +1, y cuanto más alejado de 0 esté el coeficiente mayor es la fuerza de la relación. Sin embargo, el lector habrá advertido que las cosas no pueden ser tan fáciles, y que una aproximación semejante está llena de problemas. En primer lugar, las variables no suelen estar medidas en la misma escala, por lo que la fuerza de la correlación va a depender más de la correlación que haya entre las variables de mayor escala de ambos conjuntos. Más concretamente, si E está medida en una escala de 0 a 100, D en una escala de 0 a 20, P en una escala de 0 a 100 y S en una escala de 0 a 10, la correlación entre las sumas dependerá más de las variables E y P que de las demás. Este problema puede tener una solución sencilla, habrá advertido ya el lector, basta con estandarizar las variables de ambos conjuntos, con lo que todas tendrán la misma escala y el problema desaparecerá. Esta solución, sin embargo, no es suficiente por una razón importante adicional: la suma simple daría a todas las variables el mismo valor para determinar la correlación, pero es perfectamente razonable pensar que es posible que algunas sean más importantes que otras para la relación. Por ejemplo, si la alta ansiedad deteriora más la cognición que la depresión, entonces parece razonable adjudicar un mayor peso a la primera que a la segunda.

En la correlación canónica se suman las variables de cada conjunto por separado, pero la suma es ponderada, cada variable tiene asociado un peso, cada variable tiene asociada una importancia determinada. La suma ponderada de las variables de un conjunto se llama variado canónico del conjunto. Esto es, la suma ponderada de las variables de personalidad es el variado canónico

de personalidad, y la suma ponderada de las variables P y S es el variado canónico cognitivo. Por tanto, la correlación canónica es la correlación lineal entre dos variados canónicos. En términos más formales, podemos expresar estas ideas como:

$$\begin{aligned} V_{L1} &= b_{11}Y_1 + b_{12}Y_1 + \dots + b_{1q}Y_q \\ \text{y} \quad V_{R1} &= a_{11}X_1 + a_{12}X_1 + \dots + a_{1p}X_p \end{aligned}$$

donde  $V_{L1}$  es el primer variado canónico resultante de la mezcla ponderada (los  $b$  son los pesos) de las variables del conjunto Y, y  $V_{R1}$  es el primer variado canónico resultante de la mezcla ponderada (los  $a$  son los pesos) del conjunto X. Habitualmente un conjunto de variables se considera predictor (el conjunto X) y el otro se considera predicho o criterio (el conjunto Y). Por tanto, el análisis de correlación canónica puede asimilarse a un análisis de regresión múltiple multivariante, en el que es preciso predecir más de una variable a partir de más de un predictor.

Ahora quedan dos problemas más por resolver, en primer lugar cómo se determinan los pesos, y en segundo lugar cuántos conjuntos de pesos es posible obtener. La primera cuestión es indeterminada, no puede resolverse a menos que adoptemos un criterio. Evidentemente el criterio más razonable que puede adoptarse es que los pesos sean tales que la correlación entre los dos conjuntos sea lo más alta posible. Como veremos más adelante este criterio es fácil de cumplir, dadas ciertas condiciones. La segunda cuestión tiene una respuesta clara también. Hay varias combinaciones lineales posibles de las variables de cada conjunto, esto es hay varios variados canónicos posibles en cada conjunto de variables. Esto quiere decir que no hay una correlación canónica entre dos conjuntos de variables, sino varios coeficientes de correlación canónica. Esto puede dificultar la interpretación de los resultados, aunque, como veremos más adelante, no todos los coeficientes son significativos y a medida que se han obtenido coeficientes, los siguientes por obtener son cada vez menores.

En las páginas siguientes presentaremos primero los problemas de investigación que la correlación canónica ayuda a resolver, a continuación resumiremos los aspectos de cálculo más importantes y finalmente haremos referencia a la interpretación de los resultados y sus limitaciones y condiciones de aplicación. El capítulo concluirá con la secuencia de pasos necesarios para aplicar el análisis.

## 2. LAS CUESTIONES QUE EL ANÁLISIS DE CORRELACIÓN CANÓNICA AYUDA A RESOLVER

La correlación canónica es una técnica descriptiva, no explicativa. Esto implica que esta técnica no permite resolver cuestiones de tipo causal como: ¿son las variables de personalidad causantes de alteraciones en la actividad cognitiva? En su lugar, las clases de cuestiones son relativas, como en el caso de la

correlación de Pearson, a si puede decirse que hay alguna relación (lineal) entre los dos conjuntos de variables. Esto no implica, ni mucho menos, que la técnica debe ser despreciada. En muchas situaciones de investigación el investigador no puede manipular variables en sentido estricto. Por ejemplo, no podemos manipular la ansiedad rasgo, ni la depresión (y tampoco debemos), de la misma manera que podemos manipular (por lo menos fácilmente) la habilidad para resolver problemas o la capacidad para realizar dos tareas a la vez (aunque como se sabe esto puede lograrse con entrenamiento suficiente). Por tanto, las cuestiones de investigación que el análisis nos ayuda a responder serían las que siguen.

## 2.1. NÚMERO DE VARIADOS CANÓNICOS

¿Cuántas combinaciones posibles entre las variables de cada conjunto es posible realizar teniendo en cuenta que la correlación entre pares canónicos sea significativa? Si interpretamos los variados canónicos como dimensiones, la pregunta sería ¿cuántas dimensiones producen una correlación significativa entre los dos conjuntos de variables? La cuestión es, sin duda, interesante. Piénsese, por ejemplo, que en una combinación posible de las variables de personalidad damos mayor importancia a la ansiedad y en la otra a la depresión. Y piénsese también que en una combinación de las variables cognitivas damos más importancia al razonamiento y en la otra a la coordinación de tareas. Sería estimulante encontrar, por ejemplo, que cuando la ansiedad tiene mayor peso, la correlación es alta con el variado en el que el razonamiento también lo tiene, y que cuando es la depresión la que tiene mayor peso, la correlación es más elevada con el variado en el que la coordinación de tareas también lo tiene. Esto nos lleva al segundo problema.

## 2.2. INTERPRETACIÓN DE LA CORRELACIÓN CANÓNICA

¿Qué implica que un variado de un conjunto corraleione mejor con un variado determinado del otro conjunto? Para responder a esta cuestión es necesario primero determinar qué valor tiene cada variable en cada variado. Obviamente si la correlación es grande cuando el mayor peso corresponde a la ansiedad de un lado y al razonamiento por otro, parece razonable pensar que la correlación se produce precisamente debido a esas dos variables. La interpretación completa requiere, por tanto, examinar la correlación entre los variados, la correlación entre un variado y las variables que lo componen y la correlación entre un variado y las variables del otro variado.

## 2.3. PUNTUACIONES EN LOS VARIADOS

Los variados son combinaciones lineales de variables, de modo que podemos calcular la puntuación de cualquier sujeto en cualquier variado simplemente realizando la mezcla lineal adecuada de sus puntuaciones en las variables que componen el variado. Estas puntuaciones en el variado pueden usarse dentro del propio análisis, por ejemplo, para evaluar los sujetos que son extremos respecto de los demás, o en otros análisis, por ejemplo, para clasificar a los sujetos de acuerdo con esas puntuaciones y, por tanto, realizar una manipulación indirecta de variables.

## 3. EL ANÁLISIS DE LA CORRELACIÓN CANÓNICA

El análisis de correlación canónica es aplicable, teniendo en cuenta sus limitaciones y condiciones de uso, siempre que se hayan medido a varios sujetos dos conjuntos de variables diferentes, y el interés resida en determinar si los conjuntos correlacionan. En lo que sigue trabajaremos con el ejemplo que hemos esbozado anteriormente. Supongamos que hemos medido a ocho sujetos en las dos variables de personalidad (ansiedad estado, E, y depresión, D) y en las dos cognitivas (velocidad para resolver problemas, P, y realización de tareas simultáneas, S) y que los resultados fueron los siguientes:

Sujeto	E	D	P	S
1	12	6	53	19
2	12	7	47	18
3	9	11	51	14
4	13	6	49	21
5	14	19	49	20
6	17	18	42	27
7	13	20	50	26
8	14	17	51	23
Media	13	13	49	21

A partir de estas puntuaciones debe ser obvio que pueden calcularse varios tipos de correlaciones lineales: 1) entre las variables de personalidad (llamemos Y a ese grupo de variables, y sea a el número de variables), 2) entre las variables cognitivas (llamemos X a ese grupo de variables, y sea b el número de variables) y 3) entre las variables de personalidad (Y) y las cognitivas (X). En términos generales, podremos construir, pues, tres matrices de correlación, la primera de dimensión a-a (llamémosla  $R_{YY}$ ), la segunda de dimensión b-b ( $R_{XX}$ ), y la tercera de dimensión axb ( $R_{YX}$ ). Aún podemos añadir una cuarta que contendría las correlaciones de Y con X, y que, por supuesto, será simétrica de  $R_{YX}$ .

Así, definiremos la supermatriz M (su proceso de cómputo aparece en el Cuadro 6.1.), que contendrá todas las correlaciones posibles, esto es:

$$M = \begin{pmatrix} R_{YY}^{-1} & R_{YX} \\ R_{XY} & R_{XX}^{-1} \end{pmatrix}$$

Cuadro 6.1. Correlación canónica. Cómputo de las matrices R y de sus valores propios correspondientes

1) Computamos las puntuaciones diferenciales respecto de la media total de cada variable = Matriz Yt	<table border="1"> <thead> <tr> <th></th><th>E</th><th>D</th><th>P</th><th>S</th></tr> </thead> <tbody> <tr><td>1</td><td>-1</td><td>-7</td><td>4</td><td>-2</td></tr> <tr><td>2</td><td>-1</td><td>-6</td><td>-2</td><td>-3</td></tr> <tr><td>3</td><td>-4</td><td>-2</td><td>2</td><td>-7</td></tr> <tr><td>4</td><td>0</td><td>-7</td><td>0</td><td>0</td></tr> <tr><td>5</td><td>1</td><td>6</td><td>0</td><td>-1</td></tr> <tr><td>6</td><td>4</td><td>5</td><td>-7</td><td>6</td></tr> <tr><td>7</td><td>0</td><td>7</td><td>1</td><td>5</td></tr> <tr><td>8</td><td>1</td><td>4</td><td>2</td><td>2</td></tr> </tbody> </table>		E	D	P	S	1	-1	-7	4	-2	2	-1	-6	-2	-3	3	-4	-2	2	-7	4	0	-7	0	0	5	1	6	0	-1	6	4	5	-7	6	7	0	7	1	5	8	1	4	2	2
	E	D	P	S																																										
1	-1	-7	4	-2																																										
2	-1	-6	-2	-3																																										
3	-4	-2	2	-7																																										
4	0	-7	0	0																																										
5	1	6	0	-1																																										
6	4	5	-7	6																																										
7	0	7	1	5																																										
8	1	4	2	2																																										
2) Multiplicamos la traspuesta de Yt por Yt = Matriz de Sumas de Cuadrados y Productos cruzados	<table border="1"> <thead> <tr> <th></th><th>E</th><th>D</th><th>P</th><th>S</th></tr> </thead> <tbody> <tr><td>E</td><td>36</td><td>51</td><td>-36</td><td>58</td></tr> <tr><td>D</td><td>51</td><td>264</td><td>-40</td><td>113</td></tr> <tr><td>P</td><td>-36</td><td>-40</td><td>78</td><td>-49</td></tr> <tr><td>S</td><td>58</td><td>113</td><td>-49</td><td>128</td></tr> </tbody> </table>		E	D	P	S	E	36	51	-36	58	D	51	264	-40	113	P	-36	-40	78	-49	S	58	113	-49	128																				
	E	D	P	S																																										
E	36	51	-36	58																																										
D	51	264	-40	113																																										
P	-36	-40	78	-49																																										
S	58	113	-49	128																																										
3) La correlación se obtiene dividiendo por la raíz de los productos de las sumas de cuadrados	<table border="1"> <thead> <tr> <th></th><th>E</th><th>D</th><th>P</th><th>S</th></tr> </thead> <tbody> <tr><td>E</td><td>1</td><td>0,523</td><td>-0,68</td><td>0,854</td></tr> <tr><td>D</td><td>0,523</td><td>1</td><td>-0,28</td><td>0,615</td></tr> <tr><td>P</td><td>-0,68</td><td>-0,28</td><td>1</td><td>-0,49</td></tr> <tr><td>S</td><td>0,854</td><td>0,615</td><td>-0,49</td><td>1</td></tr> </tbody> </table>		E	D	P	S	E	1	0,523	-0,68	0,854	D	0,523	1	-0,28	0,615	P	-0,68	-0,28	1	-0,49	S	0,854	0,615	-0,49	1																				
	E	D	P	S																																										
E	1	0,523	-0,68	0,854																																										
D	0,523	1	-0,28	0,615																																										
P	-0,68	-0,28	1	-0,49																																										
S	0,854	0,615	-0,49	1																																										
4) Matriz M	<table border="1"> <thead> <tr> <th></th><th>1,377</th><th>-0,72</th><th>-0,68</th><th>0,854</th></tr> </thead> <tbody> <tr><td></td><td>-0,72</td><td>1,377</td><td>-0,28</td><td>0,615</td></tr> <tr><td></td><td>-0,68</td><td>-0,28</td><td>1,317</td><td>0,646</td></tr> <tr><td></td><td>0,854</td><td>0,615</td><td>0,646</td><td>1,317</td></tr> </tbody> </table>		1,377	-0,72	-0,68	0,854		-0,72	1,377	-0,28	0,615		-0,68	-0,28	1,317	0,646		0,854	0,615	0,646	1,317																									
	1,377	-0,72	-0,68	0,854																																										
	-0,72	1,377	-0,28	0,615																																										
	-0,68	-0,28	1,317	0,646																																										
	0,854	0,615	0,646	1,317																																										
5) Obtención de la matriz R	<table border="1"> <thead> <tr> <th></th><th>R<sub>YY</sub><sup>-1</sup>R<sub>YX</sub></th><th>R<sub>XX</sub><sup>-1</sup>R<sub>XY</sub></th><th>R</th></tr> </thead> <tbody> <tr><td></td><td><table border="1"> <tr><td>-0,73</td><td>0,734</td></tr> <tr><td>0,106</td><td>0,231</td></tr> </table></td><td><table border="1"> <tr><td>-0,34</td><td>0,03</td></tr> <tr><td>0,686</td><td>0,629</td></tr> </table></td><td><table border="1"> <tr><td>0,755</td><td>0,439</td></tr> <tr><td>0,122</td><td>0,148</td></tr> </table></td></tr> </tbody> </table>		R <sub>YY</sub> <sup>-1</sup> R <sub>YX</sub>	R <sub>XX</sub> <sup>-1</sup> R <sub>XY</sub>	R		<table border="1"> <tr><td>-0,73</td><td>0,734</td></tr> <tr><td>0,106</td><td>0,231</td></tr> </table>	-0,73	0,734	0,106	0,231	<table border="1"> <tr><td>-0,34</td><td>0,03</td></tr> <tr><td>0,686</td><td>0,629</td></tr> </table>	-0,34	0,03	0,686	0,629	<table border="1"> <tr><td>0,755</td><td>0,439</td></tr> <tr><td>0,122</td><td>0,148</td></tr> </table>	0,755	0,439	0,122	0,148																									
	R <sub>YY</sub> <sup>-1</sup> R <sub>YX</sub>	R <sub>XX</sub> <sup>-1</sup> R <sub>XY</sub>	R																																											
	<table border="1"> <tr><td>-0,73</td><td>0,734</td></tr> <tr><td>0,106</td><td>0,231</td></tr> </table>	-0,73	0,734	0,106	0,231	<table border="1"> <tr><td>-0,34</td><td>0,03</td></tr> <tr><td>0,686</td><td>0,629</td></tr> </table>	-0,34	0,03	0,686	0,629	<table border="1"> <tr><td>0,755</td><td>0,439</td></tr> <tr><td>0,122</td><td>0,148</td></tr> </table>	0,755	0,439	0,122	0,148																															
-0,73	0,734																																													
0,106	0,231																																													
-0,34	0,03																																													
0,686	0,629																																													
0,755	0,439																																													
0,122	0,148																																													
6) Ecuación determinantal para obtener valores propios del conjunto Y	$\begin{vmatrix} 0,755-\rho & 0,439 \\ 0,122 & 0,148-\rho \end{vmatrix} = 0 = (0,755-\rho)(0,148-\rho)-(0,439 \cdot 0,122)$																																													
7) Obtención de las soluciones del sistema determinantal	$\rho^2 - 0,903\rho + 0,0582 = 0; \text{ y con la fórmula cuadrática } \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ Tendremos que: $\rho_1 = 0,834$ y $\rho_2 = 0,07$																																													

A partir de M podemos obtener una matriz R de correlaciones que nos servirá para computar la correlación canónica, del siguiente modo (véase Cuadro 6.1., arriba):

$$R = R_{YY}^{-1}R_{YX}R_{XX}^{-1}R_{XY} \quad (\text{Ec. 6.1})$$

donde  $R_{YY}^{-1}$   $R_{XX}^{-1}$  son las inversas de las matrices de correlaciones de cada conjunto de variables. El análisis continúa obteniendo los valores propios de R, y posteriormente los vectores propios para cada conjunto de variables. Para ello, es preciso resolver primero los dos sistemas determinantales siguientes:

$$|R_{YY}^{-1}R_{YX}R_{XX}^{-1}R_{XY} - \rho^2 I| = 0 \quad (\text{Ec. 6.2})$$

para obtener los valores propios correspondientes a Y, y

$$|R_{XX}^{-1}R_{XY}R_{YY}^{-1}R_{YX} - \rho^2 I| = 0 \quad (\text{Ec. 6.3})$$

para los valores propios correspondientes a X.

Estos cálculos (ilustrados en el Cuadro 6.1.) pueden realizarse fácilmente mediante PopTools ([www.cse.csiro.au/CDG/poptools](http://www.cse.csiro.au/CDG/poptools)), una vez que hemos obtenido las matrices R. Es importante caer en la cuenta de que los valores propios obtenidos son los cuadrados de los coeficientes de correlación canónica. Por tanto, la correlación entre el primer par de variados canónicos será 0,913 (la raíz cuadrada del primer valor propio), mientras que la correlación entre el segundo par de variados será 0,264 (la raíz cuadrada del segundo valor propio). Además, la varianza compartida por el primer par será de 83,4%, pero la del segundo será sólo del 7%.

La significación de cualquier coeficiente de correlación canónica puede contrastarse mediante el test de Bartlett, que se aproxima a una chi-cuadrado según la siguiente expresión:

$$\chi^2_{ab} = -\left[N - 1 - \left(\frac{p+q+1}{2}\right)\right] \ln \Lambda \quad (\text{Ec. 6.4})$$

donde p y q son el número de variables en los conjuntos X e Y, respectivamente, N es el número total de sujetos, y  $\Lambda = \prod_{i=1}^s (1 - \lambda_i)$ . En nuestro caso, podemos contrastar la hipótesis de que hay alguna correlación canónica significativa aplicando ambas ecuaciones. Primero obtendremos  $\Lambda$ :

$$\Lambda = (1 - 0,834)(1 - 0,07) = 0,15438$$

después calculamos la chi-cuadrado:

$$\chi^2_4 = -[8-1-(2+2+1)/2] \ln(0.15438) = 8.40,$$

Examinando la tabla de  $\chi^2$  teórica, para un nivel de significación de 0,05, comprobaremos que ninguna de las dos correlaciones difiere significativamente de cero.

Cuadro 6.2. Obtención de los vectores de pesos para el conjunto Y de variables

Partiendo de R sustituimos  $\rho^2$  por su primer valor, 0,834.

$$\begin{pmatrix} 0,755 - 0,834 & 0,439 \\ 0,122 & 0,148 - 0,834 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = 0 = \begin{pmatrix} -0,079 & 0,439 \\ 0,122 & -0,686 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

de donde tendremos que:

$$\begin{aligned} -0,079 \cdot b_1 + 0,439 \cdot b_2 &= 0 \\ 0,122 \cdot b_1 - 0,686 \cdot b_2 &= 0 \end{aligned}$$

que es un sistema indeterminado en el que sólo podemos saber que:

$$b_1 = -5,557 \cdot b_2$$

Puede alcanzarse una solución asumiendo alguna restricción sobre los valores de los b, por ejemplo, que su suma de cuadrados sea la unidad:  $b_{12} + b_{22} = 1$ , por tanto,  $b_1 = 0,984$  y  $b_2 = 0,177$ . Estos coeficientes deben ser reescalados teniendo en cuenta la varianza de la combinación lineal, que será:

$$(0,984 \quad 0,177) R_{XX} \begin{pmatrix} 0,984 \\ 0,177 \end{pmatrix} = 1,182, \text{ y, por tanto, su raíz cuadrada, la desviación típica será } 1,087.$$

$$\text{Ahora } b_1 = 0,984/1,087 = 0,905 \text{ y } b_2 = 0,177/1,077 = 0,161.$$

A continuación procederíamos con el segundo valor propio, y luego con el segundo grupo de variables (X).

A continuación es preciso obtener los pesos para combinar linealmente las variables de cada conjunto. Para ello tendremos que operar en los dos sistemas siguientes:

$$(R_{YY}^{-1} R_{YX} R_{XX}^{-1} R_{XY} - \rho^2 I) b = 0; \quad (\text{Ec. 6.4})$$

para obtener los pesos del conjunto Y

$$(R_{XX}^{-1} R_{XY} R_{YY}^{-1} R_{YX} - \rho^2 I) \alpha = 0; \quad (\text{Ec. 6.5})$$

para obtener los pesos del conjunto X

Los cálculos necesarios aparecen ejemplificados en el Cuadro 6.2. (véase página anterior). Nótese que se trata simplemente de sustituir  $\rho^2$  por cada valor obtenido y resolver los sistemas correspondientes.

Estos análisis no pueden realizarse mediante menús de ventanas desplegables de SPSS, pero sí a través del editor de sintaxis. Para cargar el editor es preciso desplegar el menú Archivo, donde encontraremos «Sintaxis» en Nuevo. La secuencia de instrucciones sería: Archivo → Nuevo → Sintaxis. En la ventana que se despliega a continuación escribiremos la siguiente secuencia de comandos:

```
manova E D with P S /discrim all alpha(1) /print=sig(eigen dim)
```

que indica a SPSS que realice un análisis de correlación canónica en el que se considerarán como variables predichas E y D y como predictores P y S. Además, le pedimos que cree las variables canónicas (discrim) y que queremos la salida completa, con independencia de si la función es o no significativa. La significación se pide mediante el comando print. Los resultados aparecen resumidos en el Cuadro 6.3.

Cuadro 6.3. Análisis de correlación canónica mediante STATISTICA

1) Valores propios

Raíces	Raíz 1	Raíz 2
Variables	0,833791	0,070008

2) Vectores de pesos

Variables	Raíz 1	Raíz 2	Variables	Raíz 1	Raíz 2
E	-0,9058	-0,7457	P	0,33479	1,0975
D	-0,1616	1,1621	S	-0,7923	0,8299

3) Cargas canónicas

Variables	Raíz 1	Raíz 2	Variables	Raíz 1	Raíz 2
E	-0,990	-0,138	P	0,723	0,690
D	-0,636	0,772	S	-0,956	0,292

Una idea más intuitiva del análisis que acabamos de realizar puede obtenerse examinando los diagramas que aparecen en la Figura 12.1. (véase Capítulo XII), donde aparecen las relaciones entre las variables observadas y las combinaciones lineales que pueden realizarse de las mismas. Nótese que las variables contribuyen cada una con un peso determinado (representado por flechas de una sola dirección) a cada uno de los variados canónicos del conjunto Y, y que lo mismo ocurre con las variables P y S, respecto de los variados del conjunto X. Las flechas bidireccionales indican las correlaciones existentes entre los variados canónicos.

Los pesos canónicos son útiles para interpretar la importancia que cada variable tiene en sus variados correspondientes. Sin embargo, con muestras pequeñas estos pesos estandarizados son inestables, por lo que resulta más apropiado utilizar la correlación entre cada variable y sus variados, esto es, las cargas canónicas, también llamadas cargas estructurales. En nuestro ejemplo, las cargas aparecen en el Cuadro 6.3.

Finalmente, es importante computar la varianza de las variables observadas en cada conjunto por cada variado canónico. La razón más importante estriba en que, como cualquier coeficiente de correlación, la correlación canónica puede ser significativa, aunque sea extraordinariamente baja, cuando el número de sujetos es muy elevado. Esto implica que la varianza de un conjunto de variables explicada por el otro conjunto puede ser significativa, pero despreciable en términos prácticos. Una medida usada con frecuencia en este contexto es el índice de redundancia (Stewart y Love, 1968), que puede ser calculada para cada variable canónica. Esto es, tendremos redundancias de los variados canónicos de X explicados desde Y, y de los variados de Y explicados desde X. La redundancia total es la suma de las redundancias de los variados. El cálculo puede realizarse en varios pasos simples, primero obtendremos la varianza de Y que es explicada por sus variados canónicos (y la de X por los suyos). Sean esos índices  $p_Y$  y  $p_X$ , respectivamente. En ambos casos

$$p_Y = \sum_{i=1}^b \frac{c_{Li}^2}{b} \quad (\text{Ec. 6.6})$$

$$\text{y} \quad p_X = \sum_{i=1}^a \frac{c_{Ri}^2}{a} \quad (\text{Ec. 6.7})$$

donde los  $c$  son las cargas canónica del conjunto Y (L) y del conjunto X (R). Por tanto, en nuestro ejemplo, para el primer variado canónico de cada conjunto, será:

$$\text{y} \quad p_Y = ((-0,99)^2 + (-0,64)^2)/2 = 0,692;$$

$$p_X = (0,723^2 + (-0,956)^2)/2 = 0,719;$$

por tanto, el primer variado del conjunto Y extrae un 69,2 por 100 de la varianza de las variables E y D, mientras que el primer variado del conjunto X, extrae un 71,9 por 100 de la varianza de P y S. Esto implica que el segundo variado de Y extraerá el 30,8 por 100 de varianza restante (el máximo que puede extraerse es el 100 por 100 y el conjunto de todos los variados extrae siempre el máximo), mientras que en el conjunto X, el segundo variado extraerá el 28,1 por 100.

Ahora, la redundancia de cada variado de cada conjunto puede obtenerse como:

$$R_{V_i/W} = p_W \lambda_i, \quad (\text{Ec. 6.8})$$

donde W designa a cualquiera de los dos conjuntos de variables,  $V_i$  es un variado y  $\lambda_i$  es la raíz asociada al variado i. En nuestro ejemplo, la redundancia del primer variado de Y será:

$$R_{V_1/Y} = 0,692 * 0,834 = 0,578$$

El lector puede comprobar que la redundancia del segundo variado será 0,021, mientras que las redundancias ligadas a los variados del conjunto X serán: 0,599 y 0,019.

La redundancia total, esto es, la varianza total de un conjunto de variables explicada por el otro conjunto será:

$$R_{Y/X} = \sum R_{V_i/W_i}, \quad (\text{Ec. 6.9})$$

Esto es la suma de las redundancias del conjunto Y. Por tanto, en nuestro ejemplo tendremos que:

$$R_{Y/X} = 0,578 + 0,021 = 0,599.$$

Así, la varianza de Y explicada por X es del 59,9 por 100.

#### 4. LIMITACIONES Y SUPUESTOS DEL ANÁLISIS DE CORRELACIÓN CANÓNICA

El análisis de correlación canónica es una técnica que suele considerarse de carácter exploratorio, en lugar de explicativo o predictivo. La razón más importante estriba en que puede haber una infinidad de combinaciones lineales de las variables de cada conjunto, pero no está claro si esas combinaciones tienen significado psicológico de algún tipo, o dicho de otra manera, la significación estadística de la correlación no es garantía de que esa correlación tenga importancia teórica.

Desde un punto de vista estadístico, esta técnica es un caso particular del Modelo Lineal General, y por tanto, tiene las mismas limitaciones y supuestos que éste. En primer lugar, esto implica que no puede ser empleada cuando las variables no son métricas, sino categóricas. Otras limitaciones importantes que describiremos debajo son las siguientes: linealidad de la relación, ausencia de multicolinealidad, normalidad multivariada, número de sujetos y ausencia de puntos extremos.

#### 4.1. LINEALIDAD

En el análisis de correlación canónica estándar se asume que la relación entre las variables dentro y a través de conjuntos es lineal. El no cumplimiento de este supuesto producirá bajas correlaciones, aunque de hecho la correlación entre conjuntos sea alta. SPSS incorpora un programa (OVERALS), que puede emplearse para realizar correlación canónica no lineal.

#### 4.2. MULTICOLINEALIDAD Y SINGULARIDAD

Si las variables dentro de un conjunto están muy correlacionadas los pesos canónicos no serán fiables, esto se debe a que la varianza explicada por esas variables es también explicada por las variables que tienen alta correlación con ellas. El segundo inconveniente de la multicolinealidad es que no pueda obtenerse una solución adecuada, debido a que la matriz de correlaciones sea singular, no tenga determinante.

#### 4.3. NORMALIDAD MULTIVARIADA

Esta limitación se debe al test estadístico que se emplea para tomar la decisión sobre la significación de las correlaciones. Entre otras cosas implica que la combinación lineal de las variables de cualquiera de los dos conjuntos se distribuya normalmente. Afortunadamente este problema se reduce cuando el número de sujetos es suficientemente grande. En este sentido, Stevens (1986) recomienda que se empleen como mínimo 20 sujetos por cada variable medida. Esto llevaría en un ejemplo como el que hemos empleado aquí a un mínimo de 80 sujetos. Sin embargo, en general, cuanto mayor sea el número de soluciones que sea necesario interpretar (mayor número de correlaciones canónicas significativas), más alto debe ser el número de sujetos.

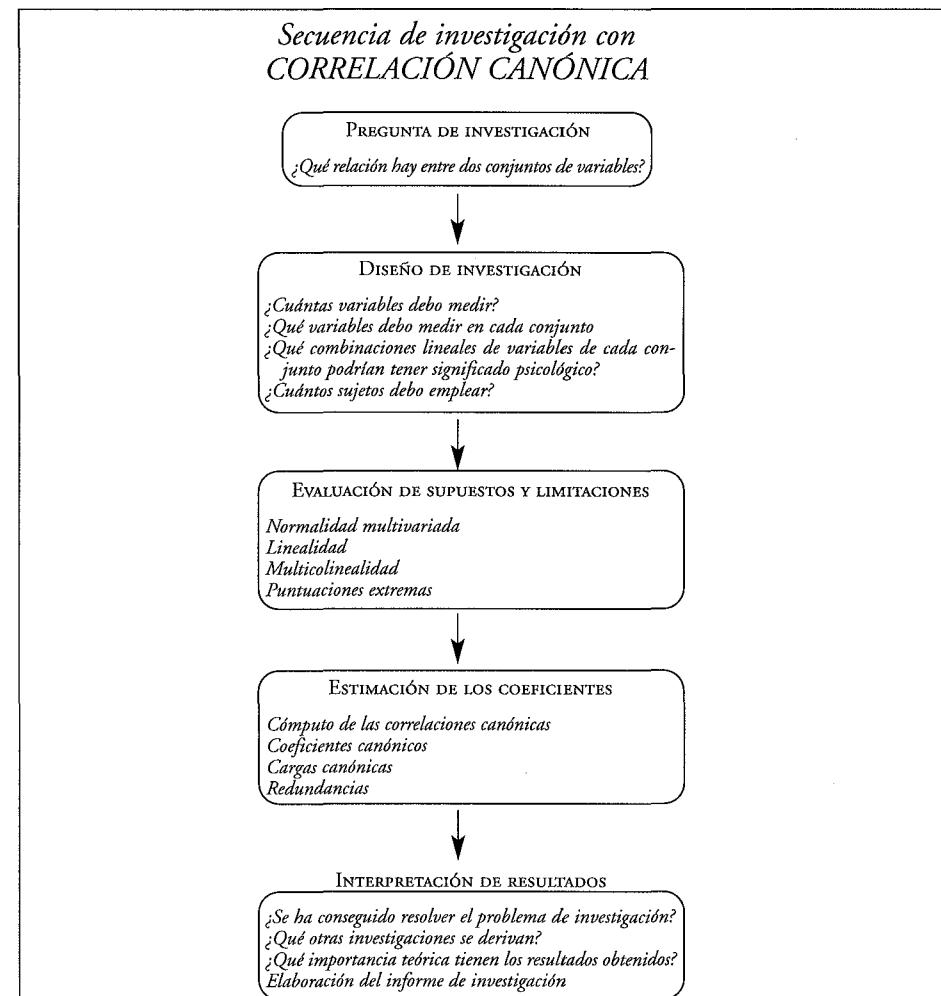
#### 4.4. AUSENCIA DE PUNTOS EXTREMOS

Como ocurre con el coeficiente de Pearson, cuanto menor es el número de sujetos, mayor es la influencia que los puntos extremos ejercen sobre el valor del coeficiente. Es recomendable, por tanto, que el investigador identifique esos puntos y trate bien de realizar transformaciones que reduzcan su influencia, bien de eliminarlos del análisis.

#### 5. LA SECUENCIA DE INVESTIGACIÓN EN EL ANÁLISIS DE CORRELACIÓN CANÓNICA

En apariencia las técnicas exploratorias exigen un menor grado de preparación de la investigación, sin embargo, es importante tener presente que, aun en ese caso, una planificación inadecuada puede llevar a la realización de un esfuerzo importante pero baldío. En el Cuadro 6.4. se presenta la secuencia característica de investigación orientada al análisis de datos mediante correlación canónica.

Cuadro 6.4. La secuencia de investigación orientada al análisis mediante correlación canónica



Como en el resto de técnicas multivariadas que presentamos en este manual, el investigador comienza planteándose un problema de investigación. El planteamiento adecuado del problema debe conducir a la adopción de decisiones sobre qué dos conjuntos de variables van a ser evaluados, cuántas variables se van a medir en cada conjunto, y, sobre todo, qué variables concretas son las que se van a emplear.

Debe tenerse presente que esas decisiones, particularmente la última, pueden determinar la importancia y el alcance de los resultados obtenidos. Además, el investigador debe decidir qué número de sujetos debe utilizar para conseguir una estabilidad adecuada en la estimación de los coeficientes. En ese sentido es importante tener presente que cuanto mayor sea el número de sujetos, más fiables serán los resultados, y también más probable es que las correlaciones resulten significativas, aunque debe considerarse también que la significación estadística no implica significación teórica.

Una vez que la investigación ha sido diseñada cuidadosamente y se han obtenido los resultados es importante comprobar hasta qué punto se cumplen los supuestos y se han conseguido evitar las limitaciones características del análisis. El investigador debe prestar especial atención a los supuestos de linealidad y multicolinealidad, puesto que pueden hacer inviable una interpretación razonable de sus datos. Así mismo, es importante que se exploren las puntuaciones en busca de sujetos extremos que, en caso de un insuficiente número total, podrían sesgar los coeficientes.

El análisis de los datos mediante SPSS es la siguiente etapa. El investigador debe prestar atención no sólo a la magnitud de los coeficientes de correlación, que pueden depender del tamaño de la muestra, sino a la varianza explicada por cada variado. La interpretación de los variados debe realizarse con preferencia a partir de las cargas canónicas. El peso de la interpretación debe vascular más sobre las variables que mayor carga (correlación con el variado) presenten.

Finalmente, el investigador debe evaluar el problema de investigación inicial y debe ponderar con detenimiento la posibilidad de realizar nuevas investigaciones que permitan un acercamiento al problema de naturaleza más explicativa que la correlación canónica.

## 6. LECTURAS RECOMENDADAS

- ATO, M., *Investigación en ciencias del comportamiento. I. Fundamentos*, Barcelona, PPU, 1991.  
 BISQUERRA, R., *Introducción Conceptual al Análisis Multivariable*, vols. I y II, Barcelona, PPU, 1989.  
 HAIR, J.y cols., *Análisis multivariable* (5.<sup>a</sup> edición). Madrid. Prentice Hall, 1999.  
 THOMPSON, B., *Canonical correlation analysis: uses and interpretation*, Sage University Papers, Quantitative Applications in the Social Sciences, Newbury Park, Sage, 1989.

## LECTURAS COMPLEMENTARIAS

- DILLON, W. R. y GOLDSTEIN, M., *Multivariate Analysis: Methods and applications*, Nueva York, Wiley, 1984.  
 JOBSON, J. D., *Applied Multivariate Data Analysis*, vol. 2.: Categorical and Multivariate Methods, Nueva York, Springer-Verlag, 1991.  
 SHARMA, S., *Applied Multivariate Analysis*, Nueva York, Wiley, 1996.

## 7. EJERCICIOS

1. Calcule la correlación canónica entre X e Y conociendo las siguientes matrices de correlación:

$$R_{XX} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$R_{YY} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$R_{XY} = \begin{pmatrix} 0,6 & 0 \\ 0 & 1 \end{pmatrix}$$

2. ¿Cuáles son los vectores propios asociados a las raíces obtenidas en la cuestión anterior?
3. Suponga que los datos del ejercicio 4 del Capítulo IV han sido obtenidos de manera que las 6 primeras variables son el conjunto Y (la preferencia por el modo de trabajo) y las 6 últimas son el conjunto X de variables que miden personalidad. ¿Qué relación hay entre preferencia y personalidad?
4. En la pregunta anterior, calcule las redundancias de X e Y.

## RESPUESTAS A LOS EJERCICIOS

1. Sólo hay una raíz de R, cuyo valor es 0,36 y la correlación canónica será 0,6.
2. El vector es (1, 0).
3. Los valores propios son (0,966 0,932 0,543 0,450 0,103 0,018).

4. La proporción de varianza extraída por cada raíz, y las redundancias son:

Conjunto Y		Conjunto X	
Prp. Var.	Redund.	Prp. Var.	Redund.
0,404968	0,391358	0,361255	0,349114
0,288286	0,268652	0,408852	0,381008
0,075913	0,041220	0,081565	0,044288
0,050567	0,022761	0,064847	0,029189
0,156892	0,016184	0,061340	0,006327
0,023375	0,000410	0,022141	0,000388

## CAPÍTULO VII

### Análisis de supervivencia

#### OBJETIVOS DE APRENDIZAJE

- 1) Conocer los contextos de investigación en los que puede emplearse el análisis de supervivencia.
- 2) Conocer el tipo de cuestiones de investigación que el análisis de supervivencia ayuda a resolver.
- 3) Comprender los conceptos básicos del análisis.
- 4) Saber interpretar los resultados (tablas de vida, funciones de supervivencia, etc.) en relación con los problemas de investigación planteados.
- 5) Conocer las limitaciones y supuestos del análisis.

#### 1. CONCEPTOS FUNDAMENTALES DEL ANÁLISIS DE SUPERVIVENCIA

El análisis de supervivencia es una técnica analítica cuyos inicios radican en la investigación médica y biológica, particularmente en relación con la supervivencia de pacientes, sin embargo, cada vez más está siendo empleada en el ámbito de las ciencias sociales con aplicaciones de naturaleza muy diferente. El origen de la técnica explica la terminología un tanto atípica que suele emplearse, y que seguiremos a lo largo de este capítulo, en este contexto. Por ejemplo, imaginemos que queremos comprobar la efectividad de un nuevo tratamiento de la depresión. Uno de nuestros intereses puede ser comprobar el número de días de terapia que los pacientes necesitan para que se produzca el alta. Es evidente que el estudio no puede prolongarse de manera indefinida en el tiempo, por lo

que al concluir el mismo es probable que todavía permanezcan pacientes en terapia (sujetos censados supervivientes), cuya mejoría no haya sido suficiente para ser dados de alta, pero tendremos también pacientes que han abandonado el tratamiento (abandonos de sujetos censados) y pacientes que han mejorado y han sido dados de alta antes de que acabe el tratamiento (fallecimientos). Está claro que el interés fundamental del investigador residirá en determinar el tiempo de terapia necesario para que se produzca el alta. Para poder realizar la evaluación será necesario tener información sobre el tiempo de observación de cada sujeto, en nuestro caso el número de días que estuvo en terapia (DÍAS); si el sujeto se mantuvo en el estudio (con lo cual no mejoró lo suficiente para ser dado de alta) o no; y si el sujeto recibió alguna intervención o no (la terapia). Un estudio de este tipo puede realizarse en un contexto formal de investigación, pero más frecuentemente se realiza en situaciones en las que el control sobre la situación es deficitario, por lo que es preciso tener en cuenta que: 1) el origen puede ser diferente para distintos sujetos, esto es, no todos entran en el estudio en el mismo momento. Por ejemplo, si se está aplicando en un gabinete o en un contexto hospitalario, está claro que los pacientes entrarán a formar parte del estudio a medida que vayan llegando al centro; 2) la última observación también puede realizarse en momentos diferentes por motivos muy distintos, por ejemplo, el paciente abandonó el centro, mejoró muy rápidamente y fue dado de alta (se produjo el evento terminal), o concluyó el período de estudio prefijado. Por lo tanto, 3) el período de observación, el tiempo que media entre el origen y la última observación, puede también ser diferente de unos sujetos a otros. A continuación vamos a utilizar este conjunto de conceptos para presentar los aspectos fundamentales del análisis de supervivencia.

## 2. ANÁLISIS DE SUPERVIVENCIA

Consideremos los datos ficticios que aparecen en el Cuadro 7.1. (véase página siguiente), relativos al ejemplo que estamos considerando sobre la eficacia del nuevo tratamiento de depresión. Supongamos que el tratamiento es recibido por la mitad de los sujetos, mientras que la otra mitad no reciben tratamiento alguno. En la primera columna de la tabla aparece el período de observación (los DÍAS de terapia) recibidos por cada uno de los sujetos. El máximo prefijado es de 45 días. La segunda columna contiene información sobre si el sujeto fue dado de alta (1) o no (0), y la tercera informa de si el sujeto recibió terapia (1) o no (0). Obsérvese que en la codificación del evento de interés no se distingue si el sujeto abandonó la terapia o fue dado de alta. Esta distinción la hará el programa de análisis de una manera muy simple: un sujeto que no reciba el alta y no acabe el tratamiento es un abandono. El análisis puede realizarse de varias maneras, mediante tablas de supervivencia, tablas de Kaplan-Meier y regresión de Cox. A continuación presentaremos los aspectos conceptuales más importantes para interpretar cada uno de ellos.

Cuadro 7.1. Datos ficticios sobre la eficacia de una nueva terapia de depresión, ordenados según el tratamiento y los días en terapia

Sujeto	DÍAS	ALTA	TRT	Sujeto	DÍAS	ALTA	TRT
1	17	1	0	15	12	1	1
2	18	1	0	16	15	1	1
3	19	0	0	17	15	0	1
4	26	1	0	18	21	1	1
5	33	1	0	19	23	1	1
6	33	0	0	20	25	1	1
7	36	1	0	21	27	1	1
8	39	0	0	22	29	1	1
9	40	0	0	23	30	1	1
10	40	0	0	24	41	1	1
11	43	0	0				
12	43	0	0				
13	45	0	0				
14	45	0	0				

## 2.1. TABLAS DE SUPERVIVENCIA

Las tablas de supervivencia pueden considerarse tablas de frecuencia extendidas. Su construcción se realiza dividiendo el rango total de tiempo en períodos temporales (Intvl Start Time, IST) cuya extensión es determinada por el investigador. Por ejemplo, puesto que en nuestro caso tenemos un máximo de 45 días, podemos dividir en intervalos de 5 días, de modo que tendríamos intervalos de 0 a 4,99, de 5 a 9,99, etc. A continuación se realiza el primer conteo básico, consistente en identificar el número de eventos de interés, altas (Number entrng this intvl, NEI) y de abandonos (Number drawn during intvl, NWI). Considerando la mitad derecha de la tabla observamos que hasta el intervalo 15-19,99 tendríamos a todos los sujetos que reciben el tratamiento 1 y no hay ningún tipo de suceso, por tanto, en los tres primeros intervalos 10 entradas, cero eventos y cero abandonos. En ese intervalo tenemos un total de dos sujetos, de los cuales uno recibe el alta (el evento terminal) y el otro abandona (puesto que su período de observación es de 15 días, menor que el máximo, y no ha recibido el alta). En el siguiente intervalo tenemos otros dos sujetos que sí reciben el alta. Y así continuaríamos hasta completar todos los intervalos. A partir de estos datos básicos podemos comenzar a realizar cálculos adicionales, por ejemplo, el número de observaciones expuestas a riesgo que serán el número de entradas menos la mitad del número de abandonos (Number exposd to risk, Risk). También podemos obtener algunas proporciones útiles para interpretar los resultados: la proporción de eventos terminales, altas (Propn Terminating, PT), la proporción de supervivencia (Propn Survigin, PS), y las funciones de supervivencia.

La proporción de eventos terminales es la razón entre el número de casos terminales (altas) y el número de casos en riesgo. La proporción de supervi-

vencia es la complementaria de la de eventos terminales. Finalmente, las funciones de supervivencia que se computan son las siguientes:

- 1) Proporción de supervivencia acumulada al final del intervalo (Cumulative survival at end, CPS). Se asume que la proporción acumulada en el primer intervalo es 1,0. El resto se computan como el producto de la acumulada hasta el intervalo anterior por la proporción de supervivencia del intervalo actual. Esto se debe a que se está asumiendo que las supervivencias son independientes de un intervalo a otro.
- 2) Densidad de probabilidad del intervalo (Probability Density, PD). Es la probabilidad de que se produzca el evento de interés en el intervalo especificado. Se computa como el cociente entre la diferencia de probabilidades acumuladas del intervalo actual respecto del precedente, dividido por la anchura del intervalo actual. Esto es:  $F_i = (p_{i-1} - p_i)/h_i$ , donde los  $p$  son las probabilidades acumuladas,  $i$  designa el intervalo y  $h$  la anchura del mismo.
- 3) Proporción de azar (Hazard rate, H). Es la probabilidad, por unidad de tiempo, de que un sujeto que haya comenzado el intervalo no lo finalice porque le suceda el evento de interés. Se computa como el número de eventos de interés por unidad de tiempo partido por el número medio de supervivientes en el centro del intervalo.
- 4) Los errores típicos de cada una de las funciones precedentes (SE of cumulative survival, SES; SE of Probability Dens, SED; y SE Hazard rate, SEH).
- 5) Mediana de supervivencia. La mediana es el valor de tiempo en el que la proporción acumulada es 0,5. Para su interpretación debe tenerse en cuenta que están incluidos los abandonos, por lo que realmente no indica el período en el cual la supervivencia es del 50 por 100.

El análisis mediante SPSS se realiza con la siguiente secuencia de comandos:

Análisis → Tablas de mortalidad → Tiempo: DÍAS → Mostrar intervalos de tiempo: 0 hasta: 45 → por: 5 → Estado: ALTA → Definir evento → Valor único: 1 → Continuar → Factor: TRATAMIENTO → Definir rango → Mínimo: 0; Máximo: 1 → Opciones → Tablas de mortalidad → Gráfico: Supervivencia → Continuar → Aceptar

Los resultados ofrecidos por el paquete estadístico aparecen en el Cuadro 7.2. (véase página siguiente). Junto a los títulos de columna en inglés hemos incluido su traducción. La eficacia de la terapia aparece de forma clara cuando comparamos sus funciones de supervivencia con las del control.

Cuadro 7.2. Tablas de mortalidad para el control (arriba) y el tratamiento (abajo)

IST	NEI	NWI	Risk	Term	PT	PS	CPS	PD	H	SECS	SEPD	SEH
0	14	0	14	0	0	1	1	0	0	0	0	0
5	14	0	14	0	0	1	1	0	0	0	0	0
10	14	0	14	0	0	1	1	0	0	0	0	0
15	14	2	13	1	0,077	0,923	0,923	0,02	0,016	0,074	0,015	0,016
20	11	0	11	0	0	1	0,923	0	0	0,074	0	0
25	11	0	11	1	0,091	0,909	0,839	0,02	0,019	0,105	0,016	0,019
30	10	1	9,5	1	0,105	0,895	0,751	0,02	0,022	0,125	0,017	0,022
35	8	1	7,5	1	0,133	0,867	0,651	0,02	0,029	0,143	0,019	0,029
40	6	4	4	0	0	1	0,651	0	0	0,143	0	0
45+	2	2	1	0	0	1	0,651	**	**	0,143	**	**

IST	NEI	NWI	Risk	Term	PT	PS	CPS	PD	H	SECS	SEPD	SEH
0	10	0	10	0	0	1	1	0	0	0	0	0
5	10	0	10	0	0	1	1	0	0	0	0	0
10	10	0	10	1	0,1	0,9	0,9	0,02	0,021	0,095	0,019	0,021
15	9	1	8,5	1	0,118	0,882	0,794	0,02	0,025	0,13	0,02	0,025
20	7	0	7	2	0,286	0,714	0,567	0,05	0,067	0,164	0,028	0,047
25	5	0	5	3	0,6	0,4	0,227	0,07	0,171	0,141	0,032	0,089
30	2	0	2	1	0,5	0,5	0,113	0,02	0,133	0,107	0,021	0,126
35	1	0	1	0	0	1	0,113	0	0	0,107	0	0
40	1	1	0,5	0	0	1	0,113	0	0	0,107	0	0

## 2.2. KAPLAN-MEIER

Los estimados de las funciones de supervivencia que se obtienen con el procedimiento de las tablas de vida dependen del agrupamiento que se realice. Una solución a este problema consiste en intentar computar las funciones sin tener que obtener previamente las tablas de vida con agrupamientos. Para ello podemos proceder construyendo una tabla en la que el intervalo sea de tal tamaño que contenga solamente a un sujeto. Las funciones pueden derivarse multiplicando las probabilidades de supervivencia de los sujetos en los cuales ocurre el evento de interés. En otras palabras, las tablas de vida y Kaplan-Meier son equivalentes cuando cada intervalo en el primer análisis contiene solamente un sujeto.

En SPSS, la secuencia de comandos necesaria para realizar el análisis es muy semejante al caso anterior, concretamente tendremos:

Análisis → Kaplan-Meier → Tiempo: DÍAS → Estado: ALTA → Definir evento → Valor único: 1 → Continuar → Factor: TRATAMIENTO → Opciones → Tablas de mortalidad → Gráfico: Supervivencia → Continuar → Aceptar

Los resultados ofrecidos por el paquete se interpretan de manera similar a como hemos mencionado respecto de las tablas de vida.

### 2.3. REGRESIÓN DE COX

En ocasiones el investigador mide una o más variables de las cuales supone que pueden explicar gran parte de la aparición de los eventos de interés, esto es, que pueden explicar la supervivencia o la mortalidad. En principio sería posible aplicar técnicas más estándares para examinar si efectivamente esa variable dependiente puede ser explicada por los supuestos predictores. Sin embargo, esos análisis (como el de regresión lineal, véase Capítulo VIII), asumen que las variables se distribuyen normalmente, lo cual es difícil que ocurra en el caso de la aparición de eventos. Pero, además, no se tiene en cuenta la diferencia entre abandonos y mortalidad (altas en nuestro ejemplo). En este sentido, el modelo de azar de Cox tiene las ventajas de los modelos de regresión, y además carece de supuestos sobre los datos. Es interesante notar que en la regresión de Cox pueden emplearse predictores continuos o categóricos (por ejemplo, el tratamiento aplicado es un predictor categórico). Matemáticamente se trata de estimar la función de azar dados los distintos covariados a partir de la función de azar cuando todos los covariados son cero. Más concretamente:

$$H_{(t,x_1,x_2,\dots,x_p)} = H_{0(t)} e^{b_1x_1 + b_2x_2 + \dots + b_px_p}, \quad (\text{Ec. 7.1})$$

Esta función exponencial puede linearizarse fácilmente:

$$\ln \left( \frac{H_{(t,x_1,x_2,\dots,x_p)}}{H_{0(t)}} \right) = b_1x_1 + b_2x_2 + \dots + b_px_p \quad (\text{Ec. 7.2})$$

lo que permite realizar una estimación sencilla de los coeficientes de regresión. La secuencia de comandos para el análisis, aplicado a nuestro ejemplo, mediante SPSS es la siguiente:

```
Análisis → Regresión de Cox → Tiempo: DÍAS → Estado: ALTA →
Definir evento → Valor único: 1 → Continuar → Covariables: TRATAMIENTO →
Categórica → Covariables categóricas: TRATAMIENTO →
Continuar → Opciones → Tablas de mortalidad → Gráfico: Supervivencia →
Continuar → Aceptar
```

La evaluación del modelo de regresión se realiza mediante una chi-cuadrado (véase la salida de SPSS en el Cuadro 7.3., página siguiente), mientras que se emplea el test de Wald para evaluar la significación de los coeficientes de regresión para cada uno de los predictores (véase Cuadro 7.3.).

### Análisis de supervivencia

Cuadro 7.3. Regresión de Cox considerando el tratamiento como un predictor categórico

	Chi-Square	df	Sig				
Overall (score)	8,811	1	,0030				
Change (-2LL) from Previous Block	7,709	1	,0055				
Previous Step	7,709	1	,0055				
Variables in the Equation	B	S.E.	Wald	df	Sig	R	Exp(B)
TRATAMIE	-1,5133	,5504	7,5611	1	,0060	-,2657	,2202

### 3. LIMITACIONES Y SUPUESTOS

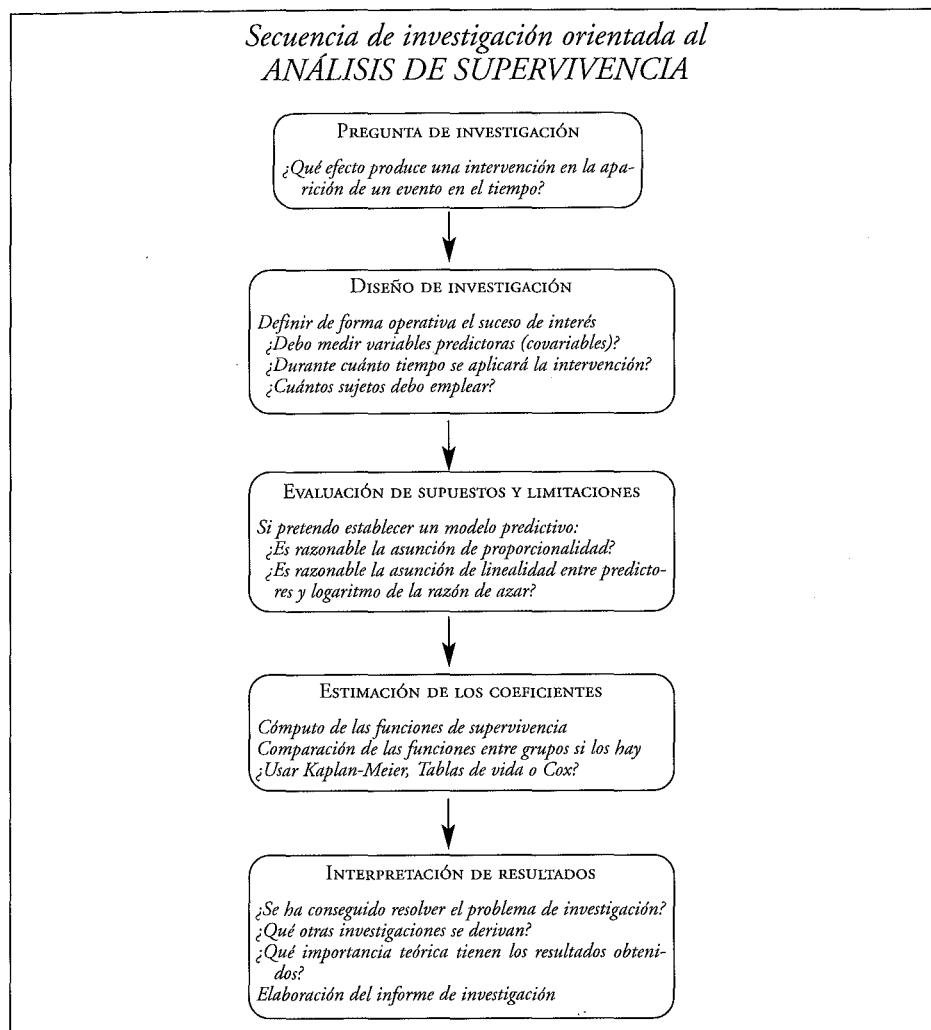
Como hemos podido comprobar, el análisis de supervivencia está, en relación con otras técnicas de análisis, relativamente libre de supuestos. Concretamente, no se hacen asunciones sobre las características de distribución de los datos, ni son relevantes la existencia de puntos extremos o de un reducido número de sujetos. Sin embargo, en el caso de la regresión de Cox, sí es importante tener presente dos supuestos, no relativos a los datos, sino a las relaciones entre los predictores y la función de azar. En primer lugar, se asume la «proporcionalidad», esto es, que si dos sujetos tienen valores diferentes, la razón entre sus funciones de azar es independiente del tiempo. El segundo supuesto implica, como puede comprobarse en la ecuación 7.2., que la relación entre el logaritmo de la razón de azar y los predictores es lineal. Por supuesto el no cumplimiento de estos dos supuestos hace el análisis inviable.

### 4. LA SECUENCIA DE INVESTIGACIÓN ORIENTADA AL ANÁLISIS DE SUPERVIVENCIA

Excepto en el caso de la regresión de Cox el análisis de supervivencia puede considerarse una técnica exploratoria, de reducción de datos. Como en el resto de este tipo de técnicas, el diseño de investigaciones orientadas a su uso no implica un alto grado de esfuerzo en el diseño. En este sentido, el establecimiento de una pregunta de investigación interesante desde un punto de vista teórico (o aplicado) es fundamental para que la investigación gane en calidad y alcance. El tipo de preguntas que pueden formularse en este contexto, aunque pueda parecer limitado a primera vista, es bastante amplio (véase Cuadro 7.4.).

En realidad, la única limitación importante que impone el análisis en ese sentido es que es preciso poder definir un suceso relevante que pueda aparecer en función del tiempo.

Cuadro 7.4. La secuencia de investigación orientada al análisis de supervivencia



Esta característica hace que la técnica sea interesante en muchos problemas relativos a la evaluación de calidad de un tratamiento, de una intervención, o de una mera campaña. La cuestión fundamental, por tanto, estriba en poder definir de manera clara un suceso de interés que pueda aparecer en función del tiempo y/o de variables que pueden ser manipuladas (como un terapia) o simplemente medidas (como la frecuencia con que se emiten ciertos anuncios en medios de comunicación). Además de medir si el suceso se produce y cuándo en cada uno de los sujetos de la muestra es preciso, cuando se han incluido variables predictoras en el diseño caer en la cuenta de que la calidad del análisis dependerá básicamente del tipo de relación entre ellas y el logaritmo de la

razón de azar. En el caso de que esas variables se incluyan en la investigación, la elección de la técnica analítica es relativamente clara, la opción razonable es usar la regresión de Cox. Sin embargo, debe caerse en la cuenta de que cuando tenemos un solo predictor categórico, pueden emplearse sin ningún menoscabo análisis de tablas de vida o de Kaplan-Meier. Recuérdese en este contexto que estas dos últimas técnicas son iguales si el intervalo temporal se define de modo que tengamos un solo individuo en cada uno de ellos.

## 5. LECTURAS RECOMENDADAS

- BISQUERRA, R., *Introducción conceptual al análisis multivariable*, Barcelona, PPU, 1991.  
 KLEINBAUM, D.G., *Survival Analysis: A Self-Learning Text (Statistics in the Health Sciences)*, Nueva York, Springer-Verlag, 1996.  
 RIVAS, M.ª J. y LÓPEZ, J., *Ánalisis de supervivencia*, Madrid, La Muralla, 2000.

## LECTURAS COMPLEMENTARIAS

- KLEIN, P. J. y MOECHBERGER, M. L., *Survival Analysis*, Springer, 1997.  
 HOSMER, D. W. y LEMESHOW, S., *Applied Survival Analysis: Regression Modeling of Time to Event Data*, Nueva York, Wiley, 1999.  
 PARMAR, M. K. B. y MACHIN, D., *Survival analysis: a practical approach*, Chichester, Wiley, 1995.

## 6. EJERCICIOS

Use SPSS para estimar la función de supervivencia para la muestra de 20 desempleados de una oficina del INEM usando tablas de vida y Kaplan-Meier. La primera columna contiene los días registrados en la oficina, la segunda si encontraron empleo (1) o no (0) y la tercera si tenían Formación Profesional (1) o no (0).

DÍAS	ESTADO	FP	DÍAS	ESTADO	FP
72	1	1	82	1	0
411	1	1	10	1	0
228	1	1	118	1	0
231	0	1	126	1	0
242	1	1	8	1	0
991	1	1	92	1	0
111	1	1	35	1	0
1	1	1	117	0	0
587	1	1	132	0	0

DÍAS	ESTADO	FP	DÍAS	ESTADO	FP
389	0	1	12	0	0
33	0	0	162	1	0
25	1	0	3	1	0
357	1	0	95	1	0
467	1	0	24	1	1
201	1	0	18	1	1
1	1	1	83	1	1
30	1	1	31	0	0
44	0	0	51	0	1
283	1	0	90	1	0
15	1	0	52	1	0
87	1	1	73	1	1
112	0	1	8	1	1
999	1	1	36	1	1
11	0	1	48	1	1
25	0	1	7	1	1
144	1	1	140	0	0
8	1	0	186	1	0
42	1	0	84	1	0
100	0	1	19	1	1
314	1	1	45	1	1
110	1	1	80	0	1

1. Calcule la función de supervivencia usando tablas de vida.
2. Calcule la función de supervivencia usando Kaplan-Meier.
3. Calcule las funciones de supervivencia teniendo en cuenta si estas personas tienen FP o no.
4. Use la regresión de Cox para determinar si el estado es predecible a partir de la FP, teniendo en cuenta el tiempo.

## RESPUESTAS A LOS EJERCICIOS

1. Tabla de mortalidad para toda la muestra (usando períodos de 100 días):

Life Table  
Survival Variable DIAS

Intrvl Start Time	Entrng ntrvl	Number Wdrawn During Intrvl	Number Exposd to Risk	Number of Events	Number Propn Termnl	Propn Terminat- ing	Cumul Propn Surv at End	Prob- ability Densty	Hazard Rate
					Propn Sur- viving	Propn Sur- viving			
,0		62,0	8,0	58,0	29,0	,5000	,5000	,5000	,0050 ,0067
100,0		25,0	5,0	22,5	7,0	,3111	,6889	,3444	,0016 ,0037
200,0		13,0	1,0	12,5	4,0	,3200	,6800	,2342	,0011 ,0038
300,0		8,0	1,0	7,5	2,0	,2667	,7333	,1718	,0006 ,0031
400,0		5,0	,0	5,0	2,0	,4000	,6000	,1031	,0007 ,0050
500,0		3,0	,0	3,0	1,0	,3333	,6667	,0687	,0003 ,0040
600,0		2,0	,0	2,0	,0	,0000	1,0000	,0687	,0000 ,0000
700,0		2,0	,0	2,0	,0	,0000	1,0000	,0687	,0000 ,0000
800,0		2,0	,0	2,0	,0	,0000	1,0000	,0687	,0000 ,0000
900,0		2,0	,0	2,0	2,0	1,0000	,0000	,0000	,0007 ,0200

The median survival time for these data is 100 ,00

Intrvl Start Time	SE of Cumul Sur- viving	SE of Proba- bility	SE of Hazard Rate	SE of
,0	,0657	,0007	,0012	
100,0	,0665	,0005	,0014	
200,0	,0641	,0005	,0019	
300,0	,0603	,0004	,0021	
400,0	,0522	,0004	,0034	
500,0	,0447	,0003	,0039	
600,0	,0447	,0000	,0000	
700,0	,0447	,0000	,0000	
800,0	,0447	,0000	,0000	
900,0	,0000	,0004	,0000	

## 2. Kaplan-Meier

Time	Status	Cumulative Survival	Standard Error	Cumulative Events	Number Remaining
1,00	1,00			1	61
1,00	1,00	,9677	,0224	2	60
3,00	1,00	,9516	,0273	3	59
7,00	1,00	,9355	,0312	4	58
8,00	1,00			5	57
8,00	1,00			6	56
8,00	1,00	,8871	,0402	7	55
10,00	1,00	,8710	,0426	8	54
11,00	,00			8	53
12,00	,00			8	52
15,00	1,00	,8542	,0449	9	51
18,00	1,00	,8375	,0471	10	50
19,00	1,00	,8207	,0490	11	49
24,00	1,00	,8040	,0508	12	48
25,00	1,00	,7872	,0524	13	47
25,00	,00			13	46
30,00	1,00	,7701	,0540	14	45
31,00	,00			14	44
33,00	,00			14	43
35,00	1,00	,7522	,0556	15	42
36,00	1,00	,7343	,0571	16	41
42,00	1,00	,7164	,0585	17	40
44,00	,00			17	39
45,00	1,00	,6980	,0598	18	38
48,00	1,00	,6796	,0610	19	37
51,00	,00			19	36
52,00	1,00	,6608	,0621	20	35
72,00	1,00	,6419	,0632	21	34
73,00	1,00	,6230	,0641	22	33
80,00	,00			22	32
82,00	1,00	,6035	,0650	23	31
83,00	1,00	,5841	,0657	24	30
84,00	1,00	,5646	,0663	25	29
87,00	1,00	,5451	,0668	26	28
90,00	1,00	,5257	,0672	27	27
92,00	1,00	,5062	,0675	28	26
95,00	1,00	,4867	,0677	29	25
100,00	,00			29	24
110,00	1,00	,4664	,0678	30	23

## Análisis de supervivencia

Time	Status	Cumulative Survival	Standard Error	Cumulative Events	Number Remaining
------	--------	---------------------	----------------	-------------------	------------------

111,00	1,00	,4462	,0678	31	22
112,00	,00			31	21
117,00	,00			31	20
118,00	1,00	,4239	,0680	32	19
126,00	1,00	,4015	,0680	33	18
132,00	,00			33	17
140,00	,00			33	16
144,00	1,00	,3764	,0682	34	15
162,00	1,00	,3514	,0681	35	14
186,00	1,00	,3263	,0677	36	13
201,00	1,00	,3012	,0670	37	12
228,00	1,00	,2761	,0660	38	11
231,00	,00			38	10
242,00	1,00	,2485	,0649	39	9
283,00	1,00	,2209	,0633	40	8
314,00	1,00	,1932	,0611	41	7
357,00	1,00	,1656	,0583	42	6
389,00	,00			42	5
411,00	1,00	,1325	,0552	43	4
467,00	1,00	,0994	,0504	44	3
587,00	1,00	,0663	,0431	45	2
991,00	1,00	,0331	,0318	46	1
999,00	1,00	,0000	,0000	47	0

Number of Cases: 62 Censored: 15 (24,19%) Events: 47

Survival Time Standard Error 95% Confidence Interval

Mean: 202,84 39,72 (125,00; 280,69)  
Median: 95,00 15,42 ( 64,77; 125,23)

## 3. Introduzca FP como factor

## 4. Regresión de Cox:

Events Censored

47 15 (24,2%)

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 294,941

Beginning Block Number 1. Method: Enter  
 Variable(s) Entered at Step Number 1.  
 FP

Coefficients converged after 3 iterations.

-2 Log Likelihood 294,038

	Chi-Square	df	Sig
Overall (score)	,912	1	,3396
Change (-2LL) from Previous Block	,904	1	,3418
Previous Step	,904	1	,3418

Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
FP	-,2913	,3061	,9059	1	,3412	,0000	,7473

Covariate Means

Variable	Mean
FP	,5323

## CAPÍTULO VIII

### Análisis de regresión lineal múltiple

#### OBJETIVOS DE APRENDIZAJE

- 1) Definir la regresión múltiple y distinguirla de otras técnicas multivariadas semejantes como análisis discriminante o regresión logística.
- 2) Discriminar los contextos en los que la regresión múltiple es adecuada para resolver el problema de investigación planteado.
- 3) ¿Cómo usar el análisis de regresión para contrastar hipótesis sobre la predicción de una variable dependiente (criterio) a partir de un conjunto de independientes (predictores)?
- 4) Conocer cómo usar variables ficticias y cómo interpretar los resultados.
- 5) Conocer las condiciones de aplicación del análisis.
- 6) Distinguir entre el análisis por etapas (*stepwise*) y el análisis simultáneo.

En este capítulo encontrará cómo usar el análisis de regresión para contrastar hipótesis relativas a la predicción de una variable dependiente o criterio a partir de un conjunto de variables independientes o predictoras. Se trata de una técnica cuya aplicabilidad es tremadamente amplia, puesto que puede aplicarse, dentro del ámbito de investigación en Psicología, en contextos muy diferentes, incluyendo en la práctica todas las áreas y casi todos los problemas de investigación. Como veremos, el análisis de regresión múltiple puede utilizarse siempre que, como mínimo, nuestras variables, predictoras y criterio sean

métricas. Esto se debe a las condiciones de aplicación específicas del análisis que veremos más adelante. Sin embargo, también puede aplicarse cuando las variables no son métricas, pero se realiza una transformación para convertirlas en métricas (codificación ficticia). Finalmente, se introducen sugerencias para la interpretación de los resultados. La exposición que realizaremos es, como en el resto de capítulos de este Manual de carácter intuitivo, procurando más que el lector comprenda el contenido que el detalle técnico. Desde nuestro punto de vista lo relevante en sí no es el detalle de la ejecución del análisis, algo que los paquetes estadísticos realizan a la perfección, cuanto que el lector esté en condiciones de controlar el flujo de procesamiento de esos paquetes.

## 1. APROXIMACIÓN CONCEPTUAL AL ANÁLISIS DE REGRESIÓN MÚLTIPLE

La regresión múltiple es una técnica que puede emplearse cuando el objetivo de la investigación es predecir una única variable criterio (dependiente) a partir de un conjunto de variables predictoras (independientes). Todas las variables han sido medidas en la investigación, y son métricas. La predicción se realiza combinando linealmente las predictoras. Recordemos que la combinación lineal es una suma ponderada de variables. Cada variable predictora tiene un valor (peso) diferente para hacer la predicción. Por ejemplo, podemos medir el deterioro en la realización de las actividades de la vida cotidiana en un grupo de personas mayores (mediante el inventario de actividades que computase la cantidad global de actividades que las personas realizan de forma autónoma), y tratar de determinar si el estado de la memoria operativa, de la atención y de la memoria prospectiva (la realización demorada de actividades planificadas previamente) puede predecir el deterioro en las actividades cotidianas. Naturalmente tendríamos que utilizar tareas para medir también las funciones cognitivas, por ejemplo, podríamos usar el test de ejecución continua (CPT) para medir la vigilancia atencional, en el que el número de errores de omisión (no detección de objetivos presentados) podría ser el indicador de la vigilancia.

Nótese que el ejemplo incluye las características fundamentales de una investigación cuyos datos pueden analizarse de forma idónea mediante regresión múltiple. En primer lugar, las variables son métricas y varían arbitrariamente, es decir, no son manipuladas, sino simplemente medidas. Aunque no podemos utilizar la regresión múltiple cuando la variable dependiente es no métrica (en este caso es preferible utilizar regresión logística o análisis discriminante), sí es posible emplearla cuando los predictores son no métricos, creando variables ficticias. En segundo lugar, claramente podemos definir una variable criterio, y un conjunto de predictores. Esta definición es *a priori* y viene ya especificada en el momento de establecer el problema de investigación: ¿Depende el deterioro de las personas mayores en la realización de actividades de la vida diaria de su estado cognitivo (atención y memoria)? En tercer

lugar, cabe sospechar que hay alguna relación entre cada predictor y el criterio, y también entre los predictores entre sí. Esto último es importante, porque implica que la regresión es aplicable cuando, como ocurre en investigaciones experimentales con un número desigual de sujetos por grupo, los diseños son no balanceados (los independientes correlacionan entre sí).

## 2. REGRESIÓN LINEAL SIMPLE

La regresión múltiple es una complejización de la regresión simple, en la que solamente hay dos variables medidas, un criterio y un predictor. En este caso, es bien conocido que la ecuación de predicción tiene la forma siguiente:

$$Y' = a + bX \quad (Ec. 8.1)$$

donde  $a$  y  $b$  son coeficientes,  $X$  es la independiente e  $Y'$  es la predicción. Recordemos que  $a$  es la predicción que podemos hacer si no hubiera variable independiente, esto es, es el punto de corte en la ordenada, el valor que adopta  $Y'$  cuando  $X$  es cero. Por su parte  $b$  es la tasa de cambio de  $Y'$  respecto de  $X$ , el coeficiente de regresión. Como se sabe, una tasa de cambio indica cuánto cambia una variable cuando la otra cambia en una unidad. Así, si  $b=0,8$ , lo que sabemos es que por cada unidad que cambia  $X$ ,  $Y'$  cambia en 0,8 unidades. Este coeficiente de regresión contiene dos elementos, uno relativo a la correlación entre  $X$  e  $Y$  y el otro relativo a la escala de las variables. Consideremos el siguiente conjunto de datos. Supongamos que hemos medido las actividades de la vida cotidiana ( $Y$ ) y la vigilancia ( $X$ ):

Y	X
4	30
6	70
5	60
2	20
7	80

Realizando algunos cálculos (véase Cuadro 8.1.), tendríamos que el coeficiente de regresión no estandarizado (usando las variables en la escala que hemos empleado para medirlas) sería  $b=0,0716$ , y que el punto de corte en la ordenada sería  $a=1,075$ . Sin embargo, si los datos hubiesen sido los siguientes:

Y	X
4	3
6	7
5	6
2	2
7	8

entonces  $b=0,716$  y  $a=1,075$ . Nótese que la escala de X ha sido dividida por 10 y que el coeficiente b es el anterior multiplicado por el cambio de escala (10). Estandarizando las variables, conseguimos suprimir la influencia de la escala sobre el coeficiente de regresión, lo que nos permite una mejor interpretación de los resultados, puesto que ahora éste contiene información relativa solamente a su capacidad para predecir Y.

Cuadro 8.1. Estimación de los coeficientes mediante mínimos cuadrados: coeficiente de regresión (b) y punto de corte en la ordenada (a). El símbolo S es la suma de la puntuaciones de los distintos sujetos. Para hacer los cálculos debemos elevar al cuadrado los valores del predictor (columna X2) y también multiplicar los pares XY. SPXY es la suma de los productos de las variables. SCX es la suma de los cuadrados de X

Sujetos	Y	X	X <sup>2</sup>	XY
1	4	3	9	12
2	6	7	49	42
3	5	6	36	30
4	2	2	4	4
5	7	8	64	56
Suma	24	26	162	144

$\Sigma P_{XY} = \Sigma XY - (\Sigma X \Sigma Y)/n = 144 - (24 \cdot 26)/5 = 19,2$   
 $SC_X = \Sigma X^2 - (\Sigma X)^2/n = 162 - (26^2/5) = 26,8$   
 $b = SP_{XY}/SC_X = 19,2/26,8 = 0,716$   
 $a = \Sigma Y/n - b(\Sigma X/n) = 24/5 - 0,716(26/5) = 1,075$

Para estandarizar b, obtenemos Beta = b(S<sub>X</sub>/S<sub>Y</sub>), donde S<sub>X</sub> y S<sub>Y</sub> son desviaciones típicas de X e Y

### 3. REGRESIÓN LINEAL MÚLTIPLE

La regresión múltiple es una extensión de la regresión simple al caso en que tenemos más de un predictor. Así, la ecuación de predicción que habremos de estimar es un poco más compleja, pero tiene un significado semejante. Suponiendo p predictores ( $X_j$ ), tendremos:

$$Y' = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \dots + b_p X_p \quad (\text{Ec. 8.2})$$

donde Y' es la predicción que hacemos de la variable criterio, a es la predicción en el caso en que no tuviésemos predictores (los valores de los X fuesen 0), y los  $b_j$  son los coeficientes de regresión de cada variable predictora. Supongamos que además de medir la vigilancia ( $X_1$ ), hubiésemos medido la memoria operativa ( $X_2$ ) y la memoria prospectiva ( $X_3$ ). Tendríamos entonces 3 predictores y un criterio. Si los datos fuesen los siguientes:

Sujetos	Y	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>
1	4	3	2	7
2	6	7	4	11
3	5	6	3	9
4	2	2	1	4
5	7	8	5	12
Suma	24	26	15	43

el análisis requeriría ahora estimar tres coeficientes de regresión mediante mínimos cuadrados, y un punto de corte en la ordenada. Los cálculos aparecen en el Cuadro 8.2. (página siguiente). Obsérvese que el procedimiento es similar al que hemos empleado para el análisis de regresión simple, pero ahora tendremos en cuenta todas las variables simultáneamente, lo que implica que tendremos que operar con matrices, el vector columna de la variable dependiente (Y) y la matriz de n (sujetos/filas) x p (predictores/columnas) (X). Los coeficientes se obtienen teniendo en cuenta que hay que minimizar los cuadrados de los errores de predicción. La ecuación matricial a resolver es  $b=(X'X)^{-1}(X'Y)$ , donde b contiene los pesos de cada variable, y X' es la transpuesta de X (las filas se cambian por las columnas). Los coeficientes serán, pues:  $a=-0,1875$ ,  $b_1=-0,1875$ ,  $b_2=0,375$ , y  $b_3=0,5625$ , de manera que nuestra ecuación de predicción será:

$$Y' = -0,1875 - 0,1875 X_1 + 0,375 X_2 + 0,5625 X_3,$$

la puntuación que predecimos para el sujeto 1, por ejemplo, será:

$$3,9375 = -0,1875 - 0,1875 \times 3 + 0,375 \times 2 + 0,5625 \times 7,$$

y el error de predicción  $e_1 = 4 - 3,9375 = 0,0625$ .

Estos cálculos pueden realizarse mediante herramientas como la de análisis de datos de Excel (*Menú Herramientas → Análisis de datos → Regresión*), o mediante paquetes estadísticos como SPSS (*Analizar → Regresión → Regresión Lineal*). La salida de SPSS aparece en el Cuadro 8.3. (véase página siguiente). Obsérvese que la salida de SPSS nos ofrece además de los coeficientes, su error estándar, los coeficientes estandarizados (que hubiésemos obtenido, como hemos visto antes, si estandarizamos los predictores sustrayendo a cada puntuación su media y dividiendo por la desviación típica correspondiente), y un contraste mediante t de Student para determinar si los coeficientes son significativamente diferentes de cero. Cuando el número de sujetos es tan reducido como el de este ejemplo ficticio es frecuente encontrar que aunque la correlación de los predictores con el criterio sea muy alta, los coeficientes no son significativos. Este punto será importante más adelante, para interpretar el significado de los coeficientes en las combinaciones lineales.



Cuadro 8.2. Cálculo de los coeficientes de regresión y punto de corte en la ordenada

Sujeto	Y	X1	X2	X3
1	4	3	2	7
2	6	7	4	11
3	5	6	3	9
4	2	2	1	4
5	7	8	5	12
Media	4,8	5,2	3	8,6

La ecuación a resolver es:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Primero obtenemos las puntuaciones diferenciales sustrayendo la media correspondiente  
 $x_j = X_j - \bar{X}_j$

$y = Y - \bar{Y}$	$-0,8$	$-2,2$	$-1$	$-1,6$
	$1,2$	$1,8$	$1$	$2,4$
	$0,2$	$0,8$	$0$	$0,4$
	$-2,8$	$-3,2$	$-2$	$-4,6$
	$2,2$	$2,8$	$2$	$3,4$

Matriz  $\mathbf{Y}$       Matriz  $\mathbf{X}$       Transpuesta de  $\mathbf{X}$ :  $\mathbf{X}'$

Después multiplicamos la traspuesta de  $\mathbf{X}$  por  $\mathbf{X}$  ( $\mathbf{X}'\mathbf{X}$ ) e invertimos la matriz, para obtener  $(\mathbf{X}'\mathbf{X})^{-1}$

$26,8$	$16$	$32,4$	$0,938$	$-0,88$	$-0,31$
$16$	$10$	$20$	$-0,88$	$4,25$	$-1,38$
$32,4$	$20$	$41,2$	$-0,31$	$-1,38$	$0,937$

Matriz  $\mathbf{X}'\mathbf{X}$       Inversa de  $\mathbf{X}'\mathbf{X}$ :  $(\mathbf{X}'\mathbf{X})^{-1}$

A continuación multiplicamos la traspuesta de  $\mathbf{X}$  ( $\mathbf{X}'$ ) por  $\mathbf{Y}$ , para obtener  $\mathbf{X}'\mathbf{Y}$ , y luego multiplicamos  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

$19,2$	$12$	$24,6$	$b_1 = -0,19$
			$b_2 = 0,375$
			$b_3 = 0,562$

Matriz  $\mathbf{X}'\mathbf{Y}$        $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

Finalmente, calculamos a como  $Y - \text{suma } bX$ :  $a = Y - (-0,1875 * X_1 + 0,375 * X_2 + 0,5625 * X_3) = -0,19$

Cuadro 8.3. Resumen de resultados ofrecido por SPSS (versión 9.0). Los coeficientes Beta son los de regresión cuando se estandarizan los predictores. Ningún coeficiente de regresión es significativamente diferente de cero. La t es de Student. La constante es a, el punto de corte en la ordenada. Por tanto, los predictores no son adecuados para predecir el criterio. Esto puede deberse al reducido número de sujetos empleado

ANALIZAR → REGRESIÓN → LINEAL → DEPENDIENTE: Y, INDEPENDIENTES: X <sub>1</sub> , X <sub>2</sub> y X <sub>3</sub>						
		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
Modelo		B	Error tip.	Beta		
1	(Constante)	-,188	,658	,252	-,285	,823
	X1	-,187	,242	,515	-,775	,580
	X2	,375	,308	,728	,728	,600
	X3	,562	,939	,2324	,2324	,259

## 3.1. CONTRASTE DE HIPÓTESIS EN LA REGRESIÓN MÚLTIPLE

¿Cómo podemos comprobar si la predicción del criterio desde los predictores es significativa? Esto es, incrementamos de forma fiable nuestra capacidad de predecir la variable dependiente usando las variables predictoras. El test para realizar el contraste lo podemos construir de una manera muy sencilla si tenemos en cuenta que lo relevante es nuestro acierto y nuestro error en la predicción. Si el acierto es suficientemente grande respecto del error, la predicción será significativa. La cuestión reside, pues, en encontrar medidas del acierto y del error. Pero eso también resulta sencillo. Basta con caer en la cuenta de que los sujetos son diferentes en el criterio, o, dicho de otro modo, que el criterio tiene una variabilidad. Llamémosla variabilidad total. Esa variabilidad la podemos explicar en parte desde los predictores (llamémosla variabilidad de regresión), pero en parte no. La variabilidad no explicada por los predictores es error. Ésto es:

$$\text{Variabilidad TOTAL} = \text{Variabilidad de REGRESIÓN} + \text{Variabilidad ERROR}$$

La variabilidad se puede medir fácilmente, en general una función de las distancias entre las puntuaciones y un punto de referencia, que suele ser la media. Así, la variabilidad total es la suma de los cuadrados de las distancias en el criterio entre los sujetos y la media del criterio ( $SC_T$ ). Teniendo en cuenta que el error de predicción es la diferencia entre la puntuación real y la predicha, la variabilidad de error o residual la definiremos como la suma de los cuadrados de los errores de predicción ( $SC_{RES}$ ). Finalmente, lo que queda, la suma de las distancias entre las puntuaciones predichas y la media del criterio, será la variabilidad de regresión ( $SC_{REG}$ ). De manera más formal, tendremos que:

$$SC_T = SC_{REG} + SC_{RES} \quad (\text{Ec. 8.3})$$

que calcularemos según las expresiones siguientes:

$$SC_T = \sum(Y - \bar{Y})^2 \quad (\text{Ec. 8.4})$$

$$SC_{REG} = \sum(Y' - \bar{Y})^2 \quad (\text{Ec. 8.5})$$

$$y \quad SC_{RES} = \sum(Y - Y')^2 \quad (\text{Ec. 8.6})$$

Ahora podemos saber la proporción de variabilidad total que es explicada por los predictores. Si dividimos variabilidad de regresión por la total tendremos:

$$R^2 = \frac{SC_{REG}}{SC_T} \quad (\text{Ec. 8.7})$$

donde  $R^2$  es el cuadrado del coeficiente de correlación múltiple (del conjunto de predictores con el criterio). Antes de continuar es importante tener presente que  $R^2$  depende del número de variables y del número de sujetos. De hecho, su valor es sobreestimado cuando el número de ambos es bajo. El ajuste de  $R^2$  puede realizarse mediante la siguiente expresión, siendo  $n$  el número de sujetos y  $p$  el de predictores:

$$R^2 \text{ ajustada} = 1 - (1 - R^2) \left( \frac{n - 1}{n - p - 1} \right) \quad (\text{Ec. 8.8})$$

Por ejemplo, si  $R^2$  es 0,8, con 5 sujetos y 3 independientes, su valor ajustado será:  $R^2$  ajustada =  $1 - (1 - 0,8)(4/1) = 0,2$ . Por el contrario, con dos predictores, su valor sería:  $R^2$  ajustada =  $1 - (1 - 0,8)(4/2) = 0,6$ . Cohen y Cohen (1983) recomiendan que el valor de  $p$  sea el de variables incluidas en la ecuación, y no el número total de predictores medidos.

El resto de pasos para tomar la decisión sobre la fiabilidad de la predicción consisten en convertir las variabilidades en varianzas (dividiendo por los grados de libertad respectivos) y en dividir la varianza de regresión por la de error. Los grados de libertad totales ( $gl_T$ ) son el número de observaciones menos 1. Si tenemos  $n$  sujetos,  $gl_T = n - 1$ . Los grados de libertad de la regresión son el número de predictores ( $p$ ), esto es,  $gl_{REG} = p$ . Por tanto, los residuales serán:  $gl_{RES} = (n - 1) - p$ . En nuestro ejemplo,  $gl_T = 5 - 1 = 4$ ,  $gl_{REG} = 3$ , y  $gl_{RES} = (5 - 1) - 3 = 1$ .

Las varianzas o medias de cuadrados son un cociente entre sumas de cuadrados y grados de libertad. Nos interesan sólo dos, la de regresión y la residual. Por tanto, tendremos que:

$$MC_{REG} = \frac{SC_{REG}}{gl_{REG}} \quad (\text{Ec. 8.9})$$

$$y \quad MC_{RES} = \frac{SC_{RES}}{gl_{RES}} \quad (\text{Ec. 8.10})$$

Ahora, obtendremos el estadístico  $F$  para la regresión, que es el cociente entre la media de cuadrados de regresión y la residual:

$$F_{REG} = \frac{MC_{REG}}{MC_{RES}} \quad (\text{Ec. 8.11})$$

que sigue una distribución teórica de  $F$  con  $p$  y  $(n - p - 1)$  grados de libertad en numerador y denominador, respectivamente. Éste es, como el lector habrá advertido ya, un análisis de varianza sobre la regresión.

Cuadro 8.4. Coeficiente de correlación múltiple, tabla superior, y análisis de varianza para la regresión (tabla inferior). La  $R$  cuadrado corregida tiene en cuenta el número de variables y el de sujetos, que tienden a incrementar artificialmente el valor de  $R$  cuadrado

ANALIZAR → REGRESIÓN → LINEAL → DEPENDIENTE: Y, INDEPENDIENTES:  $X_1$ ,  $X_2$  y  $X_3$

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,998	,996	,983	,25000000

#### ANOVA

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	14,737	3	4,913	78,600	,083
	Residual	0,0625	1	0,0625		
	Total	14,800	4			

Este conjunto de análisis puede hacerse mediante SPSS con la secuencia de comandos que hemos visto más arriba (Analizar → Regresión → Lineal). Los resultados para el ejemplo que venimos considerando aparecen en el Cuadro 8.4. (arriba). Como era de esperar a partir de los contrastes sobre los coeficientes de regresión de cada predictor, la predicción no es significativa, aunque puede considerarse marginal.

#### 3.2. VARIEDADES DEL ANÁLISIS DE REGRESIÓN MÚLTIPLE

Tener más de una variable predictora implica una decisión importante sobre cómo se va a comprobar la aportación de cada variable a la predicción. Es importante, caer en la cuenta de que los resultados pueden cambiar de forma notable si alguno de los predictores no está presente. Podemos ver claramente en los ejemplos precedentes cómo los coeficientes (y la predicción) han cambiado cuando hemos introducido tres predictores en lugar de solamente uno. La razón es sencilla, cuando tenemos más de un predictor consideramos las correlaciones entre ellos. Si la correlación entre dos predictores es alta, su predicción conjunta es redundante, por tanto, uno de ellos aporta poco si el otro ya está dentro de la ecuación de predicción. La cuestión importante es, pues, cómo vamos a introducir los predictores en la ecuación de predicción. Hay tres formas diferentes: simultánea, secuencial o jerárquica y por etapas. Se diferencian en el modo en que cada predictor es introducido (o eliminado) de la ecuación.

### 3.2.1. Regresión simultánea

Este método se conoce también como estándar. Se caracteriza porque todas las variables independientes son introducidas a la vez en la ecuación de predicción. Esto implica que cada una de ellas es evaluada en cuanto a lo que aporta para mejorar la predicción suponiendo que el resto ya está presente en la ecuación. Puede ocurrir, por tanto, que una variable que tenga una alta correlación con el criterio aparente ser irrelevante para la predicción. Esto es lo que ocurre cuando la correlación entre los predictores es muy alta. En este método, todas las variables permanecen en la ecuación, con independencia de su poder de predicción.

### 3.2.2. Regresión múltiple secuencial

En el análisis de regresión secuencial o jerárquico, los predictores se introducen en la ecuación de modo progresivo de acuerdo con un criterio especificado por el investigador. Puesto que en ella ya hay algún predictor presente, lo que se evalúa es la aportación que hace la nueva variable a la predicción. Lo que hace interesante este método es que el orden en que se introducen las variables se suele fijar en función de criterios teóricos. En el ejemplo que venimos considerando en este capítulo las consideraciones teóricas nos dirían que la memoria prospectiva y la memoria operativa son más importantes para predecir la realización de actividades de la vida cotidiana que la vigilancia. Por tanto, podríamos decidir introducir primero X3 y X2 juntas y finalmente X1.

En SPSS este análisis puede realizarse mediante el método INTRODUCIR. Para ello la secuencia de comandos sería:

Analizar → Regresión → Lineal

A continuación especificamos la variable dependiente (Y) y las variables dependientes. Puesto que hemos decidido la secuencia de introducción X3/X2 → X1, emplearemos dos bloques de variables, constituidos en este caso por X3 y X2 el primero y por X1 el segundo. Así, la secuencia de comandos sería:

Dependiente: Y  
Independientes: X3, X2 → Bloque Siguiente → X1 → Analizar

Los resultados del análisis realizado por SPSS aparecen en el Cuadro 8.5. Observemos que SPSS construye dos modelos de regresión. El primero con X3 y X2 como predictores y el segundo con los tres predictores.

Cuadro 8.5. Análisis de regresión múltiple jerárquico mediante SPSS

#### Variables introducidas/eliminadas

Modelo	Variables introducidas	Variables eliminadas	Método
1	X2, X3	.	Introducir
2	X1	.	Introducir

a Todas las variables solicitadas introducidas

#### Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,997	,993	,986	,22360680
2	,998	,996	,983	,25000000

a Variables predictoras: (Constante), X2, X3

b Variables predictoras: (Constante), X2, X3, X1

#### ANOVA

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	14,700	2	7,350	147,000	,007
	Residual	,100	2	5,000E-02		
	Total	14,800	4			
2	Regresión	14,737	3	4,913	78,600	,083
	Residual	6,250E-02	1	6,250E-02		
	Total	14,800	4			

a Variables predictoras: (Constante), X2, X3

b Variables predictoras: (Constante), X2, X3, X1

#### Coeficientes

		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
Modelo		B	Error típ.	Beta		
1	(Constante)	-,100	,580		-,172	,879
	X3	,500	,204	,834	2,449	,134
	X2	,200	,414	,164	,483	,677
2	(Constante)	-,188	,658		-,285	,823
	X3	,562	,242	,939	2,324	,259
	X2	,375	,515	,308	,728	,600
	X1	-,187	,242	-,252	-,775	,580

a Variable dependiente: Y

#### Variables excluidas

	Beta dentro	t	Sig.	Correlación parcial	Estadísticos de colinealidad	
Modelo					Tolerancia	
1	X1	-,252	-,775	,580	-,612	3,980E-02

a Variables predictoras en el modelo: (Constante), X2, X3

Es interesante caer en la cuenta de que la predicción es significativa cuando incluimos X3 y X2, pero sólo lo es marginalmente cuando introducimos también X1. Parece claro en el análisis de varianza (ANOVA) que la elección más razonable es utilizar el primer modelo y excluir X1 de la ecuación de predicción. Eso es, precisamente lo que hace SPSS. Por último, obsérvese cómo los coeficientes de regresión cambian cuando X1 está ausente respecto de cuan- do está presente.

### 3.2.3. Regresión múltiple por etapas (Stepwise)

El orden en que entran las variables en la ecuación de predicción puede establecerse por criterios estadísticos en lugar de por criterios teóricos o lógicos. Desde nuestro punto de vista eso no es muy acertado, puesto que se ignora el contenido de las variables, y ello puede dificultar mucho la interpretación final de los resultados. Además, los criterios estadísticos pueden llevar aparejada una consecuencia nefasta, puesto que diferencias mínimas entre variables pueden llevar a excluir a unas en favor de otras. Si consideramos ambas cuestiones juntas, debe quedar claro para el lector que éste es un procedimiento que hay que aplicar con cierta cautela.

Los paquetes estadísticos como SPSS introducen tres variedades de este procedimiento: por pasos sucesivos, hacia delante y hacia atrás. Veremos a continuación las características de estas técnicas, pero antes vamos a introducir un par de conceptos importantes: la F para entrar y la F para salir. La F para entrar (o su probabilidad) se emplea para decidir si una variable entra en la ecuación de predicción. La estrategia consiste en calcular para la variable candidata un valor de F, teniendo en cuenta las predictoras que ya están en la ecuación. Primero se calculan los errores de predicción con las variables ya presentes en el modelo. A continuación, se evalúa cómo la candidata reduce los residuos, mediante un análisis de regresión simple en el que los residuos son la dependiente y la candidata la independiente. La F de regresión sobre los residuos que se obtiene es la F para entrar. La regla que se aplica a continuación es la siguiente: si la F (o su probabilidad) supera un criterio arbitrario fijado por el analista, la variable entra en la ecuación. Es habitual (en SPSS es la elección por defecto) usar la probabilidad del valor de F. Así, por ejemplo, si hemos decidido que la probabilidad debe ser menor o igual que 0,05, la candidata entrará en la ecuación si su probabilidad es igual o está por debajo del criterio. Cuando se introduce una nueva variable en la ecuación hay que calcular la capacidad predictiva de cada una de las que ya estaban dentro. Para ello se procede de manera similar a lo ya explicado, es decir, con cada variable dentro se intentan predecir los residuos que se obtienen usando el resto para predecir el criterio. El estadístico que se obtiene es la F para eliminar. Las reglas para eliminar es la siguiente: si la probabilidad del valor de F para salir es mayor que el criterio de probabilidad exigido para salir, la variable es eliminada de la ecuación. El uso

adecuado de las F para entrar y para salir requiere que los valores de probabilidad no sean iguales, concretamente, el valor de entrada tiene que ser menor que el de salida, de este modo se asegura que la variable que acaba de entrar no es inmediatamente eliminada de la ecuación. Por ejemplo, si fijamos un valor de entrada de 0,10 y uno de salida de 0,05, una candidata cuya probabilidad fuera 0,07 entraría en la ecuación, pero al producir la evaluación para ver si alguna debe salir, sería eliminada de la ecuación. Finalmente, si se desea introducir muchas variables en la ecuación, es aconsejable incrementar la probabilidad (por ejemplo, 0,10 en lugar de 0,05). Por el contrario, si se desea ser exigente con las variables que quedan en la ecuación, es preferible poner probabilidades de salida bajas (digamos 0,050 en lugar de 0,10).

#### 3.2.3.1. «Regresión por etapas hacia delante»

La ecuación de predicción comienza teniendo cero variables. A continuación se realiza un análisis de regresión simple con cada una de las predictoras y el criterio. La predictora que supere el criterio estadístico y tenga mayor F (o menor probabilidad de F) es la primera en entrar. A partir de aquí, en pasos sucesivos se van introduciendo variables de acuerdo con el criterio estadístico. Si una variable tiene una F para entrar mayor que el criterio (o una probabilidad menor o igual que la criterio) entra en la ecuación. El proceso continua hasta que ya no quedan más variables para entrar. Una variable que haya entrado en la ecuación no puede ser eliminada de la misma.

Cuadro 8.6. Resultados del análisis de regresión hacia delante mediante SPSS. Beta es el coeficiente de regresión estandarizado. La correlación parcial es la correlación entre la variable y el criterio teniendo en cuenta el resto de variables

Variables en la ecuación						
		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
Modelo		B	Error típ.	Beta		
1	(Constante)	-,335	,272		-1,229	,307
	X3	,597	,030	,996	19,866	,000

Variables excluidas						
		Beta dentro	t	Sig.	Correlación parcial	Estadísticos de colinealidad
Modelo						Tolerancia
1	X1	-,148	-,580	,621	-,379	4,927E-02
	X2	,164	,483	,677	,323	2,913E-02

El Cuadro 8.6. (arriba) presenta los resultados de SPSS obtenidos con los datos del ejemplo que venimos analizando en este capítulo. La secuencia de comandos empleada ha sido la siguiente:

Analizar → Regresión → Lineal → Independiente: Y → Dependientes: X1, X2, X3 → Método: hacia delante → Opciones > Probabilidad de F: Entrada: 0,05; Salida = 0,10 → Continuar → Aceptar (Análisis)

Se ha utilizado una probabilidad de F para entrar de 0,05 y de una F para salir de 0,10. Como puede observarse, la única variable que queda en el modelo es X3. El programa no ofrece información sobre los pasos, pero realmente ésta no es demasiado relevante. Lo importante es que las variables que quedan fuera tienen un coeficiente de regresión no significativo, mientras que la única que queda dentro sí lo tiene.

### 3.2.3.2. «Regresión por etapas hacia atrás»

En este caso, todas las variables son introducidas en la ecuación en la primera etapa. A continuación se calculan las F para salir, y es eliminada aquella cuya F tiene un valor de probabilidad mayor que el de probabilidad para salir. Si hay más de una variable que cumple el criterio de eliminación, se suprime la que mayor probabilidad de F tiene. Una vez que una variable queda fuera de la ecuación ya no puede ser introducida de nuevo. El Cuadro 8.7. (véase página siguiente) presenta los resultados del análisis aplicado a los datos del ejemplo de envejecimiento con el método hacia atrás. Es interesante caer en la cuenta de que los resultados coinciden con el método hacia delante (aunque no necesariamente han de coincidir) y que la predicción es significativa sólo si se emplea X3, introducir el resto de variables hace que la predicción pueda llegar a ser no significativa (lo que ocurre en el modelo 1). La secuencia de comandos empleada ha sido la siguiente:

Analizar → Regresión → Lineal → Independiente: Y → Dependientes: X1, X2, X3 → Método: hacia atrás → Opciones > Probabilidad de F: Entrada: 0,05; Salida = 0,10 → Continuar → Aceptar (Análisis)

Cuadro 8.7. Regresión hacia atrás con SPSS

ANOVA						
Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión Residual Total	14,737 6,250E-02 14,800	3 1 4	4,913 6,250E-02	78,600	,083
2	Regresión Residual Total	14,704 9,559E-02 14,800	2 2 4	7,352 4,779E-02	153,831	,006
3	Regresión Residual Total	14,688 ,112 14,800	1 3 4	14,688 3,722E-02	394,670	,000

1 Variables predictoras: (Constante), X3, X1, X2

2 Variables predictoras: (Constante), X3, X1

3 Variables predictoras: (Constante), X3

### Coefficientes

		Coefficientes no estandarizados		Coefficientes estandarizados	t	Sig.
Modelo		B	Error típ.	Beta		
1	(Constante)	-,188	,658		-,285	,823
	X1	-,187	,242	-,252	-,775	,580
	X2	,375	,515	,308	,728	,600
	X3	,562	,242	,939	2,324	,259
2	(Constante)	-,507	,429		-1,183	,358
	X1	-,110	,190	-,148	-,580	,621
	X3	,684	,153	1,141	4,456	,047
3	(Constante)	-,335	,272		-1,229	,307
	X3	,597	,030	,996	19,866	,000

### Variables excluidas

		Beta dentro	t	Sig.	Correlación parcial	Estadísticos de colinealidad
Modelo						Tolerancia
2	X2	,308	,728	,600	,588	2,353E-02
3	X2	,164	,483	,677	,323	2,913E-02
	X1	-,148	-,580	,621	-,379	4,927E-02

a) Variables predictoras en el modelo: (Constante), X3, X1

b) Variables predictoras en el modelo: (Constante), X3

### 3.2.3.3. «Regresión por etapas»

El último procedimiento que vamos a considerar es la regresión por etapas (*stepwise*). Se trata de una combinación de los dos métodos precedentes, de modo que las variables que están fuera de la ecuación pueden entrar en ella, y las que están dentro pueden ser eliminadas. Una variable candidata entra en la

ecuación si su F es mayor que la F para entrar (o su probabilidad es menor) y sale de la ecuación si su F es menor que la F para salir (o su probabilidad es mayor). Tanto en este procedimiento como en los anteriores es posible tratar las variables en bloques. Esto implica que entran o salen de la ecuación todas las variables que forman un bloque. El análisis de SPSS aplicado a los datos del ejemplo de envejecimiento produciría los resultados que aparecen resumidos en el Cuadro 8.6. Nótese que, de nuevo, el análisis coincide con los precedentes en que X3 es la única variable que debe quedar en la ecuación. Esta coincidencia entre métodos no siempre ocurre. La secuencia de comandos es la misma que en los métodos anteriores excepto en el método definido, el de pasos sucesivos.

### 3.2.3.4. «¿Qué método seleccionar?»

En muchas investigaciones en las que simplemente se está interesado en realizar predicciones y en comprobar correlaciones, el método de elección es el simultáneo. Sin embargo, es importante tener en cuenta que el análisis está orientado a objetivos. Si nuestro objetivo es la construcción de un modelo de regresión, una buena elección es el análisis por etapas (sea hacia delante, hacia atrás o por etapas). La ventaja de este método es que utiliza criterios objetivos de selección y eliminación de variables y puede permitir descartar aquéllas que no reúnan los requisitos estadísticos mínimos que hayamos fijado (para quedarse y/o para entrar en la ecuación). Esto es más interesante cuando el número de variables es muy elevado y no tenemos modelos teóricos que guíen el análisis. Sin embargo, cuando hemos medido variables independientes por razones teóricas, puede ser más interesante el análisis secuencial, que no está orientado a la construcción de modelos, sino al contraste de modelos ya especificados antes del análisis.

## 3.3. INTERPRETACIÓN DE LOS RESULTADOS DEL ANÁLISIS DE REGRESIÓN

La interpretación de los resultados del análisis depende de las correlaciones existentes entre las variables independientes, de su correlación con la variable dependiente y del método de introducción de las predictoras en la ecuación. Por ello, es conveniente obtener la matriz de correlaciones entre variables independientes y de correlaciones parciales y semiparciales de cada una con la dependiente. Esto es necesario porque los coeficientes de regresión no estandarizados dependen de la escala de medida, y los estandarizados no son interpretables de forma unívoca. Por ejemplo, un alto coeficiente de regresión estandarizado con una baja correlación del predictor con el criterio puede deberse a que la variable predice cuando hay otras presentes en el modelo. Por tanto, para tener una pintura completa necesitamos las correlaciones de cada variable con

la dependiente, y esas correlaciones teniendo en cuenta el resto de predictores. Recordemos que  $R^2$  indica la proporción de variabilidad total explicada por los predictores y que, por tanto, un predictor que haga contribuciones significativas debe disminuir  $R^2$  si se elimina de la ecuación e incrementarlo si se introduce en ella. Una medida del cambio en  $R^2$  nos informará, pues, de la contribución del factor a la predicción. La mejor medida en este sentido es la correlación semiparcial, cuyo significado depende del método empleado en la regresión. Comenzaremos primero asumiendo que el método empleado es el estándar, y que, por tanto, todas las variables entran en la ecuación simultáneamente. Después analizaremos el caso de la introducción secuencial.

### 3.3.1. Método simultáneo

En SPSS podemos obtener información sobre las correlaciones en el método simultáneo mediante la siguiente secuencia de comandos:

Analizar → Regresión → Lineal → Dependiente: Y; Independientes: X1, X2, X3 → Método: Introducir → Estadísticos: Matriz de covarianzas; Correlaciones Parcial y semiparcial → Continuar → Aceptar (Análisis)

El comando Estadísticos: Matriz de covarianzas nos proporciona las covarianzas y correlaciones entre las variables dependientes. El comando Estadísticos: Correlaciones Parcial y Semiparcial sirve para obtener las correlaciones de las independientes con la variable dependiente. La correlación semiparcial de un predictor con el criterio es la cantidad en que se reduce  $R^2$  si ese predictor se elimina de la ecuación. Por tanto, nos indica hasta qué punto ese predictor contribuye a la correlación múltiple, a la predicción. Los resultados del análisis de los datos del ejemplo obtenidos mediante la anterior secuencia de comandos en SPSS aparecen en el Cuadro 8.8.

Cuadro 8.8. Correlaciones de las independientes con el criterio (izquierda) y entre sí (derecha). La correlación de orden cero no tiene en cuenta a los demás predictores

Correlaciones con el criterio				Correlaciones de los coeficientes		
	Orden cero	Parcial	Semiparcial	X3	X1	X2
Correlaciones	X3	1	-0,333	-0,689		
	X1	-0,333	1	-0,438		
	X2	-0,689	-0,438	1		
Covarianzas	X3	0,0586	-0,02	-0,086		
	X1	-0,02	0,0586	-0,055		
	X2	-0,086	-0,055	0,2656		

Nótese que la mayor correlación semiparcial corresponde al predictor X3, que en análisis secuenciales precedentes era el único que dejábamos en la ecuación de predicción. Este predictor tiene también la mayor correlación absoluta (orden cero, sin tener en cuenta a los demás predictores), la mayor correlación parcial y el mayor coeficiente de regresión estandarizado, por tanto, su impacto en la predicción es claramente superior al del resto de predictores.

### 3.3.2. Métodos secuenciales

En los métodos secuenciales las variables son introducidas siguiendo un orden, determinado por criterios estadísticos (regresión por pasos) o teóricos (regresión jerárquica). En cualquiera de los casos, el coeficiente de regresión semiparcial es un indicador de la cantidad en que  $R^2$  aumenta en el momento en que el predictor es incluido en la ecuación de predicción. El Cuadro 8.9. presenta los resultados obtenidos en SPSS con la siguiente secuencia de instrucciones:

Analizar → Regresión → Lineal → Dependiente: Y; Independientes: X3 → Siguiente bloque → Independientes: X2 → Siguiente bloque → Independientes: X1 → Método: Pasos sucesivos → Continuar → Aceptar (Análisis)

Como puede comprobarse, los predictores son introducidos uno a uno en la ecuación, de forma que el programa genera tres modelos de regresión, el primero con X3 como predictor único, el segundo con X3 y X2, y el tercero con los tres predictores. Inicialmente la variable X3 incrementa enormemente  $R^2$ , como prueba su correlación semiparcial, pero la introducción sucesiva de predictores disminuye su contribución. No obstante, sigue estando claro que X3 es el mejor predictor, puesto que con todos los predictores dentro, su correlación semiparcial, y también su correlación parcial sigue siendo la mayor.

Cuadro 8.9. Correlaciones absolutas, parciales y semiparciales entre predictores y criterio en cada uno de los tres modelos de regresión

		Correlaciones		
Modelo		Orden cero	Parcial	Semiparcial
1	X3	0,9962	0,9962	0,9962
2	X3	0,9962	0,866	0,1424
	X2	0,9864	0,323	0,0281
3	X3	0,9962	0,9186	0,151
	X2	0,9864	0,5883	0,0473
	X1	0,9641	-0,612	-0,05

### 3.4. EL ANÁLISIS DE VARIANZA COMO ANÁLISIS DE REGRESIÓN

¿Podemos utilizar el análisis de regresión para evaluar el efecto de variables independientes manipuladas en el marco de un diseño experimental o cuasi experimental? Podemos, puesto que ANOVA y Regresión Múltiple son casos particulares del Modelo Lineal General. Supongamos que medimos la memoria prospectiva a tres grupos de personas de diferente edad, mayores de 60 (G1), entre 51 y 60 (G) y entre 35 y 50 años (G3). ¿Cómo haríamos el análisis de varianza desde una aproximación de regresión en este caso? La respuesta es simple, tenemos una variable independiente manipulada entre grupos a 3 niveles (la edad) y una variable dependiente, la memoria prospectiva. La idea es determinar si la variable dependiente puede ser predicha a partir de la variable independiente. El análisis requiere que creamos variables ficticias (*dummy*) que sean capaces de separar los niveles de la variable independiente. Para crear estas variables ficticias necesitamos preguntarnos cuántas son necesarias para separar los grupos. Una forma fácil de responder consiste en enfocar el problema como si se tratase de realizar comparaciones entre los niveles de la variable independiente. Así pues, podemos reformularlo como ¿cuántas comparaciones necesitamos para saber qué grupos difieren de qué otros? Es evidente que con tres grupos dos comparaciones ( $j$ ) son suficientes. Por ejemplo, si la primera compara a los grupos G1 y G2, la segunda compara a esos dos juntos contra el grupo que queda, G3. Podemos expresar esas ideas en ecuaciones simples:

$$\varphi_1: G1 - G2 = 0, \text{ que indica que esperamos que } G1 \text{ y } G2 \text{ sean iguales.}$$

$$\varphi_1: (G1 + G2) - 2G3 = 0, \text{ que indica que } G1 \text{ y } G2 \text{ juntos son iguales a } G3.$$

Las comparaciones son sumas ponderadas de medias, por tanto, nos basta saber los pesos de cada media para explicitarlas. En este caso, tendríamos que la primera comparación tendría coeficientes 1, -1 y 0, respectivamente para G1, G2 y G3; mientras que la segunda tendría coeficientes 1, 1 y -2. Además, debemos tener en cuenta que, por definición, los coeficientes suman 0, y que por tanto, la constante,  $a$ , no puede ser computada de la forma ordinaria. Para calcularla basta con introducir un predictor más, cuyos coeficientes sean todos uno (llamémosle  $X_0$ ). Así, tendremos tres predictores, la constante ( $X_0$ ), la primera comparación ( $X_1$ ) y la segunda ( $X_2$ ). Si organizamos los coeficientes en una matriz ( $X$ ), y suponemos los datos ficticios de memoria prospectiva de 12 sujetos, 4 por cada grupo (Y), tendremos:

Suj.	Grupo	Y	X <sub>0</sub>	X <sub>1</sub>	X <sub>2</sub>
1	G1	6	1	1	1
2	G1	7	1	1	1
3	G1	6	1	1	1
4	G1	5	1	1	1
5	G2	9	1	-1	1
6	G2	9	1	-1	1
7	G2	7	1	-1	1
8	G2	9	1	-1	1
9	G3	13	1	0	-2
10	G3	12	1	0	-2
11	G3	11	1	0	-2
12	G3	14	1	0	-2

Los cálculos podemos hacerlos ahora siguiendo el mismo procedimiento que hemos empleado en el análisis estándar. Aparecen ilustrados en el Cuadro 8.10. (página siguiente). La numeración indica la secuencia de operaciones. Puede apreciarse que la predicción es significativa (columna Sign) y que, por tanto, podemos concluir que hay diferencias significativas entre los grupos. Obviamente, si no hubiera diferencias entre los grupos la pertenencia al grupo no sería información suficiente para realizar una predicción fiable de la memoria prospectiva (Y).

Es importante en este punto tener en cuenta tres consideraciones. Primero, el número desigual de sujetos puede hacer que las variables X<sub>1</sub> y X<sub>2</sub> correlacionen, aunque no deberían. La correlación es un problema del análisis de regresión que trataremos más adelante cuando hablamos de las condiciones y limitaciones del análisis. La segunda consideración tiene que ver con el número de factores que hemos manipulado. Supongamos que hubiésemos cruzado factorialmente dos variables independientes, digamos, la edad, como en el ejemplo anterior, y el nivel cultural de los participantes también a tres niveles (bajo, medio y alto). La cuestión está ahora en definir variables ficticias, para cada factor, pero también para la interacción de ambos.

Por tanto, la codificación ficticia debería incluir, dos variables para separar los grupos de edad (X<sub>1</sub> y X<sub>2</sub>), otras dos para separar los grupos de nivel cultural y cuatro (el producto de las del primer factor por las del segundo) para separar las condiciones experimentales obtenidas por la mezcla de edad y nivel cultural. El Cuadro 8.11. (en página 244) presenta la codificación de este ejemplo, prescindiendo del número de sujetos. Se ha utilizado A, para referirse al primer factor (edad), B para el segundo (nivel cultural) y AB para la interacción. La tabla inferior contiene la sumas de los productos de los coeficientes de todos los pares posibles de predictores. La diagonal contiene, pues, la suma de los cuadrados de los coeficientes de cada predictor. Por lo demás, el análisis procede como se ha ilustrado para el caso de un solo factor entre grupos.

Cuadro 8.10. Aproximación de regresión al ANOVA. Un factor entre grupos

Gr	Y	X0	X1	X2	Prd	Error
1	6	1	1	1	6	0
1	7	1	1	1	6	1
1	6	1	1	1	6	0
1	5	1	1	1	6	-1
2	9	1	-1	1	8,5	0,5
2	9	1	-1	1	8,5	0,5
2	7	1	-1	1	8,5	-1,5
2	9	1	-1	1	8,5	0,5
3	13	1	0	-2	12,5	0,5
3	12	1	0	-2	12,5	-0,5
3	11	1	0	-2	12,5	-1,5
3	14	1	0	-2	12,5	1,5
Media	9	1	0	0	9	0

$$\begin{aligned} \text{ML} &= \text{Const} + \text{Reg}, \text{ luego } \text{SCREG} = 1058 - 972 = 86 \\ \text{6) Prd} &= a + b_1 \cdot x_1 + b_2 \cdot x_2 \\ \text{7) Err} &= Y - \text{Prd} \\ \text{8) SC} &= (\text{Media } Y \cdot n)^2 / n \\ g_{\text{REG}} &= p; \quad g_{\text{RES}} = n - p - 1 \end{aligned}$$

9)

Sumas de cuadrados

Y	X0	X1	X2	ML	Error
1.068	12	8	24	1.058	10

FV	gl	SC	MC	F	Sign
Total	1.068				
Const.	972				
Reg.	2	86	43	38,7	0,00
Res.	9	10	1,1		

1) Matriz traspuesta de X: X'

1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	-1	-1	-1	-1	0	0	0	0
1	1	1	1	1	1	1	1	-2	-2	-2	-2

2) (X'X)

$$\begin{bmatrix} 12 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 24 \end{bmatrix}$$

3) (X'X)<sup>-1</sup>

$$\begin{bmatrix} 0,08 & 0 & 0 \\ 0 & 0,13 & 0 \\ 0 & 0 & 0,04 \end{bmatrix}$$

4) (X'Y)

$$\begin{bmatrix} 108 & 108 & 108 \\ -10 & -10 & -10 \\ -42 & -42 & -42 \end{bmatrix}$$

5) b=(X'X)<sup>-1</sup>(X'Y)

$$\begin{aligned} a &= 9,0 \\ b_1 &= -1,3 \\ b_2 &= -1,8 \end{aligned}$$

La tercera es el tipo de manipulación. Hasta aquí hemos considerado diseños que implican factores entre grupos. ¿Cómo se realiza el análisis si el factor (o factores) está manipulado intrasujetos? La solución pasa por percatarse de que ahora no habrá que predecir una puntuación por sujeto, sino tantas como niveles tenga el factor repetido. Eso significa que tendremos que considerar dos clases de errores de predicción, uno relativo a la puntuación de cada sujeto en cada condición y otro respecto al total del sujeto en el experimento.

Cuadro 8.11. Codificación ficticia de un diseño factorial de dos factores entre grupos con tres niveles por factor (tabla superior). Las sumas de los productos deben ser cero (tabla inferior)

Factores	Const	Factor A		Factor B		Interacción AxB				
A	B	C	A1	A2	B1	B2	AB11	AB12	AB21	AB22
1	1	1	1	1	1	1	1	1	1	1
1	2	1	1	1	-1	1	-1	1	-1	1
1	3	1	1	1	0	-2	0	-2	0	-2
2	1	1	-1	1	1	1	-1	-1	1	1
2	2	1	-1	1	-1	1	1	-1	-1	1
2	3	1	-1	1	0	-2	0	2	0	-2
3	1	1	0	-2	1	1	0	0	-2	-2
3	2	1	0	-2	-1	1	0	0	2	-2
3	3	1	0	-2	0	-2	0	0	0	4

Sumas de cuadrados (diagonal) y sumas de productos								
	A1	A2	B1	B2	AB11	AB12	AB21	AB22
A1	6	0	0	0	0	0	0	0
A2	0	18	0	0	0	0	0	0
B1	0	0	6	0	0	0	0	0
B2	0	0	0	18	0	0	0	0
AB11	0	0	0	0	4	0	0	0
AB12	0	0	0	0	0	12	0	0
AB21	0	0	0	0	0	0	12	0
AB22	0	0	0	0	0	0	0	36

#### 4. LIMITACIONES Y CONDICIONES DE APLICACIÓN

Por lo que llevamos visto hasta aquí la predicción del criterio a partir de las variables independientes no es prácticamente nunca perfecta. Eso no aparenta tener mayor importancia si la exactitud de la predicción es suficientemente grande, si es significativa. Sin embargo, cuando los errores son grandes cabe preguntarse si se deben a que realmente los predictores no sirven para predecir el criterio, es decir, si no hay relación entre la combinación de predictores y el criterio, o, por el contrario, la relación existe pero algo ha fallado en el proceso de estimación de los parámetros. En otras palabras, puede haber relación entre predictores y criterio pero las características de los datos hacen que los coeficientes sean estimados de forma incorrecta. A lo largo de los siguientes párrafos vamos a presentar las características que deben cumplir los datos para que el análisis de regresión proporcione información fiable. Los supuestos o condiciones de aplicación del análisis son siete: Relación lineal entre predictores y criterio, homogeneidad de la varianza error, independencia de los errores, normalidad de los errores, ausencia de valores extremos tanto en las independientes como en la dependiente, ausencia de multicolinealidad. A continuación presentaremos cada uno de estos supuestos y las implicaciones de su no cumplimiento, e insuficiente número de sujetos.

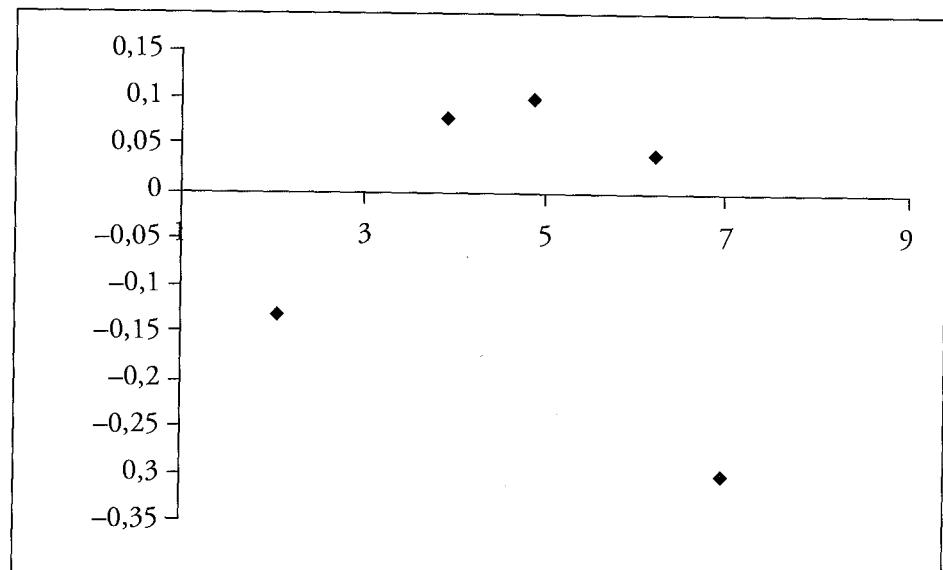
#### 4.1. LINEALIDAD

La regresión está basada en el supuesto de que la relación entre la variable dependiente y la(s) independiente(s) es lineal, esto es, en que la tasa de cambio de la primera por cada unidad de cambio en la segunda es constante a lo largo de toda la variable. La linealidad puede comprobarse fácilmente representando gráficamente los resultados predichos (eje de abcisas) respecto de los errores de predicción (eje de ordenadas). En la Figura 8.1. aparece un ejemplo hipotético. Obsérvese cómo la relación parece de tipo curvilíneo o cuadrático, es decir, las predicciones de los extremos tienen un error asociado de signo y tamaño similar entre sí, pero diferente de los valores centrales. En este caso, el análisis de regresión múltiple no es aplicable, por lo que conviene utilizar análisis alternativos, como la regresión polinómica. SPSS ofrece un buen conjunto de alternativas mediante la serie de comandos:

Analizar → Regresión → Estimación curvilínea

donde encontraremos estimación cuadrática ( $Y = a + b_1X + b_2X^2$ ), cúbica ( $Y = a + b_1X + b_2X^2 + b_3X^3$ ), logarítmica ( $Y = a + b_1\ln(X)$ ), etc. Desafortunadamente, estas aproximaciones sólo pueden emplearse cuando tenemos un solo predictor.

Figura 8.1. Valores predichos (eje de abcisas) frente a errores de predicción. La relación es curvilínea, no lineal

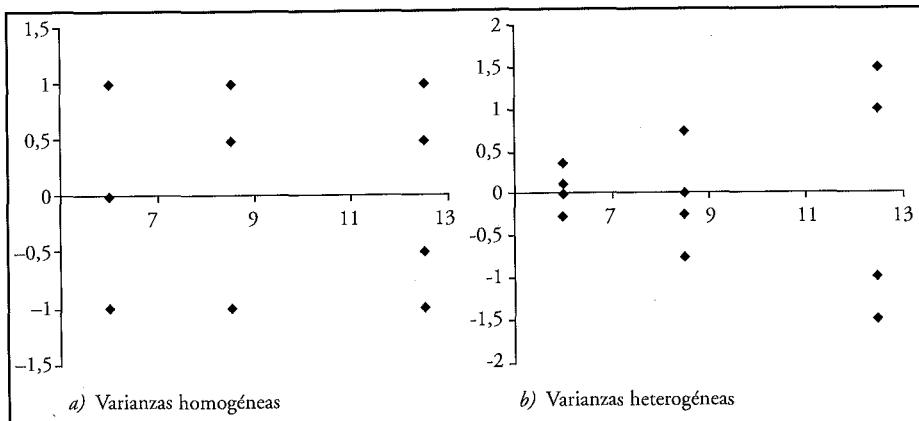


#### 4.2. HOMOGENEIDAD DEL ERROR DE PREDICCIÓN

Cuando los errores cometidos para una misma puntuación predicha son diferentes a lo largo de la escala de la variable dependiente las varianzas error no son homogéneas, hay heteroscedasticidad. La evaluación de este supuesto no puede hacerse si el número de datos que predecimos es muy bajo, particularmente cuando no hay repeticiones de valores predichos (el mínimo número de datos para calcular una varianza es dos). Una forma simple de evaluar consiste en representar gráficamente los valores predichos frente a los errores de predicción. Si la gráfica muestra algún patrón consistente, por ejemplo, una forma triangular o de diamante en cualquier dirección, el supuesto parece no cumplirse. La Figura 8.2. presenta un caso de cumplimiento (panel a) y otro de incumplimiento del supuesto (panel b). Estadísticamente SPSS y otros paquetes estadísticos ofrecen el test de Levene para comprobar la homogeneidad. La ventaja de esta prueba es que es menos sensible a las desviaciones de normalidad que otros estadísticos como la t de Student. El Cuadro 8.12. (pág. siguiente) presenta los resultados obtenidos en SPSS con la siguiente secuencia de comandos:

```
Analizar → Comparar Medias → ANOVA de un factor → Opciones:  
Estadísticos descriptivos: Homogeneidad de varianzas → Continuar → Post  
hoc → DMS → Continuar → Aceptar (análisis)
```

Figura 8.2. Valores predichos (abcisa) frente a errores de predicción



que realiza la prueba de Levene (Homogeneidad de varianzas) y comparaciones *a posteriori* (DMS: Diferencia Mínima Significativa, de Fisher) sobre los datos ilustrados en la Figura 8.2.b. El factor entre grupos ha sido la puntuación predicha tomada del Cuadro 8.10. (6, 8,5 y 12,5) y la variable dependiente el

error de predicción. Obsérvese que, tal y como sugería la inspección visual de la gráfica, las varianzas no son homogéneas: la F de Levene es significativa. Es importante caer en la cuenta de que según la t de Student (DMS de Fisher), no hay diferencias entre los grupos, y puesto que estamos analizando los errores de predicción ello implicaría que las varianzas son homogéneas. La diferencia entre las dos pruebas es fácil de resolver: el test de Levene es el más adecuado para evaluar la homogeneidad de varianza.

El no cumplimiento del supuesto puede resolverse de dos formas diferentes. En primer lugar, cabe emplear la regresión empleando un procedimiento de mínimos cuadrados ponderados. Esta aproximación pondera cada punto del criterio en función de su varianza, de hecho, el peso es la inversa de la varianza. Es fácil ver que los pesos mayores corresponderán a varianzas menores. Por ejemplo, si tenemos un punto con varianza 3 y otro con varianza 6, el peso del primero es el doble que el del segundo (1/3 frente a 1/6). En SPSS podemos utilizar esta opción con el comando MCP>>, para ello debemos crear una variable que contenga los pesos. La segunda opción consiste en utilizar transformaciones de los datos para intentar que cumplan el supuesto. El tipo de transformación más adecuado depende del tipo de datos. Las más frecuentes son la logarítmica, la raíz cuadrada y la arco seno.

Cuadro 8.12. Homogeneidad de varianzas y comparaciones entre grupos

#### 1) F de Levene

Estadístico de Levene	gl1	gl2	Sig.
11,406	2	9	.003

#### 2) Comparaciones entre grupos (t de Student: DMS de Fisher)

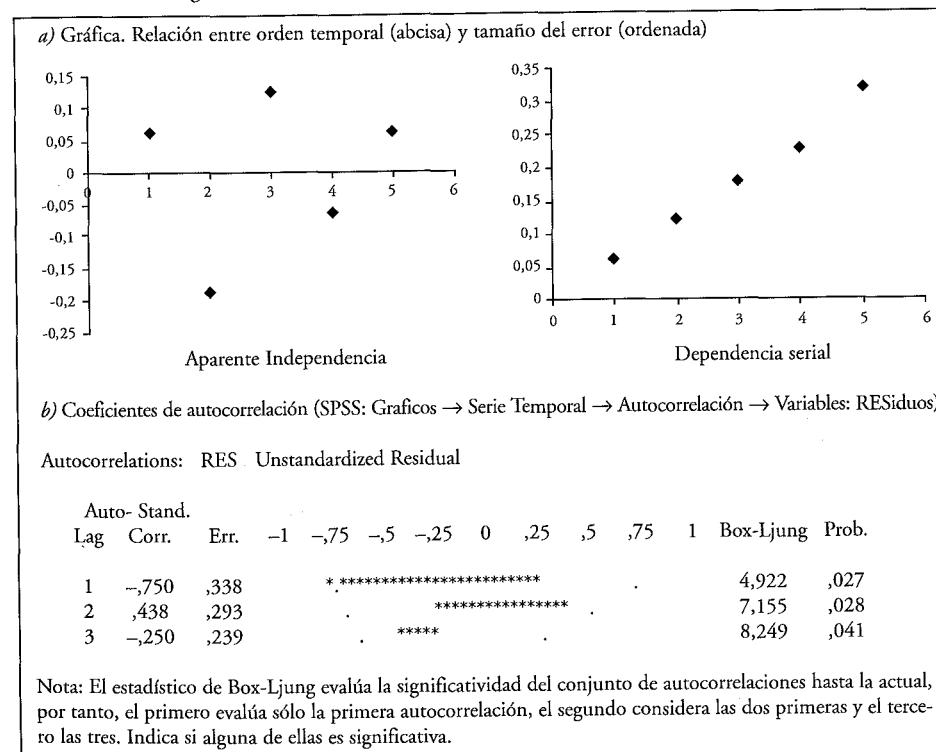
Grupos		Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%
(I) PRED	(J) PRED				Límite inferior Límite superior
6,000000	8,500000	0,875	,6741	,900	-1,4374 1,6124
	12,500000	0,250	,6741	,971	-1,4999 1,5499
8,500000	6,000000	-0,875	,6741	,900	-1,6124 1,4374
	12,500000	-0,625	,6741	,928	-1,5874 1,4624
12,500000	6,000000	-0,250	,6741	,971	-1,5499 1,4999
	8,500000	0,6250	,6741	,928	-1,4624 1,5874

Nota: PRED es la puntuación predicha.

### 4.3. INDEPENDENCIA DE LOS ERRORES

En el análisis de regresión, como en el de varianza, un supuesto fundamental es que el error de predicción asociado a una puntuación no sirva para predecir el error de predicción de otra puntuación. La causa más frecuente de incumplimiento de este supuesto es la obtención serial de los datos. Por ejemplo, si se evalúa un sujeto tras otro y los sujetos se comunican qué estrategias han utilizado para resolver la tarea es muy probable que los errores presenten dependencia serial. Otra causa frecuente es la ganancia en experiencia de quien está realizando las mediciones. La forma más fácil de evaluar el cumplimiento de este supuesto consiste en representar gráficamente la relación entre el orden en que se han adquirido los datos (el tiempo) y los errores de predicción. Si los puntos de la gráfica muestran alguna tendencia, probablemente el supuesto se ha violado. La Figura 8.3 a) presenta la gráfica para los datos del Cuadro 8.1. Asumiendo que el orden en que aparecen en la tabla representa la secuencia en que han sido medidos los datos. En la abscisa aparece el orden y en la ordenada el error de predicción. Nótese como en el panel a) no hay una relación evi-

Figura 8.3. Evaluación del supuesto de independencia de los errores



dente entre ambos, pero sí en el panel b). El análisis estadístico del supuesto puede realizarse calculando el coeficiente de autocorrelación serial (Figura 8.3. b). Este coeficiente es similar al de correlación de Pearson, pero aplicado a una sola variable. El estadístico de Durbin-Watson se emplea para evaluar la correlación serial.

El incumplimiento del supuesto es grave, puesto que si las autocorrelaciones son positivas se subestima la varianza error, lo que hace que se cometan más errores Tipo I de los prefijados, y si son negativas se sobreestima, lo que hace que se cometan menos errores tipo I de los prefijados, pero se reduce la potencia en la toma de decisiones. Una forma de afrontar el incumplimiento del supuesto consiste en transformar los datos, por ejemplo, diferenciándolos. La diferenciación consiste en calcular una nueva serie de puntuaciones que se obtiene como una puntuación menos la anterior (si la demora es 1). Desde un punto de vista de diseño también puede intentarse reducir la dependencia serial, por ejemplo separando las mediciones, dificultando la comunicación entre sujetos, o entrenando adecuadamente a quien realizará las mediciones.

### 4.4. NORMALIDAD DE LAS PUNTUACIONES ERROR

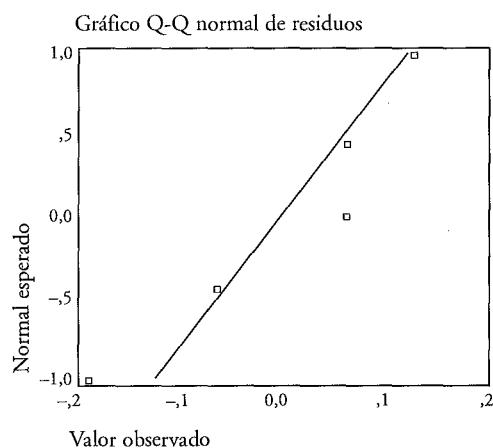
El supuesto implica que los errores se distribuyen según una normal. Es frecuente su incumplimiento, pero no tiene consecuencias muy graves. Para evaluarlo se utiliza un procedimiento gráfico o una prueba de ajuste. En el procedimiento visual se utilizan gráficos de probabilidad normal, que compara los residuos estandarizados con la normal. La normal es una diagonal, si los puntos se separan de ella, probablemente la distribución no es normal. Estadísticamente suelen emplearse bien la prueba de Kolmogorov-Smirnov, bien la de Shapiro-Wilks. SPSS las proporciona ambas, junto con el gráfico de probabilidad normal, con la siguiente secuencia de comandos:

Analizar → Estadísticos Descriptivos → Explorar → Dependientes: Residuos → Gráficos: Gráficos con pruebas de normalidad → Continuar → Aceptar (análisis)

En la Figura 8.4. (página siguiente) aparece la salida de SPSS para la variable Residuos, que contiene los errores de predicción en el ejemplo del Cuadro 8.2. Obsérvese que la desviación de la normalidad no es aparente (panel a, gráfico Q-Q) ni fiable (pruebas de Kolmogorov-Smirnov y de Shapiro-Wilks).

Figura 8.4. Pruebas de normalidad

a) Gráfica de probabilidad normal:



b) Pruebas de normalidad

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Residuos	,291	5	,191	,890	5	,379

K-S con corrección de la significación de Lilliefors

#### 4.5. PUNTOS EXTREMOS

En bastantes ocasiones hay valores del criterio que no son bien predichos. Los errores grandes de predicción producen problemas importantes sobre todo porque disminuyen el coeficiente de correlación múltiple, produciendo la impresión de que la regresión es menos adecuada de lo que en realidad es. Aunque no está claro cuándo un error de predicción es un punto de apalancamiento, suele usarse un criterio asociado al número de puntos de la variable dependiente. En general, se considera que, con los errores estandarizados, un punto situado más allá de 3 desviaciones típicas a la izquierda o a la derecha de la media (que es cero) es un punto de apalancamiento. Una de las recomendaciones más frecuentes respecto de este tipo de casos es eliminarlos (Wilcox, 1997).

#### 4.6. MULTICOLINEALIDAD

Hemos podido comprobar que para poder estimar los coeficientes de la ecuación de regresión múltiple es preciso invertir una matriz que contiene las sumas de cuadrados y productos cruzados de las variables dependientes ( $X'X$ ). Como sabemos, estas sumas son los numeradores de los coeficientes de correlación. Cuando las correlaciones son muy altas y tienden a ser iguales hay multicolinealidad. La consecuencia más importante es que una matriz multicolineal puede ser singular, esto es, no puede invertirse, por lo que no será posible estimar los parámetros. Aun cuando sea posible invertir la matriz, puede haber un segundo problema, a saber: el error típico de los coeficientes de regresión es mayor cuanto mayor es la correlación entre predictores (Berry, 1993), lo que dificulta enormemente declararlos significativos. El diagnóstico de la multicolinealidad se realiza a través de la tolerancia, que es el complementario del cuadrado del coeficiente de correlación múltiple, considerando todas las variables introducidas en la ecuación hasta el momento:

$$\text{Tolerancia} = 1 - R^2$$

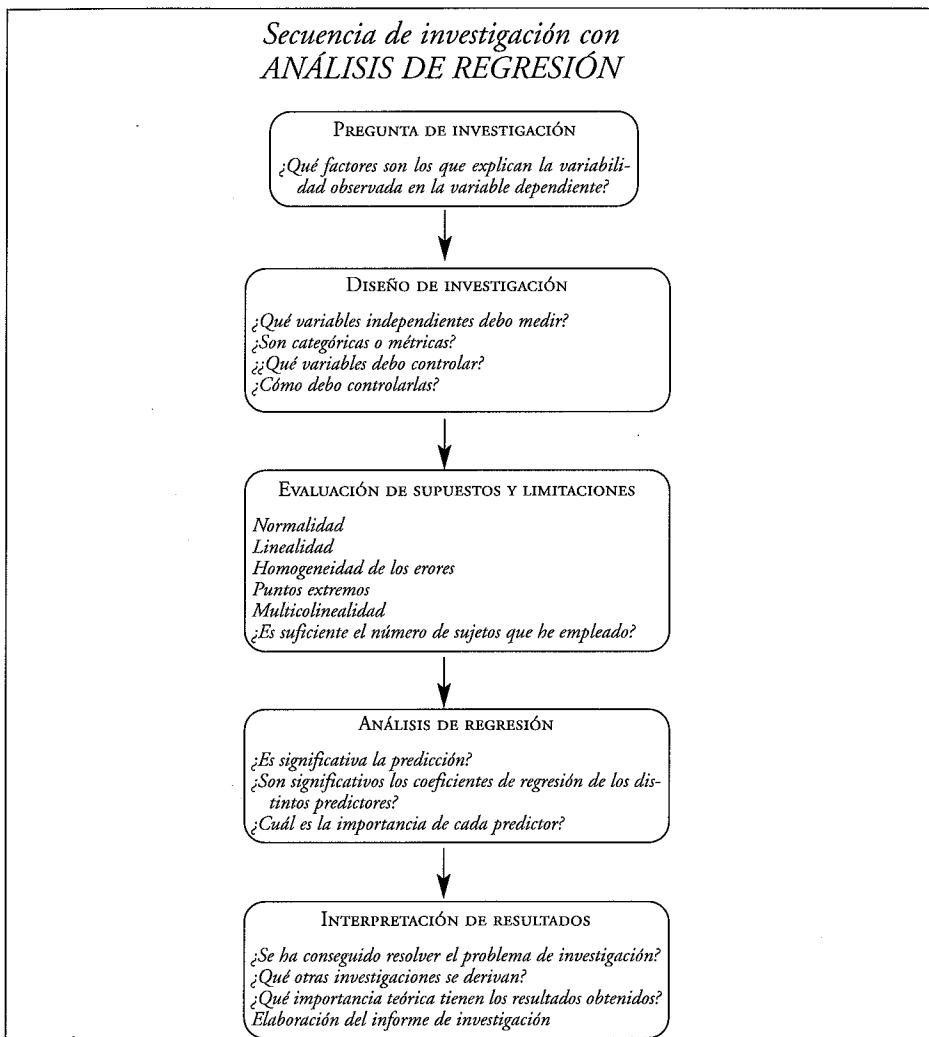
Una tolerancia alta significa que el predictor depende poco del resto de predictores. En el Cuadro 8.6. puede observarse que la tolerancia de  $X_2$  y  $X_3$  es extremadamente baja, lo que implica que tienen una muy alta correlación con la variable incluida en el modelo ( $X_3$ ) y que su aportación a la predicción es prácticamente nula (de ahí que no sean incluidas en la ecuación de predicción). Cuando la multicolinealidad es alta, es aconsejable emplear bien una regresión sesgada (*ridge regression*), bien una regresión por componentes principales. Sin embargo, la regresión sesgada, que intenta estabilizar los parámetros manipulando las varianzas, ha sido criticada seriamente (Fox, 1997), por lo que es más aconsejable emplear regresión por componentes principales. La idea aquí es aprovechar la alta correlación entre predictores para definir variados que son combinaciones lineales de los predictores y emplear los variados como nuevos predictores del criterio.

#### 4.7. NÚMERO DE SUJETOS

Si el número de sujetos en relación con el número de variables dependientes no es muy elevado, el análisis de regresión puede carecer de sentido, puesto que las predicciones pueden llegar a ser perfectas de forma espuria. Una regla razonable es la siguiente:  $n > 50 + 8p$ , siendo  $p$  el número de predictores. Sin embargo, no es una buena estrategia incrementar sin más el número de sujetos, puesto que, como es bien sabido, los coeficientes de correlación pueden llegar a ser significativos, aunque en la práctica sean despreciables (véase Green, 1991, para una discusión en profundidad de este tópico).

## 5. LA SECUENCIA DE INVESTIGACIÓN ORIENTADA AL ANÁLISIS MEDIANTE REGRESIÓN MÚLTIPLE

Cuadro 8.13. La secuencia de investigación orientada al análisis mediante regresión múltiple



La secuencia de análisis en regresión múltiple aparece esquematizada en el Cuadro 8.13 (arriba). Como en el resto de técnicas de análisis es conveniente seguir un conjunto organizado de pasos para optimizar los resultados. En primer lugar es importante tener presente que esta técnica debe ser utilizada sólo cuando se deseé determinar si una o más variables independientes pueden ser

empleadas para predecir una única variable dependiente. El interés de los resultados dependerá de la pregunta de investigación que se haya planteado. En cualquier caso, en ocasiones la técnica puede emplearse para determinar la importancia que cada variable independiente tiene para predecir el criterio.

En segundo lugar, la elaboración del diseño de investigación tiene que desarrollarse de acuerdo con las exigencias que plantea la pregunta inicial. Es importante seleccionar variables independientes adecuadas, determinar cómo van a ser medidas. Finalmente, es preciso elegir un número de sujetos suficiente para garantizar que los test estadísticos van a tener potencia suficiente para detectar efectos de la manipulación si los hubiere.

En tercer lugar, para garantizar que las inferencias basadas en el test estadístico se realizan con tasas de error tipo I y de potencia aceptables es importante evaluar los supuestos y las limitaciones del análisis de regresión. Especial atención debe dedicarse a la inspección de los datos en busca de casos extremos, y al incumplimiento de los supuestos de independencia, de linealidad y multicolinealidad y de homogeneidad de varianzas-covarianzas. A continuación puede aplicarse la prueba estadística, y determinar la importancia de cada predictor evaluando la matriz de estructura.

## 6. LECTURAS RECOMENDADAS

- BERRY, W. D., *Understanding regression assumptions*, Newbury Park, CA, Sage, 1993.  
 COHEN, J. y COHEN, P., *Applied multiple regression/correlation analysis for the Behavioral Sciences*, Nueva York, Erlbaum, 1983.  
 DRAPER, N. R. y SMITH, H., *Applied Regression Analysis*, Nueva York, Wiley, 1998.  
 EDWARDS, A. L., *An introduction to lineal regression and correlation*, San Francisco, Freeman, 1976.  
 FOX, J., *Applied Regression Analysis, Linear Models, and Related Methods*, CA, Sage, 1997.  
 GUILLÉN, M. F., *Ánalisis de regresión múltiple*, Madrid, Centro de Estudios Políticos y Constitucionales, 1992.

## LECTURAS COMPLEMENTARIAS

- ALLISON, P. D., *Multiple Regression: A Primer*, Pine Forge Press, 1999.  
 DILLON, W. R. y GOLDSTEIN, M., *Multivariate analysis: Methods and applications*, Nueva York, Wiley, 1984.  
 GREEN, S. B., «How many subjects does it take to do a regression analysis?», *Multivariate Behavioral Research*, 27, 1991, págs. 346-354.  
 PEDHAZUR, E. J., *Multiple Regression in Behavioral Research*, Wadsworth, 1997.  
 WOODWARD, J. A. y OVERALL, J. E., «Multivariate analysis of variance by multiple regression methods», *Psychological Bulletin*, 82, 1975, págs. 21-32.

## 7. EJERCICIOS

Use SPSS para analizar la siguiente tabla de datos:

Depresión	54	62	5	3	15	34	8	8	46	23	35	11	63	9	33	12
Estatus Ec.	100	35	330	190	165	90	710	130	41	155	101	205	25	305	151	83
Trabajo	0	0	1	1	0	0	1	0	0	1	0	1	1	1	0	1
Edad	28	23	42	50	27	58	54	63	22	27	37	46	42	28	46	15

- Realice un análisis de regresión para determinar si puede predecirse de manera significativa la depresión a partir las variables estatus y edad.
- ¿Qué variables tienen efectos significativos?
- Aplique regresión por pasos incluyendo también la variable trabajo.
- Aplique una regresión jerárquica asumiendo que estatus y trabajo deben entrar juntos y antes que la edad.

### RESPUESTAS A LOS EJERCICIOS

- La predicción es significativa:  $F(2,13) = 6,875$ ,  $p < 0,05$
- Únicamente la variable estatus es significativa: Beta estandarizado = -0,672.
- La única variable que debe quedar en la ecuación es estatus. El resto no realizan aportaciones significativas a la predicción.
- En el primer bloque, sólo la variable estatus hace una aportación significativa, la variable trabajo debe ser excluida de la ecuación. En el segundo bloque, la variable edad no hace una aportación significativa. Por tanto, sólo el estatus debe ser incluido en la ecuación de predicción.

## CAPÍTULO IX

### Análisis multivariado de varianza (MANOVA) y covarianza (MANCOVA)

#### OBJETIVOS DE APRENDIZAJE

- Conocer las ventajas y desventajas del análisis multivariado de varianza.
- Conocer en qué contextos de investigación es idóneo MANOVA/MANCOVA.
- Conocer las características de las hipótesis que MANOVA/MANCOVA permite contrastar.
- Conocer la lógica del MANOVA/MANCOVA.
- Conocer los diferentes estadísticos de contraste que se pueden emplear en MANOVA/MANCOVA.
- Saber aplicar MANOVA en diseños de medidas repetidas.
- Saber qué tipos de comparaciones pueden hacerse una vez que el MANOVA/MANCOVA ha resultado significativo.
- Interpretar los resultados de MANOVA/MANCOVA y de las comparaciones, con una o más de una variable independiente.
- Conocer las limitaciones y supuestos de MANOVA/MANCOVA.

## 1. ANÁLISIS UNIVARIADO (ANOVA) FRENTE A MULTIVARIADO (MANOVA)

El análisis de varianza (ANOVA) es una técnica aplicada con frecuencia en el análisis de datos de experimentos en áreas tan diversas como la Psicología, la Educación o la Sociología. El análisis es apropiado cuando el diseño se restringe a la medición de una sola variable dependiente (un solo aspecto del comportamiento de las unidades experimentales), y se ha manipulado al menos una variable independiente. Sin embargo, su aplicación y utilidad disminuyen cuando en el diseño se incluye más de una variable dependiente, más de una variable criterio. En este caso la técnica analítica de elección, al menos, como veremos más adelante, cuando las variables dependientes están relacionadas entre sí, es el análisis multivariado de varianza (MANOVA) o de covarianza (MANCOVA). Desde un punto de vista estadístico puede considerarse a ANOVA un caso particular de MANOVA y a ambos como casos particulares del modelo lineal general, lo que claramente aparece recogido en paquetes profesionales de análisis como SPSS, STATISTICA o SYSTAT. En las páginas que siguen presentaremos primero la base conceptual del análisis multivariado, después la técnica analítica y, finalmente se exemplificará y contextualizará en relación con otras técnicas multivariadas.

### 1.1. BASES CONCEPTUALES DE MANOVA

Muchas investigaciones pueden realizarse empleando solamente un aspecto del comportamiento como criterio para juzgar el efecto de la variable independiente. En estas circunstancias, la técnica de elección es, sin duda, ANOVA. Sin embargo, pueden encontrarse bastantes casos en áreas muy distintas de la Psicología en que a cada sujeto se le miden por lo menos dos variables criterio. Podemos distinguir dos casos en los cuales MANOVA/MANCOVA es una técnica más apropiada que ANOVA: cuando tenemos varias variables dependientes diferentes y al menos un factor manipulado entre grupos y cuando tenemos uno más factores intrasujeto (o de medidas repetidas), haya o no factores entre grupos.

Por ejemplo, si se quiere examinar el efecto de un conjunto de programas terapéuticos de ansiedad, sería lógico aplicar cada programa a un grupo diferente de sujetos, y además medir mediante cuestionarios varios aspectos de la ansiedad, como la ansiedad ante los exámenes, la ansiedad en situaciones sociales y la ansiedad ante eventos estresantes. Tendríamos, pues un diseño que incluye un factor manipulado entre grupos a tres niveles, las tres terapias, y tres variables dependientes, las diferentes medidas de ansiedad. La variable independiente se mide en escala no métrica (es una variable nominal), mientras que las dependientes se miden es escala métrica (la puntuación en los cuestionarios

es una variable escalar discreta). El análisis de este tipo de diseños unifactoriales entre grupos puede realizarse de dos maneras diferentes, mediante ANOVA o mediante MANOVA.

Si nuestro interés reside en comparar los grupos entre sí en cada variable dependiente por separado, la alternativa de análisis apropiada es ANOVA, puesto que nos permite contrastar las medias de grupo en cada una de las variables dependientes. Así, podríamos saber en qué medidas de ansiedad difieren los grupos y cuáles difieren y cuáles no. Sin embargo, debemos caer en la cuenta de que esto implica realizar, en nuestro ejemplo, tres análisis diferentes, uno por variable. En general, si hemos medido  $p$  variables dependientes, tendremos que realizar  $p$  análisis de varianza. Es fácil percatarse de que eso implica que se han de tomar  $p$  decisiones sobre la eficacia de los tratamientos. Y naturalmente, también es fácil darse cuenta de que, en general, cuanto mayor sea el número total de decisiones que es necesario adoptar, mayor es el riesgo de equivocarse en al menos una de ellas. Si las variables dependientes no están relacionadas entre sí, la probabilidad de cometer al menos un error tipo I (rechazar una hipótesis nula verdadera) en el conjunto de las decisiones, lo que llamaremos error alfa a lo largo del experimento, puede llegar a ser anormalmente alta (podemos calcularla mediante la aproximación de Bonferroni). Cuando las variables no están relacionadas no es posible determinar esta tasa de error, pero en general, crece a mayor número de decisiones.

Un segundo aspecto importante de esta aproximación tiene que ver con la correlación entre las variables dependientes. El análisis separado ignora la correlación sin más, sin embargo, puede existir y ser teóricamente relevante. Por ejemplo, si la ansiedad es un rasgo de los individuos la correlación entre medidas debe ser alta, puesto que la expresión concreta de la misma es una función del rasgo. Considerar las variables dependientes por separado supone, en este ejemplo, adoptar el supuesto teórico de que la ansiedad es situacional, específica del contexto de prueba.

Hay otra serie de diseños en los que tradicionalmente se ha venido empleando ANOVA, aun a pesar de que no es siempre la técnica más aconsejable. Efectivamente, en algunas investigaciones a cada sujeto se le mide la misma variable dependiente en distintas ocasiones a lo largo del tiempo, lo que debe hacernos sospechar que las mediciones estén correlacionadas. Por ejemplo, podemos medir la ansiedad de un grupo de personas antes de comenzar un tratamiento, a la semana del tratamiento, a las tres semanas y a los tres meses. Tendríamos, pues, un diseño de un solo factor manipulado ahora intrasujeto. Los paquetes estadísticos consideran que en este caso cada una de las mediciones repetidas es una variable dependiente. Para aplicar ANOVA a estos diseños el requisito básico es que las variables no estén correlacionadas, o al menos, que su correlación sea homogénea, y sus varianzas sean iguales, esto es, que se cumpla el supuesto de composición simétrica o mejor aún el de esfericidad (Riba, 1990; Vallejo, 1991). Puede repasar el Cuadro 9.1. ahora, si se siente cómodo en el manejo de matrices de varianzas-covarianzas o más adelante cuando éstas

matrices sean discutidas en el contexto de MANOVA. El Cuadro 9.1. presenta el supuesto de esfericidad y su comprobación de manera formal. Un supuesto que se incumple con bastante frecuencia. Su incumplimiento acarrea como consecuencia que el cociente entre medias de cuadrados siga una distribución F desplazada hacia la derecha y por tanto la tasa de error tipo I cometida puede ser sustancialmente mayor que la que se cree cometer. La solución a este problema propuesta a finales de los años cincuenta por Geisser y Greenhouse (1958, Greenhouse y Geisser, 1959) consistió precisamente en emplear una corrección (una prueba F conservadora) para los grados de libertad de la distribución. Así, antes de la difusión de los paquetes estadísticos era habitual emplear la corrección máxima, consistente en dividir los grados de libertad asociados a numerador y denominador de la F por los grados de libertad del numerador. Sin embargo, otra solución que el investigador debe sopesar antes de decidir el análisis que aplicará consiste en utilizar análisis multivariado de varianza para el análisis de estos diseños intrasujetos (O'Brien y Kaiser, 1985).

Cuadro 9.1. Supuesto de esfericidad en ANOVA

El nombre del supuesto se debe a la forma que adopta la nube de puntos en el espacio multidimensional. De hecho, el supuesto de esfericidad implica que las variables tienen igual varianza y su correlación es nula. Técnicamente su comprobación se realiza a través de la prueba de esfericidad de Mauchly:

$$W = \frac{\sum Y_T}{\left[ \frac{\text{Traza} \sum Y_T}{g-1} \right]^{g-1}},$$

donde el numerador es el determinante de la matriz de varianzas-covarianzas de error total, y  $g$  es el número de grupos. La traza de esa matriz es la suma de los elementos de la diagonal principal. La W de Mauchly se puede aproximar a una chi cuadrado según:

$$\chi^2 = - \left[ (n-1) - \frac{2g^2 - 3g + 3}{6(g-1)} \right] \ln(W);$$

$$\text{siendo } v = \frac{g(g-1)}{2} = -1$$

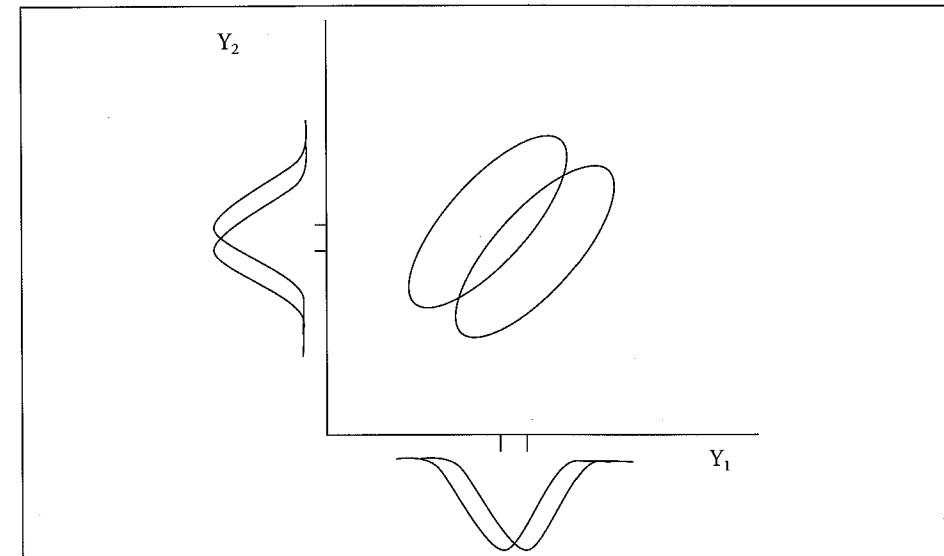
W contrasta la hipótesis nula de que la matriz de covarianzas error de las puntuaciones transformadas es proporcional a la matriz identidad (que tiene 1 en la diagonal y 0 fuera de ella).

En resumen, tanto si el diseño incluye más de una variable dependiente, como si incluye una variable de medidas repetidas, puede aplicarse ANOVA. En el primer caso realizando un análisis por cada variable, en el segundo tratando las distintas mediciones como una variable intrasujeto. Pero en ambas ocasiones debe tenerse en cuenta que se está suponiendo que las correlaciones

entre variables o mediciones no existen o, al menos, son iguales, y además son teóricamente irrelevantes.

¿Qué utilidad tiene MANOVA frente a ANOVA? MANOVA es una técnica que tiene en cuenta todas las variables dependientes de forma simultánea, lo que implica, primero, que la correlación entre ellas, si existe, se tiene en cuenta en el análisis. Por lo que ya hemos visto hasta aquí, el lector se habrá percatado inmediatamente de que decir que se consideran varias variables dependientes de forma simultánea es lo mismo que decir que se mezclan linealmente (recordemos que la mezcla lineal es una suma ponderada). Esto es, en MANOVA el análisis se realiza sobre variables creadas a partir de las dependientes, lo que hemos llamado ya variados. Por tanto, MANOVA nos permite usar más información para tomar decisiones que ANOVA. En segundo lugar, MANOVA puede ser más sensible para capturar los efectos de la manipulación. En algunas circunstancias, es cierto que raras, puede ocurrir que no haya diferencias entre los grupos en ninguna de las variables dependientes consideradas de forma separada, pero sí las haya cuando se consideran simultáneamente. La Figura 9.1. exemplifica este punto en un caso con dos grupos y dos variables dependientes. Nótese que en cada variable, cuyas medias se proyectan sobre el eje horizontal y sobre el vertical, respectivamente, los grupos están muy próximos. Sin embargo, las elipses, que representan la combinación de las dos variables apenas se solapan, lo que indica que los grupos son diferentes cuando las variables se combinan.

Figura 9.1. MANOVA es superior a ANOVA en algunas circunstancias. Es posible que las variables dependientes por separado (medias proyectadas sobre la abscisa y la ordenada) no discriminén a los grupos, pero sí la combinación de ambas (elipses)



En tercer lugar, no necesariamente todas las variables dependientes son útiles para discriminar entre los grupos. Mediante MANOVA podemos discernir cuáles son relevantes y cuáles no para capturar el efecto de la manipulación. Finalmente, MANOVA realiza un contraste sobre el conjunto de las variables dependientes, por lo que la tasa de error tipo I puede mantenerse más fácilmente bajo control.

En algunas ocasiones las condiciones en que ha de realizarse la investigación no permiten que los grupos puedan formarse con el grado apropiado de control experimental. Esto implica que los grupos pueden no ser equivalentes, es decir, pueden diferir, además de en nuestra manipulación, en otros aspectos que podrían ser relevantes para explicar los resultados. Por ejemplo, si los grupos de terapia ya están constituidos como tales cuando vamos a realizar el estudio y en uno de ellos están incluidas personas con fobia social y en otro personas con ansiedad ante los exámenes, es obvio que las posibles diferencias entre ellos tras la terapia podrían explicarse mejor por sus características que por la terapia en sí. En estas situaciones en que el control experimental no puede realizarse de forma suficiente, la alternativa es realizar un control estadístico. Como hemos visto en el Capítulo II, el control estadístico consiste en medir, antes de aplicar los tratamientos, variables (llamadas covariados) que estén relacionadas con las dependientes. En nuestro ejemplo podría bastar con evaluar la ansiedad ante los exámenes, la ansiedad social y la ansiedad ante eventos estresantes antes de comenzar a aplicar las terapias. La estrategia es simple: lo que pueda ser explicado por los covariados no se debe a la manipulación, por tanto, el análisis consiste en contrastar las diferencias entre tratamientos en lo que no puede ser explicado por los covariados. Esta estrategia es conocida como análisis de covarianza (ANCOVA). Como en el caso de ANOVA, ANCOVA puede emplearse para evaluar diferencias entre tratamientos considerando las variables dependientes de forma separada. En este caso, cada una es predicha desde su covariado correspondiente. Sin embargo, también puede intentarse un análisis multivariado (MANCOVA) que considerará simultáneamente todos los covariados y todas las variables dependientes.

En resumen, MANOVA/MANCOVA son recomendables cuando en el diseño se han incluido varias variables dependientes métricas y al menos una variable independiente (que puede ser métrica o no), y el investigador (Bray y Maxwell, 1985, Huberty y Morris, 1989):

1) Pretende protegerse contra el crecimiento del error alfa a lo largo del experimento que implica el análisis univariado de cada uno de los criterios. Para ello puede bastar realizar previamente un análisis multivariado (Riba, 1990). Sin embargo, contra este uso, Huberty y Morris (1989) han señalado acertadamente que un MANOVA significativo no implica necesariamente que alguno de los distintos ANOVA sea o no significativo. No obstante, hay que tener en cuenta que no todo son ventajas en MANOVA/MANCOVA: no siempre un análisis multivariado previo protege del crecimiento del error tipo I (Bird y Hadzi-Pavlovic, 1983).

- 2) Tiene interés en determinar cuáles de entre las variables medidas pueden emplearse para la interpretación de los efectos. El objetivo básico aquí es el de *seleccionar* un subconjunto de variables dependientes.
  - 3) Tiene interés en la aportación relativa de cada criterio a la diferenciación entre los tratamientos, a los efectos obtenidos. La cuestión fundamental ahora es la de *ordenar* la importancia de cada variable dependiente.
  - 4) Tiene razones teóricas para suponer que la correlación entre las variables dependientes es debida a una variable subyacente, a un constructo teórico (por ejemplo, la ansiedad rasgo) del cual depende la expresión de las variables.

## 1.2. EL CONTRASTE DE HIPÓTESIS EN MANOVA/MANCOVA

El análisis mediante ANOVA implica siempre el contraste de tantas hipótesis como variables dependientes diferentes se hayan medido. Todas las hipótesis afirman algo semejante: la igualdad de las medias de tratamiento. Así, si se aplicaron, tres terapias y tres medidas diferentes de ansiedad, se realizarían, como hemos señalado ya, tres análisis. En el primero se probaría la hipótesis de que la ansiedad ante los exámenes es la misma en las tres terapias (es decir:  $H_0: \mu_{11} = \mu_{12} = \mu_{13}$ , donde el primer subíndice denota el número de variable dependiente y el segundo el grupo). En el segundo se contrastaría la hipótesis de que la ansiedad social es la misma en las tres terapias (es decir,  $H_0: \mu_{21} = \mu_{22} = \mu_{23}$ ). Y en el tercero, que la ansiedad ante eventos estresantes es la misma en las tres terapias (es decir,  $H_0: \mu_{31} = \mu_{32} = \mu_{33}$ ).

El análisis mediante MANOVA contrasta una sola hipótesis: que las medias de los g grupos son iguales en las p variables dependientes. Si tenemos en cuenta que en cada grupo tendremos tantas medias como variables dependientes, esto es, tendremos un vector de medias, llamado *centroide*, y que tenemos g grupos, diremos sencillamente que contrastamos la hipótesis de que los g centroides de grupo son iguales. Esto es asumiendo g=3 y p=3:

$$H_0 : \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \end{pmatrix} = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{pmatrix} = \begin{pmatrix} \mu_{31} \\ \mu_{32} \\ \mu_{33} \end{pmatrix}$$

El problema esencial aquí es cómo puede realizarse un contraste de esta naturaleza. Un modo sencillo de entenderlo, y, por tanto, una aproximación intuitiva a MANOVA, es considerar que se crea una nueva variable (*un varia-dor*), que es la combinación lineal de las variables dependientes. Puede comprenderse esto por referencia a un test de inteligencia. Puede obtenerse una puntuación global sumando todos los ítem del test. La puntuación global es una combinación lineal de los ítem en la que todos tienen el mismo valor. El peso es 1 en este caso. Sin embargo, es posible que no todos los ítem tengan

el mismo peso. Por ejemplo, los dígitos hacia delante y hacia atrás del WAIS se supone que miden memoria a corto plazo, pero los dígitos hacia atrás implican más a los mecanismos de control cognitivo que los dígitos hacia delante. En las normas de corrección se concede un punto por cada serie repetida correctamente. Sin embargo, el peso de los dígitos hacia atrás debería ser mayor.

Desde un punto de vista formal, para cada sujeto obtenemos una (o más) nueva puntuación (el variado, que denotaremos como  $V_{ijk}$ , donde  $i$  es el sujeto,  $j$  el variado y  $k$  el grupo), que es simplemente la suma ponderada de sus puntuaciones en cada variable dependiente. De este modo tendremos tantas medidas como niveles tenga la variable independiente (tres en nuestro ejemplo de las terapias) y la hipótesis a contrastar será:

$$H_0 : (v_1 = v_2 = v_3)$$

donde las  $v_k$  son las medias de cada tratamiento en el variado. Después se realiza un análisis de varianza estándar sobre las puntuaciones de los sujetos en el variado. Naturalmente, el problema más importante es cómo se determinan los pesos para construir el variado. Conceptualmente la solución es sencilla. Nuestro interés reside prácticamente siempre en encontrar diferencias entre los tratamientos. De modo que un objetivo adecuado es construir un variado que haga máximas las diferencias entre los tratamientos aplicados. En este sentido, MANOVA puede capturar más fácilmente diferencias entre tratamientos que cuando se considera cada variable por separado, y además, en bastantes ocasiones la interpretación de los resultados puede realizarse de manera más simple. Sin embargo, debe tenerse en cuenta que la técnica no es poderosa cuando las variables dependientes no están correlacionadas y la muestra es pequeña (Bray y Maxwell, 1985).

En resumen, el análisis multivariado de varianza es una técnica que tiene la ventaja de permitirnos contrastar hipótesis sobre los efectos de los tratamientos, pero nos permite también determinar la importancia de cada variable dependiente en el efecto observado. Finalmente, debe recordarse que su uso está recomendado siempre que la correlación entre los criterios es importante en la interpretación de los resultados obtenidos.

## 2. ANÁLISIS DE DATOS MEDIANTE MANOVA/MANCOVA

En esta sección expondremos de la forma más intuitiva posible cómo se analizan los datos mediante una aproximación multivariada de varianza. Teniendo en cuenta que el análisis es una extensión de ANOVA/ANCOVA en nuestra exposición utilizaremos la analogía entre la técnica multivariada y la univariada. En primer lugar presentaremos los diseños que incluyen sólo factores entre grupos, comenzando por el de un único factor y terminando por el

bifactorial. A continuación detallaremos los análisis de diseños que incluyen factores intrasujeto. En cada diseño discutiremos primero el análisis global y después las comparaciones entre condiciones experimentales.

### 2.1. EL ANÁLISIS DE DISEÑOS CON UN FACTOR MANIPULADO ENTRE GRUPOS

El análisis multivariado de varianza puede realizarse obteniendo diversos estadísticos, basados en diferentes supuestos matemáticos, que pueden llevar a distintas decisiones sobre la hipótesis nula de igualdad de los centroides de tratamiento. En los siguientes párrafos presentaremos la lógica del análisis y los diferentes test multivariados y sus características.

Supongamos que hemos aplicado el diseño que hemos presentado más arriba en el que se trataba de comprobar la eficacia diferencial de tres terapias de ansiedad y en el que se proponía medir a cada sujeto del estudio tres variables dependientes: ansiedad ante los exámenes, ansiedad social y ansiedad ante eventos estresantes. Cada terapia es aplicada a un grupo diferente de sujetos. Llamaremos a las terapias G1, G2 y G3, y a las variables dependientes Y1, Y2 e Y3. Supongamos que los datos obtenidos tras la aplicación de los tratamientos a 5 sujetos por grupo han sido:

Sujeto	G1			G2			G3		
	Exam	Social	Estrés	Exam	Social	Estrés	Exam	Social	Estrés
1	13	60	18	15	62	24	17	52	26
2	10	57	20	16	66	22	20	59	31
3	13	65	19	11	61	23	23	59	26
4	16	63	21	12	63	22	19	58	25
5	13	60	22	16	68	19	21	52	27
Medias	13	61	20	14	64	22	20	56	27
SD	2.121	3.082	1.581	2.345	2.915	1.871	2.236	3.674	2.345

La aproximación univariada que hemos comentado previamente trataría estos datos realizando un análisis de cada variable dependiente por separado. Así, puesto que tenemos tres diseños de un solo factor entre grupos (uno por cada variable dependiente) podríamos emplear el programa Comparar Medias: ANOVA de un factor de SPSS. La secuencia de comandos sería:

Analizar → Comparar Medias → Anova de un Factor → Independiente: Grupos → Dependientes: Exam, Social, Estrés → Aceptar (análisis)

Los resultados ofrecidos por SPSS aparecen en el Cuadro 9.2. (página siguiente). Obsérvese que hay diferencias significativas (columna Sig.) entre las

terapias en cada una de las variables dependientes. Recordemos que la variabilidad intergrupos captura el efecto y que la variabilidad error es la intragrupo. Podemos suponer razonablemente que los tratamientos producen efectos diferenciales sobre cada una de las tres variables consideradas por separado. Pero nótese que hemos tenido que tomar tres decisiones, una por cada dependiente, y que la tasa de error tipo I a lo largo del experimento podría ser elevada si hubiésemos tenido que tomar las decisiones a un alfa cercano a 0,05.

Cuadro 9.2. Análisis de varianza para un factor entre grupos en cada variable dependiente

		Suma Cuadrados	gl	Media cuadrática	F	Sig.
EXAM	Intergrupos	143,33	2	71,66	14,33	,00
	Intragrupos	60,00	12	5,00		
	Total	203,33	14			
SOCIAL	Intergrupos	163,33	2	81,66	7,77	,00
	Intragrupos	126,00	12	10,50		
	Total	289,33	14			
ESTRÉS	Intergrupos	130,00	2	65,00	16,95	,00
	Intragrupos	46,00	12	3,83		
	Total	176,00	14			

La aproximación multivariada contrasta simultáneamente las diferencias sobre las tres variables dependientes. En la aproximación que vamos a presentar a continuación emplearemos el álgebra mínima imprescindible. Exposiciones más detalladas pueden consultarse, por ejemplo en Bock (1975) o Dillon y Goldstein (1984). El conjunto de cálculos necesario aparece en el Cuadro 9.3., con los pasos numerados según el orden en que es preciso aplicarlos. Recordemos que nuestra intención es ilustrativa y que SPSS o cualquier otro paquete estadístico los realiza por nosotros. Llamaremos  $Y_{ijk}$  a la puntuación obtenida por el sujeto  $i$ , del grupo  $k$  en la variable dependiente  $j$ ,  $Y$  será la matriz de datos que resulta de poner un grupo debajo del otro, y por tanto tiene tantas filas como sujetos y tantas columnas como variables dependientes, esto es, la matriz tiene dimensión  $N \times p$ . Asumiremos que cada puntuación de  $Y$  es obtenida por la suma de dos efectos: tratamiento y error. Si restamos la media obtenida por la suma de dos efectos: tratamiento y error. Si restamos la media total de cada puntuación y premultiplicamos la nueva matriz por su traspuesta tendremos,  $YY'$ , una nueva matriz ( $T$ ) de dimensiones  $p \times p$ , en la que los elementos de la diagonal principal serán las sumas de cuadrados totales (sin restar la gran media) de los criterios y los elementos fuera de la diagonal serán las sumas de productos de las variables. La matriz contiene, por tanto, los numeradores de las varianzas (llamadas medias de cuadrados, MC, en el contexto de ANOVA) y covarianzas totales. Por tanto:

$$YY' = T = \begin{pmatrix} \sum(Y_{11k} - \bar{Y}_{1k})(Y_{11k} - \bar{Y}_{1k}) & \sum(Y_{11k} - \bar{Y}_{1k})(Y_{12k} - \bar{Y}_{2k}) & \sum(Y_{11k} - \bar{Y}_{1k})(Y_{13k} - \bar{Y}_{3k}) \\ \sum(Y_{12k} - \bar{Y}_{2k})(Y_{11k} - \bar{Y}_{1k}) & \sum(Y_{12k} - \bar{Y}_{2k})(Y_{12k} - \bar{Y}_{2k}) & \sum(Y_{12k} - \bar{Y}_{2k})(Y_{13k} - \bar{Y}_{3k}) \\ \sum(Y_{13k} - \bar{Y}_{3k})(Y_{11k} - \bar{Y}_{1k}) & \sum(Y_{13k} - \bar{Y}_{3k})(Y_{12k} - \bar{Y}_{2k}) & \sum(Y_{13k} - \bar{Y}_{3k})(Y_{13k} - \bar{Y}_{3k}) \end{pmatrix} \quad (\text{Ec. 9.1})$$

Cuadro 9.3. Cálculo de la lambda de Wilks

- 1) Obtenemos la matriz
- $Y$
- sustrayendo la media total de la puntuación de cada sujeto

- 2) Trasponemos la matriz
- $Y$
- cambiando filas por columnas

$$Y^T = \begin{bmatrix} -1 & -5 & -3 & -2 & -4 & 0 & -1 & -6 & -3 & 0 & 5 & 3 & 6 & 6 & 5 \\ -1 & -3 & 3 & 2 & -1 & 2 & 5 & 2 & 4 & 7 & -8 & -1 & -1 & -2 & -8 \\ -5 & -3 & -4 & -2 & -1 & 1 & -1 & 0 & -1 & -4 & 3 & 8 & 3 & 2 & 4 \end{bmatrix}$$

- 3) Multiplicamos
- $Y^T$
- por
- $Y$
- .
- $T$
- contiene las sumas de cuadrados y productos cruzados totales. Cada elemento de la fila de
- $Y^T$
- se multiplica por el elemento correspondiente de las columnas de
- $Y$
- . Se suman los productos.
- 
- (Por ejemplo, fila 1 x columna 1:
- $-1 \times -1 + -5 \times -5 + \dots + 5 \times 5 = 232$
- )

$$T = Y^T Y$$

$$\boxed{\begin{bmatrix} 232 & -123 & 133 \\ -123 & 256 & -107 \\ 133 & -107 & 176 \end{bmatrix}}$$

- 4) Obtenemos la matriz
- $Y_e$
- sustrayendo de la puntuación de cada sujeto la media de su grupo en la variable dependiente correspondiente

- 5) Trasponemos la matriz
- $Y_e$

$$Y_e^T = \begin{bmatrix} 2 & -2 & 0 & 1 & -1 & 2 & 1 & -4 & -1 & 2 & 0 & -2 & 1 & 1 & 0 \\ -1 & -3 & 3 & 2 & -1 & -2 & 1 & -2 & 0 & 3 & -4 & 3 & 3 & 2 & -4 \\ -2 & 0 & -1 & 1 & 2 & 2 & 0 & 1 & 0 & -3 & -1 & 4 & -1 & -2 & 0 \end{bmatrix}$$

- 6) Obtenemos la matriz
- $W$
- , de sumas y productos cruzados error y de la del efecto (
- $B$
- )

$$W = Y_e^T Y_e$$

$$\boxed{\begin{bmatrix} 42 & 17 & -22 \\ 17 & 96 & -7 \\ -22 & -7 & 46 \end{bmatrix}}$$

Puesto que  $T = B + W \rightarrow B = T - W$ 

$$\boxed{\begin{bmatrix} 190 & -140 & 155 \\ -140 & 160 & -100 \\ 155 & -100 & 130 \end{bmatrix}}$$

- 7) Obtenemos la
- $\Lambda$
- de Wilks:
- $|\mathbf{W}| / |\mathbf{T}|$

$$\Lambda = 0,03$$

- 8)
- $\Lambda$
- expresa la proporción de
- $T$
- explicada por el error. Valores cercanos a cero indican que el error no contribuye mucho a
- $T$
- . Valores cercanos a 1 indican que gran parte de
- $T$
- se debe al error, no al tratamiento

Por otra parte, si considerásemos las puntuaciones residuales calculadas restando la media de tratamiento de la puntuación de cada sujeto, obtendríamos una matriz que llamaríamos  $Y_E$  de dimensiones  $n \times p$ . El producto de esta matriz por su transpuesta nos proporcionará la matriz,  $W$ , de dimensiones  $p \times p$ , que contendrá en la diagonal principal las sumas de cuadrados dentro de tratamientos y fuera de la diagonal las sumas de productos residuales entre variables. La matriz contiene, pues, los numeradores de las varianzas y covarianzas error.

Las dos matrices anteriores,  $T$  y  $W$ , nos permiten estudiar ahora fácilmente la correlación entre variables. En efecto, si consideramos los datos teniendo en

$$Y = \begin{bmatrix} -1 & -1 & -5 \\ -5 & -3 & -3 \\ -3 & 3 & -4 \\ -2 & 2 & -2 \\ -4 & -1 & -1 \\ 0 & 2 & 1 \\ -1 & 5 & -1 \\ -6 & 2 & 0 \\ -3 & 4 & -1 \\ 0 & 7 & -4 \\ 5 & -8 & 3 \\ 3 & -1 & 8 \\ 6 & -1 & 3 \\ 6 & -2 & 2 \\ 5 & -8 & 4 \end{bmatrix}$$

$$Y_e = \begin{bmatrix} 2 & -1 & -2 \\ -2 & -3 & 0 \\ 0 & 3 & -1 \\ 1 & 2 & 1 \\ -1 & -1 & 2 \\ 2 & -2 & 2 \\ 1 & 1 & 0 \\ -4 & -2 & 1 \\ -1 & 0 & 0 \\ 2 & 3 & -3 \\ 0 & -4 & -1 \\ -2 & 3 & 4 \\ 1 & 3 & -1 \\ 1 & 2 & -2 \\ 0 & -4 & 0 \end{bmatrix}$$

cuenta los grupos (los tratamientos), encontraremos que las correlaciones intra-grupo entre las tres variables dependientes ( $r_{ykyk}$ ) serán las sumas de productos entre variables divididas por la raíz cuadrada de los productos de las sumas de cuadrados de las variables cuya correlación queremos calcular. Esto es:

$$R_{y1y2(w)} = SP_{y1y2(w)} / (SC_{y1(w)} SC_{y2(w)})^{1/2} = 17 / (42 * 96)^{1/2} = 0,27$$

$$R_{y1y3(w)} = SP_{y1y3(w)} / (SC_{y1(w)} SC_{y3(w)})^{1/2} = -22 / (42 * 46)^{1/2} = -0,50$$

$$y R_{y1y2(w)} = SP_{y1y2(w)} / (SC_{y1(w)} SC_{y2(w)})^{1/2} = -7 / (96 * 46)^{1/2} = -0,11$$

### 2.1.1. Partición de variabilidad

Continuemos ahora con el análisis. Si recordamos la ecuación fundamental de ANOVA (la suma de cuadrados total se obtiene sumando las sumas de cuadrados entre e intra:  $SC_T = SC_{EG} + SC_{IG}$ ), podemos postular, por analogía, que la matriz de sumas de cuadrados y productos totales se obtiene sumando la matriz de sumas de cuadrados y productos cruzados entre (B) con la matriz intra (W):

$$T = B + W \quad (\text{Ec. 9.2})$$

luego las sumas de cuadrados y productos cruzados entre grupos las obtendremos como:

$$B = T - W \quad (\text{Ec. 9.3})$$

de modo que B es también una matriz cuadrada,  $p \times p$ , que contiene en la diagonal principal las sumas de cuadrados y fuera de la diagonal las sumas de productos entre tratamientos (véase Cuadro 9.3., punto 6).

Este mismo resultado habría sido obtenido si hubiésemos trabajado con las puntuaciones diferenciales obtenidas sustrayendo la media total de las medias de tratamiento, esto es hubiésemos computado  $Y_b$ , y multiplicásemos la nueva matriz por su transpuesta y por el número de sujetos por tratamiento, n (véase Cuadro 9.4., en la página siguiente).

Cuadro 9.4. Obtención de la matriz de efectos en el diseño de un único factor entre grupos. Cada elemento de  $Y' B$  se obtiene como media de grupo en una variable dependiente

	$Y'_B$				$Y_B$	
	G1	G2	G3	Exam	Social	Estrés
-3	-3	0	-3			
0	-2	4	-1			
-3	5	-4	4			
				n $Y'_B Y_B$		
	190	-140	155			
	-140	160	-100			
	155	-100	130			

### 2.1.2. La lógica del contraste de hipótesis en MANOVA: variados, valores propios y coeficientes de los variados

En ANOVA, para calcular el estadístico de contraste dividimos variabilidad entre grupos, debida a los tratamientos, por variabilidad error. Estas variabilidades en MANOVA vienen dadas por B y W, respectivamente. Sin embargo, no podemos dividirlas entre sí porque la división entre matrices no existe, pero sí podemos realizarla obteniendo la inversa de W, y multiplicando por B obtendríamos la matriz  $W^{-1}B$ . Esta matriz será el equivalente de la división de  $SC_{ENTRE}/SC_{INTRA}$  necesaria para determinar la F de Snedecor en la aproximación univariada. En MANOVA, no obstante, no trabajamos con las variables dependientes, sino con combinaciones de éstas, los variados. Los valores propios de la matriz  $W^{-1}B$  nos proporcionan los cocientes entre  $SC_{ENTRE}$  y  $SC_{INTRA}$  en los variados. Por tanto, a continuación determinaremos los valores propios (o autovalores) de la matriz resolviendo la ecuación determinantal siguiente:

$$|W^{-1}B - \lambda I| = 0 \quad (\text{Ec. 9.4})$$

En nuestro ejemplo, obtendríamos dos valores propios (la ecuación determinantal tiene dos soluciones): 17,577 y 0,715. El número máximo de autovalores es  $s = \min(g-1, p)$ , es decir, el menor valor entre el número de grupos menos 1 y el número de variables dependientes. Puesto que en nuestro ejemplo tenemos 3 grupos y 3 variables dependientes tendremos dos, pues  $s = \min((3-1), 3) = 2$ . Los valores propios pueden obtenerse en SPSS empleando el programa para análisis discriminante con la siguiente secuencia de comandos:

Analizar → Clasificar → Discriminante → Variable de agrupación: grupos → Definir rango: Mínimo: 1, Máximo: 3 → Continuar → Independientes: Exam, Social, Estrés → Aceptar (análisis)

La salida que ofrece SPSS aparece en el Cuadro 9.5, junto con la ecuación determinantal planteada (panel superior).

Cuadro 9.5. Ecuación determinantal desarrollada y cómputo de los autovalores mediante el análisis discriminante realizado en SPSS

a) Ecuación determinantal

$$\begin{vmatrix} 11 & -7,1 & 7,7 \\ -2,6 & 3,5 & -2,0 \\ 7,5 & -5,2 & 7,2 \end{vmatrix} - \lambda \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 0 = \begin{vmatrix} 11-\lambda & -7,1 & 7,7 \\ -2,6 & 3,5-\lambda & -2 \\ 7,5 & -5,2 & 7,2-\lambda \end{vmatrix}$$

b) Valores propios (SPSS)

Autovalor

Función	Autovalor	% de	%	Correlación canónica
1	17,57 <sup>a</sup>	96,	96,	,97
2	,71 <sup>a</sup>	3,9	100,	,64

Para comprender apropiadamente el significado de los valores propios vamos a suponer que queremos construir una nueva variable combinando las dos dependientes que hemos medido a cada sujeto. Definiremos  $V_{11k}$  como la suma ponderada de las tres conocidas. Formalmente

$$V_{ij1} = a_{11}Y_{11k} + a_{21}Y_{12k} + a_{31}Y_{13k}$$

siendo  $a_{11}$ ,  $a_{21}$ , y  $a_{31}$  los pesos, seleccionados para cada variable dependiente (primer subíndice) en el primer variado (segundo subíndice), de manera que los tratamientos queden situados lo más distantes posible. Sin embargo, no será raro que este variado no informe adecuadamente de toda la variabilidad presente en los datos (puede pensarse, por ejemplo, que teniendo sólo dos grupos un variado es suficiente para distinguirlos, pero teniendo más de dos, pueden necesitarse más para discriminarlos). O dicho de otra forma, y persiguiendo la analogía con ANOVA, cuando tenemos tres grupos necesitamos dos comparaciones para saber qué grupos difieren de qué otros, una puede comparar, por ejemplo G1 y G2, y la otra compara a estos dos juntos contra G3. Del mismo modo, podemos pensar que el primer variado, separa, por ejemplo a dos grupos y el segundo separa a esos dos del tercero, dado que el número máximo de variados que pueden construirse que sean ortogonales entre sí es  $s=\min(g-1, p)$ , es decir, el menor valor entre el número de grupos menos 1 o el número de

variables dependientes, que es el número de soluciones de la ecuación determinantal. Construiríamos, por tanto, un nuevo variado que sería ortogonal (no correlacionado) con el anterior:

$$V_{ij2} = a_{12}Y_{11k} + a_{22}Y_{12k} + a_{32}Y_{13k}$$

El cálculo de los coeficientes de los variados es tedioso, y será tratado en el capítulo de análisis discriminante. No obstante, el lector interesado puede remitirse al capítulo X o Riba (1990). Una forma fácil de obtenerlos consiste en emplear el programa de análisis discriminante de SPSS en el que introducimos como variable de agrupamiento los grupos y como independientes las tres medidas de ansiedad (nótese que la terminología cambia y que las dependientes de MANOVA son independientes en Discriminante). La secuencia de comandos a emplear es la siguiente:

Analizar → Clasificar → Discriminante → Variable de agrupamiento: Grupo → Definir rango: Mínimo=1, Máximo = 3 → Continuar → Independientes: Exam, Social, Estrés → Estadísticos → Coeficientes de la función: No tipificados → Continuar → Aceptar (análisis)

La salida de SPSS aparece parcialmente reproducida en el Cuadro 9.5. Concretamente interesan los coeficientes que SPSS llama de las funciones canónicas discriminantes, que son los coeficientes de los variados que andamos buscando. Ahora podemos combinar las puntuaciones de cada sujeto en las dos variables dependientes para obtener los variados,  $V_1$  y  $V_2$  y realizar las mismas operaciones que hemos visto antes respecto de las variables dependientes. Es evidente que las sumas de productos serán nulas (puesto que los variados son ortogonales entre sí) y las matrices presentarán solamente las sumas de cuadrados de los variados. El Cuadro 9.6. en su parte inferior presenta los cálculos necesarios. Primero se obtiene cada variado en cada sujeto, después se procede como ya hemos visto antes: se obtienen las matrices  $T_V$ ,  $W_V$  y  $B_V$ , que contienen las sumas de cuadrados y productos cruzados de los variados. Finalmente calculamos los cocientes entre suma de cuadrados entre grupos (diagonal de matriz  $B$ ) y suma de cuadrados intragrupo (diagonal de matriz  $W$ ). Podemos observar cómo, efectivamente los valores propios son cocientes entre ambas sumas de cuadrados.

Los valores propios de la matriz  $W^{-1}_V B$  indican, por tanto, el tamaño de la separación entre tratamientos en el variado correspondiente: cuanto mayor sea el valor propio más efecto de la variable independiente refleja el variado. Dicho de otra forma, podemos obtener, por ejemplo, cada variado y realizar un análisis de varianza ordinario y separado para cada uno de ellos. El resultado nos indicará si las terapias difieren entre sí o no.

Cuadro 9.6. Coeficientes de los variados e interpretación de los valores propios de la matriz  $W^{-1}B$ 

1) Coeficientes de los variados obtenidos mediante análisis discriminante en SPSS.

Coeficientes de las funciones canónicas

	Función	
	1	2
EXAM	,570	,34
SOCIAL	-,157	,324
ESTRÉS	,450	,257
(Constante)	-10,053	-25,910

2) Obtención de cada variado para cada sujeto, matrices de sumas de cuadrados y productos error de los variados e interpretación de los valores propios.

$$V_{1i} = -10,053 + 0,570 * \text{Exam}_i - 0,157 * \text{Social}_i + 0,450 * \text{Estrés}_i$$

$$V_{2i} = -25,910 + 0,340 * \text{Exam}_i + 0,324 * \text{Social}_i + 0,257 * \text{Estrés}_i$$

V	Ve		Centroides de grupo	
	V1	V2	V1	V2
-2,67	-1,66	0,397	-0,77	G1 -3,063 -0,888
-3,73	-1,93	-0,67	-1,04	G2 -2,221 0,956
-3,98	-0,17	-0,92	0,715	G3 5,275 -0,113
-2,36	0,051	0,706	0,939	
-2,58	-0,73	0,487	0,156	
0,133	0,89	2,354	-0,07	
-1,81	1,314	0,413	0,358	
-3,74	0,429	-1,52	-0,53	
-2,79	0,922	-0,57	-0,03	
-2,9	1,225	-0,68	0,269	
5,453	-1,67	0,178	-1,55	
5,464	1,819	0,189	1,932	
4,924	0,636	-0,35	0,749	
4,631	0,055	-0,64	0,168	
5,903	-1,41	0,628	-1,3	

$$T_V = V'V$$

222,69	0,00
0,00	20,57

$$W_V = Ve'Ve$$

11,99	0,00
0,00	11,99

$$B_V = T_V - W_V$$

210,70	0,00
0,00	8,58

$$\text{Luego } SC_{Bv1}/SC_{Wv1} = 210,70/11,99 = 17,57 \rightarrow F_{V1} = 105,46$$

$$\text{y } SC_{Bv2}/SC_{Wv2} = 8,58/11,99 = 0,72 \rightarrow F_{V2} = 4,29, p < 0,04$$

## 2.1.3. Estadísticos de contraste

Una vez que está clara la interpretación de los valores propios, es conveniente presentar un modo de tomar decisiones sobre el efecto de la variable independiente. Principalmente se han propuesto cuatro estadísticos de contraste que difieren en la forma de combinar la información proporcionada por los valores propios.

La lambda ( $\Lambda$ ) de Wilks puede interpretarse como la proporción de varianza total *no explicada* por la variable independiente. Su valores oscilan entre 0 y 1. Cuanto menor, más cercano a 0, sea lambda más probable es que el efecto del tratamiento sea significativo. Formalmente:

$$\Lambda = \frac{|W|}{|T|} = \prod_{v=1}^S \frac{1}{1 + \lambda_v} \quad (\text{Ec. 9.5})$$

En nuestro ejemplo tendremos, pues:

$$\Lambda = \frac{1}{1+17,577} \frac{1}{1+0,715} = 0,031$$

lo que indica que la varianza generalizada total es poco explicada por la varianza error. Esto es, que sólo el 3,1% de la varianza total se debe al error.

La traza de Pillai puede interpretarse como la suma de las varianzas explicadas por la variable independiente en los variados. Su ecuación es:

$$V = \sum_{v=1}^S \frac{\lambda_v}{1 + \lambda_v} \quad (\text{Ec. 9.6})$$

y aplicando a nuestro ejemplo obtendremos:

$$V = \frac{17,577}{1+17,577} + \frac{0,715}{1+0,715} = 0,999$$

La mayor raíz característica de Roy es equivalente conceptualmente a las anteriores excepto en que trabaja únicamente a partir del primer variado (el que tiene asociado un mayor valor propio). Por tanto, puede interpretarse como la proporción entre varianza entre grupos y varianza error solamente en el primer variado. Formalmente:

$$R = \lambda_{MAX}$$

y aplicando a nuestros datos obtendremos  $R = 17,577$

El último de los estadísticos que veremos es el Hotelling-Lawley, que se define como:

$$T = \sum_{v=1}^S \lambda_v \quad (\text{Ec. 9.7})$$

y es conceptualmente el más próximo a la F de Snedecor: es equivalente a sumar los cocientes de las sumas de cuadrados entre respecto de las intra. En nuestro ejemplo:

$$T = 17,577 + 0,715 = 18,292$$

SPSS proporciona los cuatro estadísticos (véase Cuadro 9.7.) mediante la siguiente secuencia de comandos:

Anализar → Modelo Lineal General → Multivariante → Dependientes:  
Exam, Social, Estrés → Independiente: Grupo → Aceptar (análisis)

Además, presenta la aproximación a F de Snedecor y el nivel de significación, por tanto, la adopción de una decisión sobre los efectos del tratamiento puede realizarse, a partir de la aproximación a F, como en el ANOVA ordinario.

Cuadro 9.7. Resultados de MANOVA en SPSS. Los datos correspondientes a GRUPO son los que indican el efecto de la manipulación. Intercept hace referencia a la constante

Contrastes multivariados								
Efecto	Valor	F	Gl de la hipótesis	Gl del error gl	Sig.	Eta cuadrado	Parámetro de no centralidad	Potencia observada <sup>a</sup>
Intersección	Traza de Pillai	,999	3239,568 <sup>b</sup>	3,000	,000	,999	9718,703	1,000
	Lambda de Wilks	,001	3239,568 <sup>b</sup>	3,000	,000	,999	9718,703	1,000
	Traza de Hotelling	971,870	3239,568 <sup>b</sup>	3,000	,000	,999	9718,703	1,000
	Raíz mayor de R	971,870	3239,568 <sup>b</sup>	3,000	,000	,999	9718,703	1,000
GRUPO	Traza de Pillai	1,363	7,848	6,000	,000	,682	47,085	,998
	Lambda de Wilks	,031	15,482 <sup>b</sup>	6,000	,000	,823	92,890	1,000
	Traza de Hotelling	18,292	27,438	6,000	,000	,901	164,626	1,000
	Raíz mayor de R	17,577	64,448 <sup>c</sup>	3,000	,000	,946	193,344	1,000

<sup>a</sup> Calculado con alfa = ,05  
<sup>b</sup> Estadístico exacto  
<sup>c</sup> El estadístico es un límite superior para la F el cual ofrece un límite inferior para el nivel de significación  
<sup>d</sup> Diseño: Intercept+GRUPO

Los cuatro estadísticos pueden emplearse para adoptar decisiones sobre el efecto de la variable independiente. Sin embargo, no siempre proporcionarán iguales resultados. De hecho, solamente coincidirán cuando el número de soluciones no triviales, s, de la ecuación determinantal (el número de valores propios de  $W^{-1}B$  distintos de cero) sea 1; esto es, cuando el número de variables dependientes sea 1 o cuando el número de grupos sea 2, y  $(g-1)=1$ . Cuando s es mayor que 1 puede ocurrir que con un estadístico se rechace la hipótesis nula, mientras que con otro no pueda rechazarse.

La decisión sobre el estadístico que pueda adoptarse para realizar el contraste de hipótesis puede depender de los intereses específicos del investigador. En este sentido, conviene tener presente que la investigación Montecarlo sobre la alteración en la tasa de error tipo I y la potencia de los estadísticos, realizados en la década de los años 70, no ha arrojado resultados claros. Así, por ejemplo, Olson (1976, 1979) ha indicado que la traza de Pillai-Bartlett (V) es menos sensible, pero más potente que  $\Lambda$ , mientras que la traza de Hotelling-Lawley es equivalente a V cuando se trabaja con grandes muestras. Sin embargo, Stevens (1979) señala que las diferencias de varianzas usadas por Olson no suelen presentarse habitualmente y que cuando las diferencias no son tan extremas la tasa de error Tipo I es similar para todos los estadísticos, y que V es más apropiada en condiciones especiales. Puede concluirse, pues que la traza de Pillai-Bartlett, seguida de T y  $\Lambda$  son los estadísticos más apropiados.

#### 2.1.4. Potencia

Como se recordará, en ANOVA para estudiar tanto la magnitud de los efectos como la potencia de la F, es imprescindible conocer el parámetro de no centralidad de la distribución, que sencillamente refleja la separación entre la distribución correspondiente a la hipótesis nula y la correspondiente a la alternativa. En el caso multivariado la cuestión es la misma, es necesario conocer el parámetro de no centralidad,  $\delta^2$ . Stevens (1980) ha indicado que este parámetro deberá ser obtenido a partir de  $W^{-1}B$ , y en el caso de g grupos será:

$$\delta^2 = (N-g) \text{ traza } (W^{-1}B) \quad (\text{Ec. 9.8})$$

entendiendo traza como la suma de los valores propios de la matriz, luego el parámetro de no centralidad depende de T de Hotelling-Lawley. Los Cuadros 9.8. y 9.9., adaptados de Stevens (1980), presentan la potencia en función del número de variables dependientes (p), del número de grupos (g), del número de sujetos por grupo (n), de la magnitud del efecto o parámetro de no centralidad (traza), según sea la estructura de la solución difusa (los grupos difieren en más de un variado, Cuadro 9.8.) o concentrada (los grupos difieren sólo en el primer variado, Cuadro 9.9.). Por ejemplificar, supongamos un

investigador que tras realizar un experimento, obtiene en SPSS el valor de T de Hotelling-Lawley de 0,80 y que tiene, como en nuestro ejemplo, tres grupos y 3 variables dependientes con 5 sujetos por grupo. En el Cuadro 9.8. encontraríamos la potencia de los distintos test si su estructura fuera concentrada (solamente el primer valor propio es significativo), que oscila entre 0,416 para la V de Pillai-Bartlett, y 0,449 para la  $\Lambda$  de Wilks. Si su estructura fuera difusa, en el Cuadro 9.7. encontraríamos que la mayor potencia corresponde a V de Pillai-Bartlett, alrededor de 0,546 de potencia, mientras que la menor potencia correspondería a la R de Roy (0,362).

En definitiva, Stevens concluye su análisis indicando que la potencia de los estadísticos depende en gran medida del tamaño de las correlaciones intragrupo entre variables: cuanto mayores sean, mayor es la potencia de los estadísticos, pero que la violación del supuesto de igualdad a través de grupos de las matrices de covarianzas produce alteraciones notables en la estimación de la potencia.

SPSS nos proporciona parámetros de no centralidad diferentes para cada estadístico multivariado, además de indicarnos la potencia exacta en cada caso.

Cuadro 9.8. Potencia de distintos estadísticos multivariados para estructuras difusas sin violación de supuestos.  $a=0,05$

P	n	g	traza	R	T	$\Lambda$	V
2	5	3	,83	528	557	567	613
2	5	3	3,33		994-987		
2	10	3	,37	605	652	660	666
2	10	3	1,48		995-999		
2	50	3	,068	664	708	713	714
2	50	3	,272		1		
2	10	6	,185	382	504	512	510
2	10	6	,74	983	996	996	996
3	5	3	,83	362	436	498	546
3	5	3	3,33	948	985	992	994
3	10	3	,37	482	589	609	615
3	10	3	1,48	987	1	1	1
3	10	6	,185	355	430	435	443
3	10	6	,74	947	992	992	993
6	10	3	,37	341	405	426	419
6	10	3	1,48	933	981	984	984
6	10	6	,185	179	273	282	295
6	10	6	,74	716	940	945	948
10	10	3	,37	202	251	248	254
10	10	3	1,48	774	898	909	927
10	10	6	,185	140	225	216	202
10	10	6	,74	557	847	862	857

Nota: Tomado de Stevens (1980). Los puntos decimales en los valores de potencia han sido suprimidos. Una estructura difusa se produce cuando los grupos difieren a través de varias dimensiones: las diferencias se pueden encontrar a través de varios valores propios, de varios variados. La traza se refiere a la suma de los valores propios de  $W^{-1}B$ .

Cuadro 9.9. Potencia de distintos estadísticos multivariados para estructuras concentradas sin violación de supuestos.  $a=0,05$

p	n	g	traza	R	T	$\Lambda$	V
2	5	3	,83	562	542	520	520
2	5	3	3,33		993-983		
2	10	3	,37	667	644	632	630
2	10	3	1,48		1		
2	10	6	,185	496	493	478	455
2	10	6	,74		995-1		
2	50	3	,068	709	689	690	691
2	50	3	,272		999		
2	50	6	,034	547	551	550	548
2	50	6	,136		996-998		
3	5	3	,83	430	445	449	416
3	5	3	3,33	964	965	959	925
3	10	3	,37	526	552	562	538
3	10	3	1,48		996-1		
3	10	6	,185	430	413	410	391
3	10	6	,74		992-979		
6	10	3	,37	386	379	380	334
6	10	3	1,48	979	972	956	911
6	10	6	,185	289	268	247	245
6	10	6	,74	948	912	881	801
10	10	3	,37	230	242	204	194
10	10	3	1,48	869	851	790	640
10	10	6	,185	171	215	202	172
10	10	6	,74	848	824	743	598

Nota: Tomado de Stevens (1980). Los puntos decimales en los valores de potencia han sido suprimidos. Una estructura concentrada se produce cuando los grupos difieren solamente en una dimensión: las diferencias se pueden encontrar a través de un solo valor propio. La traza se refiere a la suma de los valores propios de  $W^{-1}B$ .

Los resultados obtenidos con la secuencia de comandos siguiente aparecen en el Cuadro 9.7. (pág. 272):

Analizar → Modelo Lineal General → Multivariante → Dependientes: Exam, Social, Estrés → Independiente: Grupo → Opciones → Potencia Observada → Continuar → Aceptar (análisis)

Nótese cómo, la potencia observada oscila entre 0,998 para la Traza de Pillai y 1 para el resto de estadísticos. La estructura de nuestro ejemplo es difusa, puesto que encontramos diferencias significativas entre los grupos tanto en el primer como en el segundo variado.

### 2.1.5. Magnitud del efecto

Para terminar, de la misma forma que en ANOVA la razón de correlación es un índice de la magnitud del efecto (recordemos que es el cociente entre la suma de cuadrados entre grupos y la total, y que indica la proporción de variancia total explicada por la variable independiente), puede definirse el mismo concepto en el contexto multivariado. Cramer y Nicewander (1979), han propuesto una medida basada en la traza de Pillai-Bartlett:

$$R_m^2 = \frac{V}{s} \quad (\text{Ec. 9.9})$$

donde  $s$  es el número de valores propios no triviales, calculado como  $s = \min\{(g-1), p\}$ . Como en el caso de ANOVA,  $R_m^2$  es afectado por el número de variables y grupos, y por tanto, es conveniente ajustarlo. Serlin (1982) ha propuesto la siguiente ecuación para el coeficiente ajustado:

$$R_m^2(\text{aj}) = 1 - (1 - R_m^2) \frac{N-1}{N-b-1} \quad (\text{Ec. 9.10})$$

siendo  $b=\max\{(g-1), p\}$ , y  $N$  el número total de sujetos.

En nuestro ejemplo, la magnitud del efecto sería  $R_m^2 = 1,363/2 = 0,682$ , que ajustada sería  $R_m^2(\text{aj}) = 0,596$ . SPSS proporciona, además de la basada en la traza de Pillai, medidas de magnitud basadas en el resto de estadísticos multivariados. La secuencia de comandos necesarios para obtener la magnitud sería

Analizar → Modelo Lineal General → Multivariante → Dependientes:  
Exam, Social, Estrés → Independiente: Grupo → Opciones → Estimaciones  
del Tamaño del Efecto → Continuar → Aceptar (análisis)

Los resultados aparecen en el Cuadro 9.7. en la columna Eta cuadrado. Nótese cómo la magnitud del efecto cambia según el estadístico de contraste.

### 2.1.6. ¿Qué estadístico seleccionar?

Como hemos podido comprobar los estadísticos pueden diferir en cuanto a potencia y tamaño del efecto, por tanto cabe preguntarse ¿cuál de ellos será más apropiado para tomar una decisión sobre el efecto de nuestra manipulación? Cuando podemos rechazar la hipótesis nula empleando cualquiera de ellos la elección es irrelevante en la práctica. No obstante, puede estar guiada

técnicamente. En este caso se debería seleccionar el estadístico que mayor potencia tuviese y a igualdad de potencia, el que mayor magnitud de efecto estime. En nuestro ejemplo, el estadístico de elección sería la Raíz máxima de Roy. Sin embargo, también puede estar guiada conceptualmente. Por ejemplo, la Raíz de Roy tiene en cuenta sólo el primer variado, por lo que hace una evaluación incompleta de la solución multivariada. Podría ser mejor, pues, emplear un estadístico que tenga en cuenta todas las soluciones. En cualquier caso, puesto que no existe una guía adecuada para decidir entre estadísticos, el investigador puede seleccionar en función de cuál de ellos le sea más cómodo de interpretar y, por supuesto le proporcione resultados significativos.

### 2.1.7. Contrastes Multivariados

La siguiente etapa en el análisis, una vez que el MANOVA global ha resultado significativo, consiste en la realización de contrastes que concreticen los niveles de la variable independiente que producen las diferencias encontradas (Bird y Hadzi-Pavlovic, 1983; Stevens, 1973, Sharma, 1996). Como en el análisis univariado, un contraste es una comparación entre medias de grupo, esto es, un contraste es una suma ponderada de centroides, específicamente:  $\Phi = CM = c_{i1}\mu_1 + c_{i2}\mu_2 + \dots + c_{ik}\mu_k$ , donde  $C$  es un vector de dimensión  $g$  (por el número de grupos) y cumple  $\sum c_i = 0$ ,  $M$  es la matriz que contiene las  $p$  medias de los  $g$  grupos. La hipótesis que se contrasta en este caso hace referencia a la igualdad de los centroides de los grupos definidos en los coeficientes;  $H_0: \Phi = 0$ .

Como en el caso de la aproximación univariada, el contraste será declarado significativo siempre que la probabilidad de obtener un valor igual o mayor que el estadístico de contraste sea inferior a 0,05. El estadístico de contraste empleado normalmente es el análogo multivariado de la  $t$  de Student, la  $T^2$  de Hotelling. Este se define como:

$$T^2 = \frac{1}{\left( \sum_{j=1}^k \frac{c_{ij}^2}{n_j} \right)} \Phi' \Sigma_W^{-1} \Phi \quad (\text{Ec. 9.11})$$

donde  $\Sigma_W^{-1}$  es la inversa de la matriz de varianzas-covarianzas intragrupo mezcladas. Como en el caso de los estadísticos globales,  $T^2$  se approxima a  $F$  mediante la siguiente expresión:

$$F = \left( \frac{gl_e - p + 1}{gl_e * p} \right) T^2 \quad (\text{Ec. 9.12})$$

que se distribuye como una F de Snedecor con  $p$  y  $gl_e + p - 1$  grados de libertad, siendo  $gl_e$  los grados de libertad del error.

Asumamos, para exemplificar que queremos contrastar si la primera terapia es igual de eficaz que la segunda. En ese caso,  $C = (1 \ -1 \ 0)$ . El conjunto de cálculos necesario para tomar la decisión sobre el contraste aparece en el Cuadro 9.10. Estos cálculos han sido realizados con Excel. Puede comprobarse que una  $F = 2,742$ , no es significativa, consultando en las tablas de F con  $gl_{efecto} = 2$  y  $gl_{denominador} = gl_{error} + p - 1 = 13$ . Parece, pues que no podemos rechazar la hipótesis de que ambas terapias son igualmente efectivas.

Cuadro 9.10. Cálculo de la  $T^2$  de Hotelling para comparar la terapia 1 con la 2

- 1) Primero computamos los productos de coeficientes por centroides:  $C'M$

$$\begin{matrix} C' & & M & & C'M \\ \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} & \begin{bmatrix} 13 & 60 & 20 \\ 14 & 64 & 22 \\ 21 & 56 & 27 \end{bmatrix} & = & \begin{bmatrix} -1 & -4 & -2 \end{bmatrix} \end{matrix}$$

- 2) Despues obtenemos la matriz de covarianzas (dividiendo W por los grados de libertad del error) y su inversa

$$\begin{matrix} W & & \Sigma & & \Sigma^{-1} \\ \begin{bmatrix} 42 & 17 & -22 \\ 17 & 96 & -7 \\ -22 & -7 & 46 \end{bmatrix} & \begin{bmatrix} 3,5 & 1,417 & -1,83 \\ 1,417 & 8 & -0,58 \\ -1,83 & -0,58 & 3,833 \end{bmatrix} & \begin{bmatrix} 0,407 & -0,06 & 0,186 \\ -0,06 & 0,135 & -0,01 \\ 0,186 & -0,01 & 0,348 \end{bmatrix} \end{matrix}$$

- 3) Calculamos  $T^2$

$$\begin{matrix} C' & & \Sigma^{-1} & & C & & T^2 \\ \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} & \begin{bmatrix} 0,407 & -0,06 & 0,186 \\ -0,06 & 0,135 & -0,01 \\ 0,186 & -0,01 & 0,348 \end{bmatrix} & \begin{bmatrix} -1 & -4 & -2 \end{bmatrix} & = & 4,113 \end{matrix}$$

- 4) Obtenemos la aproximación a F de Snedecor

$$F = 2,742 \quad p > 0,09 \quad (\text{No hay diferencias significativas})$$

Desafortunadamente SPSS no proporciona una forma sencilla de realizar estos contrastes. Una manera aproximada de obtenerlos puede consistir en seleccionar sólo los grupos que deseamos comparar. Por ejemplo, suponiendo que queremos comprobar si la terapia 1 y la terapia 2 difieren, podríamos filtrar los datos para eliminar la terapia 3. El filtrado se realizaría con la siguiente secuencia de comandos:

Datos → Seleccionar casos → Si se satisface la condición → Si: Grupo<3  
→ Continuar → Aceptar

Esta secuencia crea una nueva variable Filter\_\$. que adjudica 1 a los casos que serán analizados y 0 a los que serán excluidos del análisis. A continuación realizamos el análisis multivariado como hemos indicado más arriba. Los resultados proporcionados por SPSS aparecen en el Cuadro 9.11. Obsérvese que parece haber diferencias entre la terapia 1 y la terapia 2. La discrepancia entre las dos aproximaciones se debe a que la matriz de errores que emplea SPSS contiene, lógicamente, sólo los correspondientes a los dos grupos que hemos seleccionado. Esta aproximación es razonable sólo si los errores son diferentes a través de grupos. En el Cuadro 9.10. aparece también el resultado de realizar la comparación mediante STATISTICA (6.0), que sí permite incluir el error global (de los tres grupos) en el contraste. Nótese cómo, la solución dada por STATISTICA sí coincide con la obtenida por nosotros computando la  $T^2$  de Hotelling. En nuestra opinión es preferible con mucho utilizar el error global, a no ser que haya evidencia de que los errores difieren a través de grupos.

Cuadro 9.11. Contraste multivariado entre terapia 1 y terapia 2

- 1) Solución de SPSS excluyendo el grupo 3

Contrastes multivariados <sup>b</sup>						
Efecto		Valor	F	Gl de la hipótesis	Gl del error gl	Sig.
Intercept	Traza de Pillai	1,000	4059,833 <sup>a</sup>	3,000	6,000	,000
	Lambda de Wilks	,000	4059,833 <sup>a</sup>	3,000	6,000	,000
	Traza de Hotelling	2029,916	4059,833 <sup>a</sup>	3,000	6,000	,000
	Raíz mayor de Roy	2029,916	4059,833 <sup>a</sup>	3,000	6,000	,000
GRUPO	Traza de Pillai	,729	5,374 <sup>a</sup>	3,000	6,000	,039
	Lambda de Wilks	,0271	5,374 <sup>a</sup>	3,000	6,000	,039
	Traza de Hotelling	2,687	5,374 <sup>a</sup>	3,000	6,000	,039
	Raíz mayor de Roy	2,687	5,374 <sup>a</sup>	3,000	6,000	,039

<sup>a</sup> Estadístico exacto.

<sup>b</sup> Diseño: Intercept+GRUPO

- 2) STATISTICA, empleando error global

	Estadístico	F	gl <sub>efecto</sub>	gl <sub>error</sub>	Sig.
Wilks	0,538555	2,856073	3	10	0,090851
Pillai's	0,461445	2,856073	3	10	0,090851
Hotelling	0,856822	2,856073	3	10	0,090851
Roy's	0,856822	2,856073	3	10	0,090851

Por supuesto, un análisis de contrastes alternativo consistiría en examinar el efecto del factor en cada variable dependiente por separado, reduciendo el diseño a un conjunto de diseños univariados. Remitimos al lector al Capítulo II (donde se presentan los contrastes en diseños de una sola variable dependiente).

## 2.2. MANOVA CON MÁS DE UN FACTOR ENTRE GRUPOS

Supongamos ahora que sospechamos que el tiempo que los sujetos llevan padeciendo el problema de ansiedad podría ser un factor importante en la eficacia de la terapia. Por ejemplo, que la terapia 3 es eficaz sólo cuando el problema se ha sufrido durante un tiempo corto, pero que las otras dos terapias lo son cuando se ha sufrido durante más tiempo. Imaginemos que los datos obtenidos en este estudio hubieran sido los que aparecen más abajo. Por lo que sabemos, la manipulación de más de un factor entre grupos no cambia la lógica seguida hasta aquí. Particularmente es importante recordar que hemos seguido la secuencia de análisis de ANOVA, y que la partición de variabilidad sigue manteniéndose. Restringiéndonos, a dos factores entre grupos, en el caso univariado la variabilidad total se reparte entre variabilidad de los efectos principales, la de la interacción y el error intragrupo, esto es:  $SC_{TOTAL} = SC_A + SC_B + SC_{AxB} + SC_{ERROR\ INTRA\ GRUPO}$ , y de forma semejante ocurre en el multivariado: La variabilidad total se deberá a variabilidad producida por cada factor ( $H_A$  y  $H_B$ ), por la interacción de ambos ( $H_{AxB}$ ) y al error ( $W$ ), esto es:  $T = H_A + H_B + H_{AxB} + W$ , siendo las  $H$  matrices de sumas de cuadrados y productos cruzados de efecto (en el caso de un solo factor la matriz  $B$ ) y  $W$  la de sumas de cuadrados y productos cruzados error.

Tiempo de problema

Grupo	Poco			Mucho		
	Exam	Social	Estrés	Exam	Social	Estrés
1	15	59	18	26	67	22
1	11	57	20	16	63	28
1	13	63	19	18	69	25
1	14	62	21	18	66	26
1	12	59	22	16	63	30
2	16	62	24	22	66	31
2	15	65	22	20	70	26
2	10	62	23	15	69	28
2	13	64	22	19	68	29
2	16	67	19	20	74	24
3	21	52	26	22	60	25
3	19	59	31	18	63	32
3	22	59	26	21	63	27
3	22	58	25	23	62	24
3	21	52	27	21	56	28

Lógicamente habrá que computar tres estadísticos de contraste, uno para cada efecto principal, cada factor, y otro para su interacción. En cada caso para obtener la solución multivariada habrá que resolver, como hemos visto, la ecuación determinantal,

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

Y proceder después al cómputo de los diferentes estadísticos como en el caso unifactorial. Afortunadamente SPSS y otros paquetes estadísticos hacen este trabajo por nosotros, sin embargo, no proporcionan los valores propios de la matriz. Aunque no son necesarios para continuar los cálculos no deja de ser una insuficiencia incomprensible. Nos limitaremos, pues a presentar los análisis que SPSS nos ofrece. Comencemos por examinar los datos visualmente. La Figura del Cuadro 9.12. (en página siguiente) presenta las medias de cada variable dependiente según el grupo de terapia y el tiempo de problema. Obsérvese que cabe esperar una interacción significativa, puesto que en la terapia 3 no hay cambio en dos variables dependientes en función del tiempo, pero sí lo hay en las terapias 1 y 2. Por otra parte, el análisis multivariado realizado mediante SPSS con la siguiente secuencia de comandos corrobora esa impresión (véase tabla de resultados en el Cuadro 9.12.).

Analizar → Modelo Lineal General → Multivariante → Independientes: Grupo, Tiempo → Dependientes: Exam, Social, Estrés → Aceptar (análisis)

Puesto que la interacción es significativa, todo nuestro esfuerzo analítico debe ir encaminado, en general, a concretarla. La primera forma de hacerlo consiste en analizar los efectos simples, esto es, examinar el efecto de un factor en cada uno de los niveles del otro. Así, podríamos examinar en primer lugar el efecto de la terapia cuando consideramos sólo un tiempo de problema corto, y después ese mismo efecto cuando el tiempo es largo. Para ello, primero filtramos los datos para mantener en el análisis únicamente los datos relativos a tiempo corto. A continuación realizamos el multivariante con la secuencia de comandos siguientes:

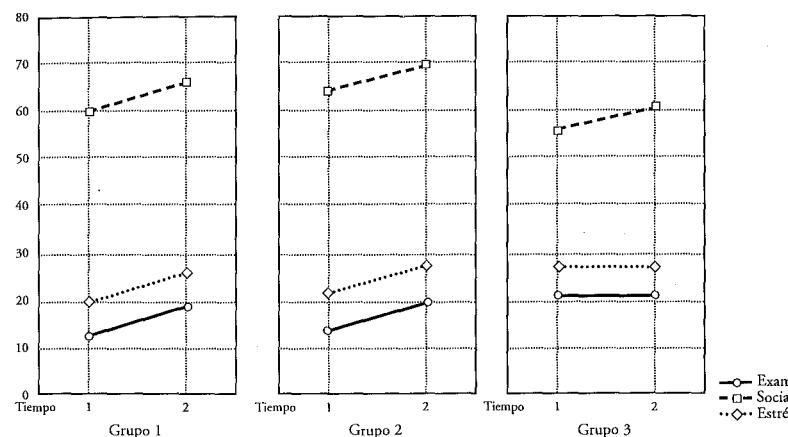
Analizar → Modelo Lineal General → Multivariante → Independientes: Grupo → Dependientes: Exam, Social, Estrés → Aceptar (análisis)

Los resultados aparecen en el Cuadro 9.13. (arriba). Por supuesto, ahora podemos realizar comparaciones para determinar qué terapias difieren de qué otras.

En segundo lugar filtramos los datos para quedarnos sólo con los correspondientes a tiempo largo, y volvemos a realizar el análisis especificado en la anterior secuencia de comandos. Los resultados aparecen en el Cuadro 9.13. abajo. Nótese que tanto el primer como este análisis de efectos simples indican que hay efecto de la terapia. Por tanto, es imprescindible realizar comparaciones entre las terapias para poder concretar más cómo afectan. Finalmente debemos tener en cuenta que es preciso analizar la interacción en la otra dirección, es decir, considerando qué efecto produce el tiempo en función de cada una de las terapias. Dejamos al lector su comprobación.

Cuadro 9.12. Análisis de los efectos de Terapia, Tiempo de problema y su interacción

1) Medias en cada variable dependiente en función de los factores.



2) Resultados proporcionados por SPSS

Contrastes multivariados

Efecto	Valor	F	Gl de la hipótesis	Gl del error gl	Sig.	Eta cuadrado	Parámetro de no centralidad	Potencia observada <sup>a</sup>
Intersección	Traza de Pillai	,999	0101,276 <sup>b</sup>	3,000	,000	,999	30303,827	1,000
	Lambda de Wilks	,001	0101,276 <sup>b</sup>	3,000	,000	,999	30303,827	1,000
	Traza de Hotelling	1377,447	0101,276 <sup>b</sup>	3,000	,000	,999	30303,827	1,000
	Raíz mayor de R	1377,447	0101,276 <sup>b</sup>	3,000	,000	,999	30303,827	1,000
GRUPO	Traza de Pillai	1,269	13,307	6,000	,000	,634	79,843	1,000
	Lambda de Wilks	,104	15,433 <sup>b</sup>	6,000	,000	,678	92,599	1,000
	Traza de Hotelling	5,046	17,661	6,000	,000	,716	105,968	1,000
	Raíz mayor de R	4,189	32,112 <sup>c</sup>	3,000	,000	,807	96,336	1,000
TIEMPO	Traza de Pillai	,848	40,866 <sup>b</sup>	3,000	,000	,848	122,598	1,000
	Lambda de Wilks	,152	40,866 <sup>b</sup>	3,000	,000	,848	122,598	1,000
	Traza de Hotelling	5,573	40,866 <sup>b</sup>	3,000	,000	,848	122,598	1,000
	Raíz mayor de R	5,573	40,866 <sup>b</sup>	3,000	,000	,848	122,598	1,000
GRUPO x TIEMPO	Traza de Pillai	,634	3,556	6,000	,006	,317	21,335	,921
	Lambda de Wilks	,366	4,782 <sup>b</sup>	6,000	,001	,395	28,694	,979
	Traza de Hotelling	1,729	6,053	6,000	,000	,464	36,317	,995
	Raíz mayor de R	1,729	13,258 <sup>c</sup>	3,000	,000	,634	39,774	,999

<sup>a</sup> Calculado con alfa = ,05

<sup>b</sup> Estadístico exacto

<sup>c</sup> El estadístico es un límite superior para la F el cual ofrece un límite inferior para el nivel de significación

<sup>d</sup> Diseño: Intercept+GRUPO+TIEMPO+GRUPO\*TIEMPO

Cuadro 9.13. Análisis multivariado de los efectos simples de terapia

1) Cuando el tiempo de problema es corto

Contrastes multivariados<sup>c</sup>

Efecto		Valor	F	Gl de la hipótesis	Gl del error gl	Sig.
Intercept	Traza de Pillai	,999	3239,568 <sup>a</sup>	3,000	10,000	,000
	Lambda de Wilks	,001	3239,568 <sup>a</sup>	3,000	10,000	,000
	Traza de Hotelling	971,870	3239,568 <sup>a</sup>	3,000	10,000	,000
	Raíz mayor de Roy	971,870	3239,568 <sup>a</sup>	3,000	10,000	,000
GRUPO	Traza de Pillai	1,363	7,848	6,000	22,000	,000
	Lambda de Wilks	,031	15,482 <sup>a</sup>	6,000	20,000	,000
	Traza de Hotelling	18,292	27,438	6,000	18,000	,000
	Raíz mayor de Roy	17,577	64,448 <sup>b</sup>	3,000	11,000	,000

<sup>a</sup> Estadístico exacto.

<sup>b</sup> El estadístico es un límite superior para la F el cual ofrece un límite inferior para el nivel de significación

<sup>c</sup> Diseño: Intercept+GRUPO

2) Cuando el tiempo de problema es largo

Contrastes multivariados<sup>c</sup>

Efecto		Valor	F	Gl de la hipótesis	Gl del error gl	Sig.
Intercept	Traza de Pillai	1,000	7093,016 <sup>a</sup>	3,000	10,000	,000
	Lambda de Wilks	,000	7093,016 <sup>a</sup>	3,000	10,000	,000
	Traza de Hotelling	2,127,905	7093,016 <sup>a</sup>	3,000	10,000	,000
	Raíz mayor de Roy	2,127,905	7093,016 <sup>a</sup>	3,000	10,000	,000
GRUPO	Traza de Pillai	,937	3,229	6,000	22,000	,020
	Lambda de Wilks	,213	3,881 <sup>a</sup>	6,000	20,000	,030
	Traza de Hotelling	2,981	4,472	6,000	18,000	,006
	Raíz mayor de Roy	2,723	9,984 <sup>b</sup>	3,000	11,000	,002

<sup>a</sup> Estadístico exacto.

<sup>b</sup> El estadístico es un límite superior para la F el cual ofrece un límite inferior para el nivel de significación

<sup>c</sup> Diseño: Intercept+GRUPO+TIEMPO+GRUPO\*TIEMPO

### 2.3. MANOVA PARA DISEÑOS DE MEDIDAS REPETIDAS

Los diseños que incluyen factores intrasujeto han suscitado un gran número de controversias respecto de su análisis (véase, Vallejo, 1991). Recordemos que en estos casos cada medición de la misma variable dependiente se considera una variable dependiente diferente (por ejemplo, si medimos en 4 días distintos los aciertos al realizar una tarea de persecución motora, cada medición es una variable dependiente diferente). La razón estriba en que el supuesto de esfericidad tiende a no cumplirse con bastante frecuencia, lo que implica, primero, que las correlaciones entre parejas de variables dependientes no son homogéneas (realmente deberían ser cero), y probablemente que las varianzas de las distintas variables tampoco. El problema surge del hecho de que la aproximación que se realiza en ANOVA no tiene en cuenta las correlaciones entre variables. Por tanto, si la correlación existe (lo cual es bastante frecuente, puesto que medimos reiteradamente a los mismos sujetos), es importante tenerla en cuenta. Justamente una de las diferencias más importantes entre MANOVA y ANOVA estriba en que MANOVA sí tiene en cuenta la relación existente entre pares de variables. Evidentemente es preferible, como diría Pascal, tener en cuenta las correlaciones, aunque sean despreciables que no considerarlas y llegar a conclusiones erróneas. En este sentido, algunos autores (por ejemplo, Riba, 1990; Pascual, Frías y García, 1996; Davidson, 1972) apuntan que MANOVA, debido a su potencia, es preferible siempre que haya más sujetos que variables dependientes, con independencia de las correlaciones. Sin embargo, como han señalado Sedlmeier y Gigerenzer (1989), muchos de los estudios que han empleado ANOVA no alcanzan a tener una potencia adecuada. No es de sorprender, pues, que paquetes estadísticos como SPSS presenten, por defecto, estadísticos multivariados cuando se les pide una aproximación univariada.

En definitiva, el lector que emplee estos diseños tendrá que tomar una decisión sobre la técnica univariada o multivariada que va a emplear. Sin embargo, no hay ningún argumento de sustancia que obligue a tomarla antes en lugar de después de realizar los análisis. SPSS presenta ambos. Por tanto, si ambas aproximaciones coinciden la elección es irrelevante. Si no coinciden, probablemente un investigador pragmático se decante por aquélla que declare significativo el efecto de su manipulación, a no ser que tenga dudas acerca de si ese efecto debiera o no existir. A continuación presentaremos el análisis del diseño unifactorial y después su generalización al caso multifactorial.

#### 2.3.1. Diseños de medidas repetidas con un solo factor

Imaginemos ahora que hemos manipulado el número total de ensayos de entrenamiento en una tarea de persecución motora en un grupo de cinco niños de secundaria. Supongamos que hemos considerado que para obtener una curva de aprendizaje razonable era suficiente con utilizar 3 niveles, sean 10, 50

y 500 ensayos. Imaginemos que hemos aplicado el entrenamiento y que hemos obtenido el número de aciertos en la tarea que aparece en la tabla siguiente.

Niño	Ensayos de entrenamiento			Tot. Niño
	10	50	500	
1	5	8	18	31
2	3	6	17	26
3	4	5	13	22
4	6	10	16	32
5	2	6	16	24
Medias	4	7	16	9

¿Produce el entrenamiento un efecto significativo? La respuesta a esta cuestión es simple, nos basta con codificar el factor mediante un conjunto de comparaciones ortogonales. Recuérdese que las comparaciones se especifican en coeficientes, que deben sumar cero. Dos comparaciones son ortogonales cuando la suma de los productos de sus coeficientes correspondientes es también cero. Esto es lo mismo que decir que son ortogonales si su correlación es nula. A continuación realizamos las comparaciones en cada sujeto. De esta forma obtendremos una matriz de diferencias  $D$ , de tamaño  $n \times c$ , donde  $n$  son los sujetos y  $c$  es el número de comparaciones del conjunto. Lo interesante es que cada elemento de esa matriz contiene efecto del factor y error. Por tanto, los pasos siguientes van encaminados a obtener estimaciones del efecto y del error. Comencemos por aclarar que contener efecto y error sumados equivale a contener variabilidad total. Por tanto, obtener las sumas de cuadrados y productos cruzados totales (la matriz  $T$ ) se reduce a multiplicar  $D'D$ . Ahora podemos construir la matriz de efectos. Ya sabemos que el efecto es la diferencia entre las medias de condición experimental en cada variable dependiente (es decir, en cada contraste). Tendremos, pues que la matriz de sumas de cuadrados y productos cruzados efecto,  $H$ , será la matriz de medias de condición por la transpuesta de coeficientes, esto es:

$$H = MC' \quad (\text{Ec. 9.7})$$

Finalmente, la matriz de sumas de cuadrados y productos cruzados error será la diferencia entre la total y la de efectos, puesto que sabemos que la variabilidad total es la suma de efecto y error:

$$E = T - H \quad (\text{Ec. 9.8})$$

El conjunto de cálculos necesarios aparece en el Cuadro 9.14. Obsérvese que la analogía con el análisis de los diseños entre grupos es grande, apenas construimos la matriz de coeficientes y obtenemos la matriz de diferencias.

Cuadro 9.14. Ejemplificación de los cálculos para obtener A de Wilks en un diseño de medidas repetidas de un solo factor

1) Primero definimos las comparaciones y comprobamos que son ortogonales

$$\begin{matrix} C & \begin{bmatrix} 1 & -1 \\ 0 & 2 \\ -1 & -1 \end{bmatrix} \\ \begin{bmatrix} 1 & 0 & -1 \\ -1 & 2 & -1 \end{bmatrix} & = \begin{bmatrix} 2 & 0 \\ 0 & 6 \end{bmatrix} = \begin{matrix} SCc1 & SP \\ SP & SCc2 \end{matrix} \end{matrix}$$

2) Calculamos las diferencias en cada sujeto,  $D = YC'$

$$\begin{matrix} Y & \begin{bmatrix} 5 & 8 & 18 \\ 3 & 6 & 17 \\ 4 & 5 & 13 \\ 6 & 10 & 16 \\ 2 & 6 & 16 \end{bmatrix} \\ \begin{bmatrix} 1 & -1 \\ 0 & 2 \\ -1 & -1 \end{bmatrix} & = \begin{bmatrix} -13 & -7 \\ -14 & -8 \\ -9 & -7 \\ -10 & -2 \\ -14 & -6 \end{bmatrix} \end{matrix}$$

3) Calculamos la matriz de sumas de cuadrados y productos cruzados total

$$\begin{matrix} D' & \begin{bmatrix} -13 & -7 \\ -14 & -8 \\ -9 & -7 \\ -10 & -2 \\ -14 & -6 \end{bmatrix} \\ \begin{bmatrix} -13 & -14 & -9 & -10 & -14 \\ -7 & -8 & -7 & -2 & -6 \end{bmatrix} & = \begin{matrix} T=D'D \\ \begin{bmatrix} 742 & 370 \\ 370 & 202 \end{bmatrix} \end{matrix} \end{matrix}$$

4) Calculamos los efectos multiplicando medias por comparaciones

$$\begin{matrix} M & \begin{bmatrix} 1 & -1 \\ 0 & 2 \\ -1 & -1 \end{bmatrix} \\ \begin{bmatrix} 4 & 7 & 16 \end{bmatrix} & = \begin{bmatrix} -12 & -6 \end{bmatrix} \end{matrix}$$

5) Calculamos la suma de cuadrados y productos cruzados de efecto:  $H=nMe'Me$

$$\begin{matrix} Me' & \begin{bmatrix} -12 \\ -6 \end{bmatrix} \\ \begin{bmatrix} 5 \end{bmatrix} & = \begin{matrix} Me \\ \begin{bmatrix} -12 & -6 \end{bmatrix} \end{matrix} \\ & = \begin{matrix} H \\ \begin{bmatrix} 720 & 360 \\ 360 & 180 \end{bmatrix} \end{matrix} \end{matrix}$$

6) Calculamos la de error:  $E=T-H$

$$\begin{matrix} \begin{bmatrix} 742 & 370 \\ 370 & 202 \end{bmatrix} & - \begin{bmatrix} 720 & 360 \\ 360 & 180 \end{bmatrix} & = \begin{matrix} E \\ \begin{bmatrix} 22 & 10 \\ 10 & 22 \end{bmatrix} \end{matrix} \end{matrix}$$

7) La lambda de Wilks es el cociente entre determinantes de error y total

$$\Lambda = 0,03$$

El análisis mediante SPSS no se realiza con el programa multivariante, sino mediante el programa de Medidas Repetidas. La secuencia de comandos en este caso sería:

Analizar → Modelo Lineal General → Medidas Repetidas → factor: Ensayos → niveles: 3 → Agregar → Definir → Variables intra sujeto: e1, e2, e3 → Aceptar (análisis)

Los resultados proporcionados por SPSS aparecen en el Cuadro 9.15. Su interpretación está clara: hay efectos significativos del factor entrenamiento.

Cuadro 9.15. Análisis multivariado de medidas repetidas. Un solo factor

Efecto	Valor	F	Gl de la hipótesis	Gl del error gl	Sig.
ENSAYO	Traza de Pillai	0,970	49,21 <sup>a</sup>	2,000	,3,000 ,005
	Lambda de Wilks	,030	49,21 <sup>a</sup>	2,000	,3,000 ,005
	Traza de Hotelling	32,81	49,21 <sup>a</sup>	2,000	,3,000 ,005
	Raíz mayor de Roy	32,81	49,21 <sup>a</sup>	2,000	,3,000 ,005

<sup>a</sup> Estadístico.

Finalmente, puesto que nuestro factor tiene más de dos niveles podemos realizar comparaciones para determinar qué niveles difieren de qué otros. La única alternativa razonable que nos ofrece SPSS es la que hemos indicado antes, seleccionar sólo los niveles que pretendemos comparar. Por ejemplo, si deseamos comparar el nivel 1 con el nivel 3 de ensayos, nos basta con definir el factor con dos niveles e indicar que se trata de e1 y e3.

### 2.3.2. Diseños de medidas repetidas con dos factores

Imaginemos que además de manipular los ensayos de entrenamiento hubiésemos manipulado la distribución de la práctica a dos niveles (masiva, distribuida) también intrasujeto. Si cruzamos factorialmente ambos factores tendríamos seis condiciones experimentales, tres tipos de ensayos con práctica masiva y otros tres con práctica distribuida. Así, tendríamos, como en el caso del diseño factorial entre grupos que calcular tres efectos, el de los ensayos (factor A), el de la distribución de la práctica (factor B) y la interacción entre ambos (interacción AxB). El esquema de partición de variabilidad es, sin embargo, más complejo, puesto que, como sabemos (véase Capítulo II) cada efecto tiene un error particular asociado. Esto es, A tiene un error propio que resulta de su interacción con los sujetos (AxS), igual ocurre con B (BxS) y con AxB (AxBxS). En pocas palabras esto hace más largo y complejo el análisis, pero no diferente.

Supongamos que hemos aplicado los factores y que los resultados, en términos de aciertos hubiesen sido los siguientes:

Niño	Masiva			Distribuida		
	10	50	500	10	50	500
1	5	8	18	8	20	19
2	3	6	17	9	18	21
3	4	5	13	6	19	20
4	6	10	16	8	15	16
5	2	6	16	7	18	19

SPSS nos proporciona la tabla de resultados que aparece en el Cuadro 9.16. cuando empleamos la siguiente secuencia de comandos (siendo m1, m2 y m3 los ensayos en práctica masiva y d1, d2 y d3 en práctica distribuida). La secuencia de comandos fue:

Anализar → Modelo Lineal General → Medidas Repetidas → factor:  
Ensayos → niveles: 3 → Agregar → Factor: práctica → Niveles: 2 → Agregar  
→ Definir → Variables intrasujeto: m1, m2, m3, d1, d2, d3 → Aceptar (análisis)

Cuadro 9.16. Análisis de un diseño de dos factores repetidos

Contrastes multivariados					
Efecto	Valor	F	Gl de la hipótesis	Gl del error gl	Sig.
Práctica	Traza de Pillai	,902	36,615 <sup>a</sup>	1,000	,004
	Lambda de Wilks	,098	36,615 <sup>a</sup>	1,000	,004
	Traza de Hotelling	9,154	36,615 <sup>a</sup>	1,000	,004
	Raíz mayor de Roy	9,154	36,615 <sup>a</sup>	1,000	,004
Ensayos	Traza de Pillai	,988	123,705 <sup>a</sup>	2,000	,001
	Lambda de Wilks	,012	123,705 <sup>a</sup>	2,000	,001
	Traza de Hotelling	82,470	123,705 <sup>a</sup>	2,000	,001
	Raíz mayor de Roy	82,470	123,705 <sup>a</sup>	2,000	,001
Práctica*Ensayos	Traza de Pillai	,943	24,988 <sup>a</sup>	2,000	,013
	Lambda de Wilks	,057	24,988 <sup>a</sup>	2,000	,013
	Traza de Hotelling	16,659	24,988 <sup>a</sup>	2,000	,013
	Raíz mayor de Roy	16,659	24,988 <sup>a</sup>	2,000	,013

<sup>a</sup> Estadístico exacto.

El análisis de la interacción se realiza como se indicó en el caso del diseño entre grupos, esto es, obteniendo los efectos simples de un factor en cada nivel del otro. Por ejemplo, los efectos del entrenamiento en práctica masiva primero y luego en práctica distribuida. Igualmente deberían obtenerse los de la

práctica en cada nivel de entrenamiento. Nótese que esta aproximación implica que estamos analizando efectos de un único factor repetido, y que, por tanto, se reduce a lo expuesto en el punto anterior.

### 2.3.3. Diseños factoriales mixtos o parcialmente repetidos

En estos diseños se manipula varios factores, al menos uno entre grupos y al menos otro intrasujetos. Por ejemplo, podríamos haber manipulado la distribución de la práctica entre grupos, de forma que tendríamos un diseño con dos grupos, uno que recibe práctica masiva y el otro distribuida y un factor intrasujeto, el entrenamiento: cada sujeto de cada uno de los dos grupos recibe los tres niveles de ensayos. Como en el resto de diseños factoriales debemos calcular los efectos principales de los factores y sus interacciones. Genéricamente, si A y B son los factores, tendremos efectos de A, de B y de la interacción AxB. En estos diseños el término error de los factores de medidas repetidas es su interacción con los sujetos dentro de los grupos. Por ejemplo, B y Ax B tendrán como error la interacción de B con los sujetos dentro de los grupos definidos por A. Por otro lado, el error de los factores entre grupos es un intragrupo.

Supongamos que los datos del ejemplo anterior han sido obtenidos con el factor práctica manipulado entre grupos. La secuencia de comandos necesaria para realizar el análisis de este ejemplo en SPSS es la siguiente:

Anализар → Modelo Lineal General → Medidas Repetidas → factor:  
práctica → niveles: 3 → Agregar → Definir → Variables intrasujeto: e1, e2,  
e3 → Factores Inter.-sujetos: práctica → Aceptar (análisis)

SPSS nos proporciona ahora dos tipos de análisis, uno multivariado para los factores repetidos (los ensayos de entrenamiento en nuestro ejemplo) y otro univariado para los factores entre grupos (la práctica en nuestro ejemplo). El Cuadro 9.17. presenta los resultados del análisis. Obsérvese que hay efecto significativo (columna Sig.) del factor entre grupos, del factor repetido y de la interacción.

El objetivo siguiente de análisis sería, por tanto, la interacción. Es importante caer en la cuenta de que ahora, cuando consideremos los efectos del factor entre grupos en cada nivel del factor repetido debemos emplear ANOVA, mientras que cuando analicemos los del factor intrasujeto en cada grupo debemos emplear MANOVA, concretamente se trataría de un diseño unifactorial como el que hemos presentado en el apartado 2.3.1.

Cuadro 9.17. Análisis de un diseño parcialmente repetido de dos factores

1) De los efectos intrasujeto					
Contrastes multivariados					
Efecto	Valor	F	Gl de la hipótesis	Gl del error gl	Sig.
Ensayos	Traza de Pillai	,971	116,317 <sup>a</sup>	2,000	,000
	Lambda de Wilks	,029	116,317 <sup>a</sup>	2,000	,000
	Traza de Hotelling	33,233	116,317 <sup>a</sup>	2,000	,000
	Raíz mayor de Roy	33,233	116,317 <sup>a</sup>	2,000	,000
Ensayos*Práctica	Traza de Pillai	,903	32,545 <sup>a</sup>	2,000	,00
	Lambda de Wilks	,097	32,545 <sup>a</sup>	2,000	,000
	Traza de Hotelling	9,298	32,545 <sup>a</sup>	2,000	,000
	Raíz mayor de Roy	9,298	32,545 <sup>a</sup>	2,000	,000

2) Del efecto entre grupos

Fuente	Suma cuadrados tipo	gl	Media cuadrática	F	Sig.
Intercept	4.272,13	1	4.272,13	818,93	,00
Práctica	258,13	1	258,13	49,48	,00
Error	41,73	8	5,21		

#### 2.4. DISEÑOS QUE INCLUYEN COVARIADOS

A menudo en un contexto no estrictamente experimental, pero también con cierta frecuencia en un contexto experimental, los investigadores miden variables sobre las cuales no suponemos que nuestra manipulación produzca ningún efecto (entre otras cosas porque se suelen medir antes de aplicarla), pero que, sin embargo, está relacionada con las variables dependientes que se supone sí reflejarán el efecto de la manipulación. ¿Qué importancia pueden tener estas variables, que hemos llamado covariados? Esencialmente son útiles en dos sentidos. Su correlación con las dependientes puede contribuir, por una parte a reducir la variabilidad error. Puesto que eliminaremos de las variables dependientes toda la variabilidad que es predecible desde el covariado, está claro que la variabilidad residual de las dependientes será siempre igual o menor que la original. Esta característica hace recomendable que se midan covariados en situaciones en las que sospechamos que las variables dependientes van a presentar una alta variabilidad. Es importante caer en la cuenta de que esto es cierto tanto si nuestra manipulación es entre grupos como si es intrasuje-

jeto. Por otra parte, fundamentalmente en situaciones en las que el control experimental no puede ser riguroso, por tanto, en investigaciones de corte cuasi experimental, no puede conseguirse la equivalencia entre los grupos. Esto sucede, por ejemplo, cuando los grupos están preformados, como ocurre cuando se manipulan atributos de los sujetos (edad, nivel cultural, sexo, patología, etc.). La no equivalencia de los grupos impide establecer con firmeza suficiente a la manipulación como causa de las diferencias observadas en las variables dependientes. Una forma de mejorar la equivalencia consiste en utilizar covariados que permitan corregir las diferencias *a priori* entre los grupos en las variables dependientes.

En los diseños de covarianza tendremos, por tanto, tres tipos de variables, independientes, dependientes y covariados. Tanto las dependientes como los covariados deben ser variables métricas. Las independientes pueden ser métricas o categóricas. Cuando tenemos una sola dependiente y un solo variado una técnica de análisis adecuada es ANOVA. Cuando tenemos varias dependientes, haya un variado o más, una técnica que el investigador debe considerar es el análisis multivariado de covarianza (MANCOVA). Presentaremos esta técnica en el contexto de un ejemplo.

Supongamos que deseamos comparar dos métodos de enseñanza del álgebra, uno que implica el uso de nuevas tecnologías (hipertexto, etc.) y otro tradicional. Disponemos de dos grupos de un curso de bachillerato. Decidimos evaluar el aprendizaje con dos pruebas, una de problemas y otra teórica. También pensamos que sería adecuado realizar las dos pruebas antes de comenzar nuestra intervención y al finalizar la misma. Es importante caer en la cuenta de que la diferencia entre las mediciones posterior y anterior a la intervención indican la mejora obtenida por cada estudiante. Supongamos, por simplificar, que los grupos están compuestos por cinco sujetos cada uno, y que los datos obtenidos fueron los que aparecen en la tabla siguiente.

GRUPO	Antes Teoría	Antes Práctica	Después Teoría	Después Práctica
1	5	7	8	9
1	4	6	6	8
1	5	5	7	6
1	5	5	7	7
1	3	6	8	7
2	3	4	6	6
2	4	4	5	5
2	4	5	4	5
2	3	4	5	3
2	4	4	4	4

En este caso para poder explicar las variables dependientes (las mediciones en teoría y práctica después de la intervención, la matriz Y) debemos considerar no sólo los tratamientos introducidos, sino también los covariados (las

mediciones en teoría y práctica antes de la intervención, la matriz X). Por tanto, podemos escribir este modelo como:

$$Y = C + bX + A + \text{Error} , \quad (\text{Ec. 9.13})$$

donde C es la constante, bX es el efecto de los covariados y A el efecto de la intervención no predecible desde los covariados. La lógica del análisis reside en calcular los parámetros de la regresión de las dependientes sobre los covariados y eliminar de las dependientes lo que puede ser predicho por éstos. La variabilidad restante se debe a la intervención y al error, que pueden separarse de la misma manera que en el MANOVA ordinario. El análisis completo del diseño mediante SPSS se realiza incluyendo además de las variables independientes y dependientes, los covariados. La secuencia de comandos sería la siguiente:

Analizar → Modelo Lineal General → Multivariante → Independientes:  
 Grupo → Dependientes: D\_T, D\_P → Covariados: A\_T, A\_P → Aceptar  
 (análisis)

Lo fundamental de la salida que SPSS ofrece se encuentra en el Cuadro 9.18., que aparece en la página siguiente. En la parte superior se presenta el análisis sin considerar los covariados. La conclusión (errónea) a la que puede llegarse es que uno de los métodos parece superior al otro (véase columna Sig.). Sin embargo, cuando se introducen los covariados, claramente no hay diferencias significativas entre ambos métodos de enseñanza del álgebra. Obviamente, las diferencias apreciadas entre los grupos en las medidas posteriores a la intervención no se deben, en lo fundamental, a ésta, sino que son explicables desde los covariados.

Cuadro 9.18. Análisis multivariado de covarianza de un diseño con un factor entre grupos y dos covariados

1) Análisis multivariado de varianza sin covariados

Contrastes multivariados <sup>b</sup>						
Efecto		Valor	F	Gl de la hipótesis	Gl del error gl	Sig.
Intersección	Traza de Pillai	,987	273,708 <sup>a</sup>	2,000	7,000	,000
	Lambda de Wilks	,013	273,708 <sup>a</sup>	2,000	7,000	,000
	Traza de Hotelling	78,202	273,708 <sup>a</sup>	2,000	7,000	,000
	Raíz mayor de Roy	78,202	273,708 <sup>a</sup>	2,000	7,000	,000
GRUPO	Traza de Pillai	,777	12,165 <sup>a</sup>	2,000	7,000	,005
	Lambda de Wilks	,223	12,165 <sup>a</sup>	2,000	7,000	,005
	Traza de Hotelling	3,476	12,165 <sup>a</sup>	2,000	7,000	,005
	Raíz mayor de Roy	3,476	12,165 <sup>a</sup>	2,000	7,000	,005

<sup>a</sup> Estadístico exacto.

<sup>b</sup> Diseño: Intercept+GRUPO

2) Análisis multivariado de covarianza

Contrastes multivariados <sup>b</sup>						
Efecto		Valor	F	Gl de la hipótesis	Gl del error gl	Sig.
Intersección	Traza de Pillai	,521	2,714 <sup>a</sup>	2,000	5,000	,159
	Lambda de Wilks	,479	2,714 <sup>a</sup>	2,000	5,000	,159
	Traza de Hotelling	1,086	2,714 <sup>a</sup>	2,000	5,000	,159
	Raíz mayor de Roy	1,086	2,714 <sup>a</sup>	2,000	5,000	,159
A_T	Traza de Pillai	,184	,564 <sup>a</sup>	2,000	5,000	,601
	Lambda de Wilks	,816	,564 <sup>a</sup>	2,000	5,000	,601
	Traza de Hotelling	,226	,564 <sup>a</sup>	2,000	5,000	,601
	Raíz mayor de Roy	,226	,564 <sup>a</sup>	2,000	5,000	,601
A_P	Traza de Pillai	,411	1,747 <sup>a</sup>	2,000	5,000	,266
	Lambda de Wilks	,589	1,747 <sup>a</sup>	2,000	5,000	,266
	Traza de Hotelling	,699	1,747 <sup>a</sup>	2,000	5,000	,266
	Raíz mayor de Roy	,699	1,747 <sup>a</sup>	2,000	5,000	,266
GRUPO	Traza de Pillai	,517	2,673 <sup>a</sup>	2,000	5,000	,162
	Lambda de Wilks	,483	2,673 <sup>a</sup>	2,000	5,000	,162
	Traza de Hotelling	1,069	2,673 <sup>a</sup>	2,000	5,000	,162
	Raíz mayor de Roy	1,069	2,673 <sup>a</sup>	2,000	5,000	,162

<sup>a</sup> Estadístico exacto.

<sup>b</sup> Diseño: Intercept+A\_T+A\_P+GRUPO

### 3. LIMITACIONES Y SUPUESTOS DEL ANÁLISIS MULTIVARIADO DE VARIANZA Y COVARIANZA

El interés del investigador que emplea MANOVA/MANCOVA es realizar inferencias de naturaleza causal, donde los agentes causales potenciales son sus manipulaciones (las variables independientes) y los efectos son los cambios observados en las dependientes. Sin embargo, que estas inferencias puedan realizarse no es algo inherente a la técnica de análisis, sino al marco teórico en el que el investigador diseña su investigación. Las técnicas estadísticas permiten tomar decisiones sobre hipótesis, pero son silenciosas respecto de si se han seleccionado variables apropiadas, si el rango de manipulación ha sido adecuado, si se han medido de forma correcta, o si se han controlado otras variables que compitan con las independientes como agentes causales potenciales. Amén de estas limitaciones conceptuales, cuya resolución satisfactoria depende del grado de conocimiento que el investigador tenga del área en que su estudio se enmarca, hay una serie de limitaciones prácticas del análisis que deben ser evaluadas para que sus resultados sean, desde un punto de vista técnico, interpretables. A continuación presentaremos estas limitaciones prácticas y su posible solución. Puesto que algunas de estas limitaciones han sido ya tratadas en el contexto de otros análisis limitaremos su exposición en el sentido de recordar su significado fundamental.

#### 3.1. INDEPENDENCIA

MANOVA/MANCOVA asumen que las puntuaciones no pueden ser predichas teniendo en cuenta el orden temporal en que han sido obtenidas. Esto es, conocida la puntuación de un sujeto, no debe ser posible predecir la de los sujetos siguiente en el orden de pasación de las pruebas. El incumplimiento de este supuesto puede deberse a muchas razones. Por ejemplo, características de la situación en que se realizan las pruebas (por ejemplo, las instrucciones, la inexperiencia del investigador, etc.) pueden hacer que las puntuaciones de un conjunto de individuos estén correlacionadas. Como hemos visto ya (véase Capítulo VIII), una forma de intentar determinar si el supuesto se cumple consiste en examinar los coeficientes de autocorrelación. Si alguno de ellos resulta significativo puede ser conveniente diferenciar la serie de observaciones para eliminar la dependencia serial. También es conveniente tomar las decisiones con un nivel de significación más estricto. Esto es, si el nivel estándar es 0,05, podemos emplear 0,01 o menos, para asegurarnos que los efectos del incumplimiento sobre la distribución del estadístico de contraste no nos hacen cometer una tasa de error tipo I mayor de lo permitido.

#### 3.2. HOMOGENEIDAD DE LAS MATRICES DE VARIANZAS-COVARIANZAS

En el caso multivariado este supuesto implica que en cada condición experimental las matrices de varianzas-covarianzas son iguales. Imaginemos que hemos manipulado dos variables independientes entre grupos, ambas con tres niveles. Si las cruzamos factorialmente tendremos nueve condiciones experimentales, nueve grupos diferentes. Supongamos ahora que a cada sujeto le hemos medido tres variables dependientes. Entonces en cada uno de los nueve grupos podemos calcular la varianza de cada dependiente y la covarianza entre cada par de dependientes (en total tres covarianzas). El supuesto implica que las varianzas son iguales en los distintos grupos y las covarianzas también. Esto es, que la matriz de varianzas-covarianzas que obtenemos para el conjunto de los grupos (combinada) no es diferente de la matriz de cada grupo.

La exigencia del cumplimiento del supuesto tiene un sentido claro: el término error ( $\mathbf{W}$ ) que empleamos en un promedio de los errores de cada grupo. Si no hay homogeneidad, el error global no es representativo de los errores de grupo, lo que hace que el test estadístico sea erróneo. No es infrecuente que este supuesto se incumpla. Afortunadamente, sin embargo, como ocurre en ANOVA, las consecuencias no son graves si el número de sujetos por grupo es igual o aproximadamente igual. El test es robusto al incumplimiento en este caso. Sin embargo, cuando el número de sujetos es desigual el test no es robusto, por lo que es importante en ese caso cerciorarse de si el supuesto se cumple o no. Una forma de comprobarlo consiste en utilizar el test  $M$  de Box (la prueba de Levene es útil sólo para comprobar si las varianzas son iguales). En esencia, el test compara la matriz de varianzas-covarianzas combinada con las matrices de condición experimental. Cuando el test es significativo a  $p < 0,001$ , MANOVA/MANCOVA proporcionarán resultados incorrectos, en el sentido de que la tasa de error tipo I real es sensiblemente diferente de la tasa de error tipo I que seleccionamos para tomar la decisión (teórico). Sin embargo, la gravedad de la situación depende de la relación entre tamaño del grupo y tamaño de las varianzas-covarianzas. Cuando la relación es directa, mayores varianzas-covarianzas en los grupos de mayor tamaño, el error tipo I real es menor que el error teórico, por lo que la decisión puede tomarse con garantía y confianza. Sin embargo, cuando la relación es inversa, el error tipo I real es mayor que el teórico, hay que sopesar con cuidado la idea de admitir que hay diferencias entre los tratamientos.

¿Qué soluciones tenemos a nuestro alcance en casos graves de incumplimiento? Primero, una solución casera consiste en hacer los grupos iguales en tamaño eliminando sujetos de forma no sistemática, al azar. Segundo, es preferible utilizar como estadístico de contraste la traza de Pillai, dado que es más robusta que otros estadísticos como la lambda de Wilks. Tercero, puede intentarse una transformación (raíz cuadrada, logarítmica, etc.) con el objetivo de reducir las varianzas.

### 3.3. NORMALIDAD MULTIVARIADA

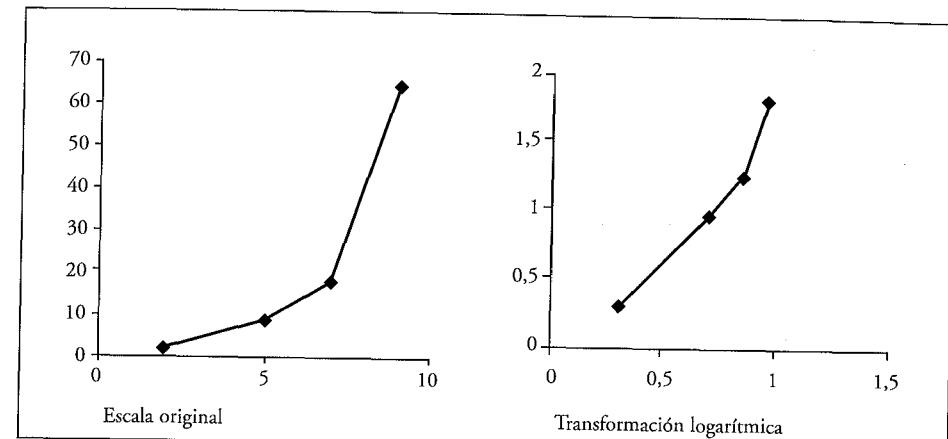
El supuesto implica que los datos se distribuyen según una normal multivariada. El significado del supuesto es triple. Primero, la combinación lineal de todas las variables dependientes debe distribuirse normalmente; segundo, la combinación lineal de cualquier subconjunto de las dependientes se distribuye también normalmente; tercero, cada una de las variables dependientes considerada aisladamente se distribuye normalmente. No hay una forma clara de comprobar este supuesto, por lo que la mayoría de los investigadores lo asumen en el tercer sentido. Así, si cada una de las variables se distribuye normalmente, se asume la normalidad multivariada. Aunque esta implicación es falsa, puesto que la normalidad de cada variable no garantiza la normalidad multivariada, sí es cierto que en ese caso, las consecuencias del incumplimiento son irrelevantes en la práctica. La comprobación del supuesto en el tercer sentido puede lograrse, como hemos mencionado en el Capítulo VIII, mediante las pruebas de Kolmogorov-Smirnov y/o Shapiro-Wilk.

### 3.4. LINEALIDAD

Las relaciones entre variables independientes y dependientes, entre covariados y dependientes y entre las propias dependientes se asumen lineales. Esto implica que todos los pares formables entre independientes y dependientes y todos los pares entre dependientes tiene relación lineal. El cumplimiento de este supuesto es, lógicamente, muy importante. La razón es sencilla: si la combinación de variables que realizamos es lineal, parece lógico pensar que la mezcla no conseguirá separar las condiciones experimentales cuando la relación es no lineal. Por otra parte, respecto de MANCOVA, si los covariados no tienen una relación lineal con las dependientes, la predicción será incorrecta y como consecuencia no se conseguirá reducir el error. En definitiva, el incumplimiento del supuesto acarrea como consecuencia que la potencia de MANOVA y MANCOVA se reduzca, con la consiguiente pérdida de información valiosa sobre el efecto real de nuestras manipulaciones. En el Capítulo VIII presentamos ya las técnicas para evaluar este supuesto. ¿Qué podemos hacer en caso de incumplimiento? La respuesta es que depende de en qué sentido se produzca. Si tenemos un covariado que tiene una relación no lineal con las dependientes, quizás la mejor solución sea eliminarlo. También podemos emplear esta estrategia cuando el problema reside en una variable dependiente. Sin embargo, antes de eso quizás debamos considerar la posibilidad de realizar una transformación para tratar de conseguir que la relación sea lineal. La Figura 9.2. presenta un ejemplo del efecto de una transformación logarítmica aplicada a dos variables cuya relación es no lineal (en realidad una variable es la otra al cuadrado). Obsérvese cómo, después de la transformación la relación es considerablemen-

te más lineal. Sin embargo, las transformaciones no siempre consiguen hacer lineal la relación y además son más difíciles de interpretar, por lo que es importante tener presente que una vez realizada, debe hacerse siempre referencia a ella y no a la puntuación original.

Figura 9.2. Efecto de una transformación logarítmica sobre la relación entre dos variables



### 3.5. MULTICOLINEALIDAD

Dos variables que tienen una correlación cercana a 1 proporcionan información redundante, puesto que una es, en la práctica, una combinación lineal de la otra. Por ejemplo, las variables Y1 (5, 6, 7) e Y2 (25, 35, 50) tienen una correlación de 0,993, y además podemos predecir casi a la perfección Y2 desde Y1, o viceversa (la ecuación de predicción de Y2 desde Y1 es:  $Y_2 = -38,3 + 12,5 Y_1$ ). Como indicábamos en el Capítulo VIII, el principal problema en este caso puede consistir en que no sea posible computar los determinantes de las matrices (sean singulares) y no pueda realizarse el análisis. Una solución, si no hay razones de peso para mantenerlas, puede ser eliminarlas de la ecuación. La estrategia consiste en fijar un nivel de tolerancia (el complementario del coeficiente de correlación múltiple) y excluir a las variables que tengan una tolerancia menor que la criterio. Cuando se desea mantenerlas, quizás la mejor opción sea emplear como variables dependientes las combinaciones lineales de éstas obtenidas tras un análisis de componentes principales.

### 3.6. HOMOGENEIDAD DE LOS COEFICIENTES DE REGRESIÓN

En MANCOVA, como en ANCOVA, un supuesto central es que los coeficientes de regresión computados en una condición experimental sean equivalentes a los obtenidos en cualquiera otra del estudio. Este supuesto es necesario debido a que las predicciones de las variables dependientes desde los covariados se realizan considerando un promedio de los coeficientes de regresión de las diferentes condiciones. Si los coeficientes no son homogéneos puede afirmarse que la relación entre covariados y dependientes es una función de las condiciones experimentales, es decir, que hay una interacción entre independientes y covariados. O dicho de otro modo, que las independientes afectan de alguna forma a los covariados. En este caso no debe emplearse MANCOVA.

Las alternativas disponibles en caso de incumplimiento del supuesto son muy variadas. Las más frecuentemente empleadas son las siguientes. Primero, transformar las variables dependientes substrayendo de ellas los covariados. Las puntuaciones de diferencia obtenidas pueden ser analizadas mediante un MANOVA ordinario. Segundo, cuando es posible, considerar a los covariados como un nivel de un factor repetido. Por ejemplo, en el caso del ejemplo de MANCOVA que hemos analizado en el apartado 2.4., bastaría con suponer que, además del factor entre grupos, tenemos un factor intrasujeto con dos niveles, medición pre y medición post-tratamiento. En cada nivel tenemos, por supuesto, las dos variables dependientes (evaluación teórica y práctica). Tercero, si el número de sujetos es suficiente podemos categorizar los covariados y convertirlos en un factor entre grupos más. Por ejemplo, podríamos considerar grupos diferentes a los que en la medida pre puntúen más de 7, a los que puntúen entre 5 y 7, y a los que puntúen por debajo de 5. De esta forma tendríamos dos factores entre grupos (método y puntuación pre) mezclados factorialmente y dos variables dependientes.

### 3.7. OBSERVACIONES EXTREMAS

MANOVA y MANCOVA son muy sensibles a puntuaciones extremas. Sus efectos sobre los estadísticos de contraste pueden alterar artificialmente las conclusiones sin que el investigador se percate, si no examina adecuadamente sus datos, de que está tomando decisiones erróneas. El programa Explorar de los Estadísticos Descriptivos de SPSS nos permite evaluar qué puntuaciones tenemos en nuestros datos que puedan considerarse extremas. La solución más recomendable en estos casos es su eliminación.

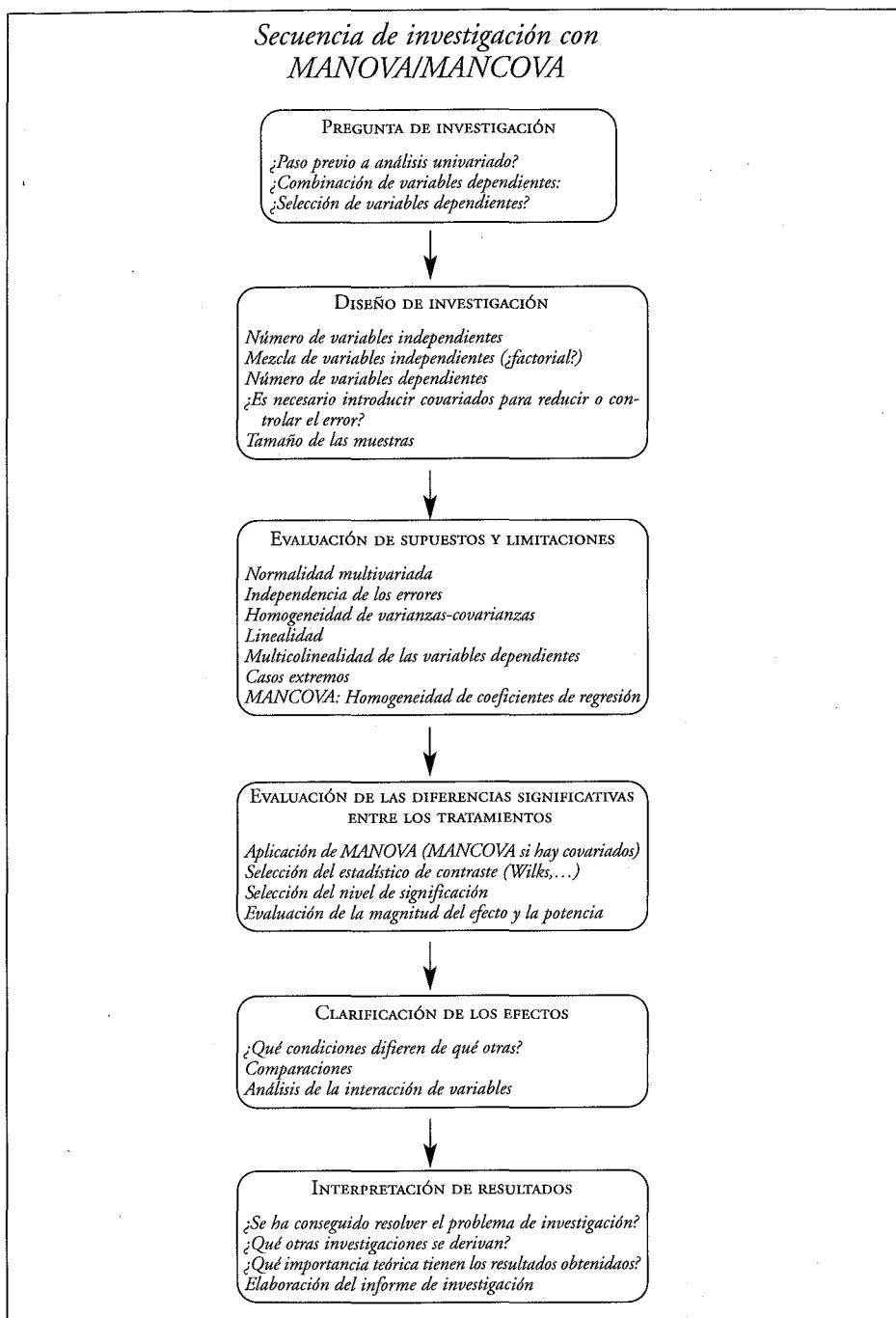
### 4. LA SECUENCIA DE INVESTIGACIÓN ORIENTADA AL ANÁLISIS MEDIANTE MANOVA/MANCOVA

La secuencia de análisis en MANOVA/MANCOVA aparece esquematizada en el Cuadro 9.19 (página siguiente). Como en el resto de técnicas de análisis es conveniente seguir un conjunto organizado de pasos para optimizar los resultados. En primer lugar es importante tener presente que esta técnica debe ser utilizada sólo cuando se deseé determinar si una o más variables independientes afectan a varias variables dependientes. Que las inferencias sean interesantes desde un punto de vista causal y teórico depende de la pregunta de investigación que se haya planteado. En cualquier caso, en ocasiones la técnica puede emplearse como un paso previo para realizar análisis de cada variable dependiente por separado, pero también a menudo el interés reside en estudiar la influencia de las manipulaciones sobre combinaciones de variables, dando lugar a comparaciones entre tratamientos en el marco de MANOVA/MANCOVA.

En segundo lugar, la elaboración del diseño de investigación tiene que desarrollarse de acuerdo con las exigencias que plantea la pregunta de la que partimos. Es importante seleccionar variables independientes adecuadas, determinar cómo van a ser manipuladas (su rango de valores, los niveles concretos que vamos a emplear, cómo se van a manipular (entre grupos, intrasujeto), cómo se van a mezclar: ¿de forma factorial?); variables dependientes sensibles a las manipulaciones, válidas, porque dependen de los procesos psicológicos que queremos estudiar y debemos prepararnos para medirlas de forma fiable. En el caso de que sea necesario introducir covariados es importante que no sean afectados por la manipulación y que correlacionen con las variables dependientes. Finalmente, es preciso elegir un número de sujetos suficiente para garantizar que los test estadísticos van a tener potencia suficiente para detectar efectos de la manipulación si los hubiere.

En tercer lugar, para garantizar que las inferencias basadas en el test estadístico se realizan con tasas de error tipo I y de potencia aceptables es importante evaluar los supuestos y las limitaciones de MANOVA/MANCOVA. Especial atención debe dedicarse a la inspección de los datos en busca de casos extremos, y al incumplimiento de los supuestos de independencia, de linealidad y multicolinealidad y de homogeneidad de varianzas-covarianzas y de coeficientes de regresión (en MANCOVA). A continuación puede aplicarse la prueba estadística. Si hemos manipulado más de una variable independiente esperaremos una interacción significativa. La evaluación de esa interacción puede hacerse mediante comparaciones multivariadas, pero también podemos realizarla reduciendo el diseño a un conjunto de univariados. Cuál sea la opción que adoptamos depende de la pregunta de investigación que pretendemos responder. Las diferencias observadas entre las condiciones experimentales nos servirán, finalmente para interpretar los resultados dotándolos de contenido teórico.

Cuadro 9.19. Preparación de la investigación orientada al análisis mediante MANOVA/MANCOVA



## 5. LECTURAS RECOMENDADAS

- HAIR, J. F. y cols., *Análisis multivariable*, Madrid, Prentice-Hall, 1998-1999.  
 HUBERTY, C. J. y MORRIS, J. D., «Multivariate versus multiple univariate analyses», *Psychological Bulletin*, 105, 1989, págs. 302-308.  
 O'BRIAN, R. G. y KAISER, M. K., «MANOVA method for analyzing Repeated Measures Design: An extensive primer», *Psychological Bulletin*, 97, 1985, págs. 316-333.  
 PASCUAL, J.; FRÍAS, D. y GARCÍA, F., *Manual de Psicología Experimental*, Barcelona, Ariel, 1996.  
 RIBA, M. D., *Modelo lineal de análisis de la varianza*, Barcelona, Herder, 1990.  
 TABACHICK, B. G. y FIDELL, L. S., *Using Multivariate Statistics*, Boston, MA, Allyn & Bacon, 2001.

## LECTURAS COMPLEMENTARIAS

- BIRD, K. D. y HADZI-PAVLOVIC, D., «Simultaneous test procedures and the choice of a test statistics in MANOVA», *Psychological Bulletin*, 93, 1983, págs. 167-178.  
 BRAY, J. H. y MAXWELL, S. E., *Multivariate analysis of variance*, Beverly Hills, CA, Sage, 1985.  
 DAVIDSON, M. L., «Univariate versus multivariate tests in repeated measures experiments», *Psychological Bulletin*, 77, 1972, págs. 446-452.  
 DILLON y GOLDSTEIN, *Multivariate Análisis*, NY, Wiley, 1984.  
 GABRIEL, K. R., «Simultaneous test procedures in multivariate analysis of variance», *Biometrika*, 55, 1968, págs. 389-504.  
 HAND, D. J. y TAYLOR, C. C., *Multivariate analysis of variance and repeated measures*, Londres, Chapman and Hall, 1987.  
 OLSON, C. L., «On choosing a test statistic in multivariate analyses of variance», *Psychological Bulletin*, 83, 1976, págs. 579-586.  
 — «Practical considerations in choosing a MANOVA test statistics: A rejoinder to Stevens», *Psychological Bulletin*, 86, 1979, págs. 1350-1352.  
 SEDLMEIER, P. y GIGERENZER, G., «Do studies of statistical power have an effect on the power of studies?», *Psychological Bulletin*, 105, 1989, págs. 309-316.  
 SERLIN, R. C., «A multivariate measure of association based on the Pillai-Bartlett procedure», *Psychological Bulletin*, 91, 1982, págs. 413-417.  
 SHARMA, S., *Applied Multivariate Techniques*, NY, Wiley, 1996.  
 STEVENS, J. P., «Power of the multivariate analysis of variance test», *Psychological Bulletin*, 88, 1980, págs. 728-737.

## 6. EJERCICIOS

Considere los dos conjuntos de datos que aparecen en las tablas siguientes.

Grupo 1		Grupo 2	
Y1	Y2	Y1	Y2
2	4	6	12
4	7	7	17
5	10	9	18
6	11	11	19

Grupo 1		Grupo 2	
Y1	Y2	Y1	Y2
2	6	7	11
3	7	8	10
6	3	10	5
6	3	11	8
4	2	10	6
4	5	8	5

1. ¿Cuál es la hipótesis nula y cuál la hipótesis alternativa en cada caso?
2. ¿Hay efectos significativos en alguno de los dos casos?
3. ¿Qué estadístico de contraste seleccionaría?
4. ¿Habrá alguna diferencia entre una aproximación univariada y otra multivariada?

### RESPUESTAS A LOS EJERCICIOS

1. La hipótesis nula es que los centroides de grupo no difieren. La alternativa es precisamente que sí lo hacen. Ambas hipótesis se aplican a los dos casos.
2. Las hay en el primero, por ejemplo,  $\Lambda=0,266$ ,  $F(2,5)=6,885$ ,  $p<0,04$ . Y también en el segundo,  $\Lambda=0,118$ ,  $F(2,5)=33,656$ ,  $p<0,01$ .
3. El estadístico de contraste es irrelevante en situaciones como la presente, en la que tenemos dos grupos y dos variables dependientes.
4. La aproximación univariada permite tomar una decisión sobre la diferencia entre grupos en cada variable dependiente. En este caso, puede rechazarse la hipótesis nula en ambos casos.

## CAPÍTULO X

### Análisis discriminante

#### OBJETIVOS DE APRENDIZAJE

- 1) Conocer en qué contextos y para qué sirve el análisis discriminante.
- 2) Distinguir entre análisis discriminante y otras técnicas de predicción como la regresión múltiple.
- 3) Conocer qué tipos de preguntas de investigación permiten resolver el análisis discriminante.
- 4) Conocer cómo se realiza el análisis discriminante y cómo se determinan los pesos de las variables independientes.
- 5) Distinguir la función discriminante lineal de la función de clasificación lineal.
- 6) Saber evaluar el poder clasificadorio de las funciones obtenidas,
- 7) Saber interpretar los resultados del análisis discriminante.
- 8) Conocer las limitaciones y supuestos del análisis discriminante.

#### 1. CONCEPTOS FUNDAMENTALES DEL ANÁLISIS DISCRIMINANTE

La característica más relevante de la actividad científica es sin duda la capacidad de predecir unas variables a partir de otras. Prácticamente toda la actividad científica está encaminada a descubrir relaciones entre variables que permitan predecir y controlar el medio. Es sin duda esto lo que ha hecho que la regresión múltiple sea una de las técnicas de análisis más extendidas. Como

hemos podido comprobar es una técnica idónea cuando pretendemos predecir una variable métrica a partir de un conjunto de variables predictoras también métricas. Sin embargo, en muchas investigaciones no estamos interesados en predecir variables métricas, sino justamente variables no métricas, categóricas. En este contexto está claro que la regresión múltiple no es útil, puesto que llevaría a predicciones absurdas. Por ejemplo, supongamos que deseamos poner a punto un test para estudiar el pensamiento en personas que padecen esquizofrenia paranoide. Lógicamente una de las cosas más importantes que el test debe hacer es distinguir adecuadamente entre este tipo de pensamiento y otros que también pueden presentar ciertas características alteradas, como puede ser el de personas alcoholizadas o personas con una personalidad límite. Así, diseñamos la prueba con un número determinados de ítem y la aplicamos a tres grupos de personas: esquizofrénicos paranoides, alcohólicos y personalidad límite. Lógicamente estamos interesados en saber si en el test los tres grupos difieren. El hecho de que difieran es fundamental porque significa que valiéndonos del test podemos clasificar a las personas como miembros de uno u otro grupo. Y ésta es la cuestión clave: clasificar. Lo que a nosotros nos interesa no es tanto la significación cuánto utilizar el test como un instrumento de clasificación, de diagnóstico. Si, de manera errónea, intentásemos predecir el grupo a partir de los ítem del test con seguridad nos llevaríamos una gran sorpresa, predeciríamos grupos inexistentes. Es justamente en este contexto donde el análisis discriminante es útil: cuando deseamos estudiar la relación entre una variable no métrica y un conjunto de variables métricas.

El análisis discriminante es una técnica idónea cuando el interés de la investigación consiste en clasificar a individuos en grupos de los cuales se conoce el criterio de clasificación de antemano. Por ejemplo, ¿podemos clasificar valiéndonos de las respuestas a una encuesta a los votantes de uno y otro partido? En este caso conocemos (porque los encuestados lo declaran en uno de los ítem) si los votantes son miembros de un grupo (votantes de un partido) u otro (votantes de otro partido). ¿Podemos clasificar correctamente a niños como pertenecientes a un grupo que tiene problemas de aprendizaje, a un grupo con problemas de motivación o a un grupo de niños sin problemas en función de sus respuestas a test de inteligencia y personalidad? De nuevo conocemos el criterio de clasificación de los niños en grupos. Sabemos de antemano quiénes son los niños con dificultades y quiénes no. Si comparamos los grupos teniendo en cuenta simultáneamente todos los ítem es obvio que estamos realizando un análisis multivariado de varianza. El análisis discriminante y el MANOVA son prácticamente iguales en este punto. Es más MANOVA es un paso previo para aplicar análisis discriminante. La razón es obvia, si los grupos no difieren difícilmente podremos clasificar correctamente a los individuos en un grupo u otro. Ahora bien, hay una cuestión semántica que los diferencia de manera importante. Desde la perspectiva de MANOVA las variables independientes son los grupos y las dependientes son los ítem. En análisis discriminante es justo al revés: la variable que predecimos es el grupo (por tanto es la depen-

diente) y los predictores (y por tanto las variables independientes) son los ítem. En pocas palabras, la diferencia reside en qué variables son las predichas y cuáles los predictores.

Cuando se intenta clasificar a individuos en grupos podemos poner el énfasis en dos cuestiones diferentes. Por un lado, podemos interesarnos solamente por encontrar una regla para realizar la clasificación lo más correctamente posible. En este caso estamos interesados en determinar lo que se denominan funciones de clasificación lineal. Nos importa poco o nada el valor que los diferentes ítem puedan tener para separar unos grupos de otros. Por otro lado, podemos enfatizar más la idea de que la separación entre los grupos puede deberse más a unos ítem que otros. Ahora el interés recae en lo que se llaman funciones discriminantes lineales. Nos interesa más estudiar la contribución de cada ítem a la predicción del grupo. Es importante no confundir ambas aproximaciones porque tanto sus fundamentos como su aplicación son bastante diferentes.

En este capítulo presentaremos de manera intuitiva y con el menor número de formalismos posible los fundamentos, el análisis y la interpretación de resultados del análisis discriminante. En primer lugar estudiaremos sus fundamentos y la clase de cuestiones de investigación que nos permite resolver. Es importante dejar bien sentado que el análisis discriminante no puede aplicarse en diseños de medidas repetidas, ni tampoco fácilmente en diseños factoriales entre grupos. Dada la semejanza con MANOVA es preferible plantear el análisis de esos diseños fuera del marco del análisis discriminante. A continuación presentaremos el análisis en sí, haciendo referencia al cálculo de las funciones discriminantes, a los tipos de análisis discriminante y a las funciones de clasificación. Finalmente abordaremos sus supuestos y limitaciones.

## 2. PREGUNTAS QUE EL ANÁLISIS DISCRIMINANTE PUEDE AYUDAR A RESOLVER

El análisis discriminante tiene un objetivo fundamental: separar grupos y encontrar funciones que permitan clasificar correctamente a individuos en los grupos. Las funciones pueden entenderse como ecuaciones de regresión: son combinaciones lineales de los predictores. Como ya sabemos la combinación lineal es una suma ponderada de variables. Y también sabemos ya que el número de combinaciones lineales ortogonales (independientes entre sí) que podemos formar depende del número de grupos y del número de predictores que tengamos. Como veremos más adelante, esto puede complicar la interpretación, pero recordemos por ahora que cuando tenemos tres grupos una combinación lineal (una función) puede separar a dos grupos y la otra a esos del tercero. El éxito de la predicción depende de un factor fundamental, de las variables predictoras que hayan sido seleccionadas. La elección inicial de variables es algo que no puede guiar el análisis, sino que debe realizarse en función

de un marco teórico o conceptual, y/o pragmático (qué variables pueden medirse realmente, qué costo tiene medir cada variable, etc.).

Consideremos antes de continuar un ejemplo que nos ayude a comprender qué cuestiones puede ayudarnos a resolver el análisis discriminante. Sigamos con el caso del pensamiento en esquizofrénicos paranoides. Supongamos que elaboramos un test que, por simplificar, tiene cuatro escalas de ítem. Creemos que las cuatro sirven para diagnosticar, para discriminar entre paranoides, alcohólicos y personalidad límite (cinco sujetos por grupo). Imaginemos los siguientes resultados:

Paranoides				Alcohólicos				Personalidad Límite			
E1	E2	E3	E4	E1	E2	E3	E4	E1	E2	E3	E4
5	8	7	2	7	4	6	8	4	4	1	6
4	9	6	5	5	4	7	9	8	2	3	5
6	7	6	3	8	2	6	6	7	3	2	5
8	8	9	2	6	3	5	9	5	2	1	4
7	8	7	3	4	2	6	8	6	4	3	5

¿Qué clase de cuestiones podemos responder sobre estos datos? En primer lugar todo depende de nuestro objetivo fundamental. Si pretendemos examinar la importancia que cada escala tiene para predecir la pertenencia a los grupos, es decir, si estamos interesados en realizar un análisis discriminante descriptivo, claramente las cuestiones serían diferentes de si pretendemos clasificar a los sujetos en grupos, es decir, si deseamos realizar un análisis discriminante predictivo.

## 2.1. CUESTIONES QUE AYUDA A RESPONDER EL ANÁLISIS DISCRIMINANTE DESCRIPTIVO

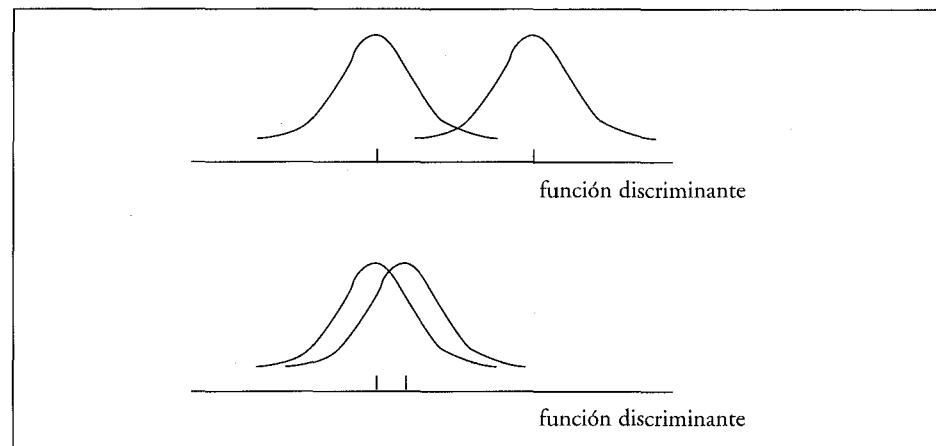
En el análisis discriminante descriptivo el interés se centra en combinar linealmente las variables predictoras (las escalas en nuestro ejemplo) para predecir el grupo. Las funciones discriminantes tienen la forma general siguiente:

$$Y_{ij} = b_0 + b_1 X_{i1} + b_2 X_{i2} + \dots + b_p X_{ip} \quad (\text{Ec. 10.1})$$

donde el primer subíndice denota al sujeto y el segundo denota a la función discriminante (una de las combinaciones posibles de variables independientes). Y es la puntuación discriminante resultante de la suma ponderada de las variables predictoras y los b son los pesos de cada una de ellas. Así, aplicando estas funciones tendremos un conjunto de puntuaciones por grupo. El promedio de las puntuaciones de grupo es el centroide de grupo.

En primer lugar podemos preguntarnos si las medias de grupo en la puntuación discriminante (los centroides) son iguales entre sí. Esta hipótesis nula establece sencillamente que no es posible predecir el grupo de pertenencia a partir de los predictores. Ésta es la cuestión fundamental, pues si se resuelve negativamente el resto de cuestiones carecen de sentido. Observemos en la Figura 10.1. cómo dos grupos pueden ser discriminados claramente en una función discriminante (arriba) o no (abajo) en función de la separación de sus centroides. Cuando la predicción del grupo es significativa estamos afirmando lo mismo que si concluimos que los grupos son diferentes entre sí. En este caso nuestro interés reside ya en las siguientes cuestiones, que se resolverán mediante MANOVA y el análisis de las funciones discriminantes lineales.

Figura 10.1. Separación de dos grupos en una función discriminante lineal



En segundo lugar puede interesarnos saber cuántas funciones discriminantes podemos construir y en cuántas de ellas se discrimina a unos grupos de otros. Cuando tenemos tres grupos (y más de dos variables predictoras) es posible definir dos funciones discriminantes. Los grupos pueden diferir en ambas o en ninguna. Qué función discrimina entre unos grupos y otros es lo que nos interesa resolver inmediatamente. Puede ocurrir, por ejemplo, que la primera función separe al grupo paranoide de los otros dos, que no difieren entre sí en ésta. La segunda puede separar a los grupos alcohólico y personalidad límite. Es importante tener presente que la potencia de separación de las funciones discriminantes depende de su orden. La primera función extraída tiene mayor poder, después la segunda, y así hasta la última. El número total de funciones (s) es el menor de dos números: el número de grupos (g) menos 1 o el número de variables predictoras (p), formalmente  $s = \min((g-1), p)$ . De este modo si la primera función no logra discriminar entre los grupos, ninguna de las demás lo hará.

Una cuestión que podemos plantearnos de manera simultánea con la anterior es si todas las variables predictoras son necesarias para predecir la pertenencia a los grupos. Es posible que algunas no realicen ninguna contribución relevante y puedan ser eliminadas de la función. Si examinamos los datos de nuestro ejemplo podremos intuir fácilmente que la primera escala probablemente no contribuya en absoluto a la predicción, o dicho de otro modo, la predicción del grupo sea igual de buena incluyendo esa escala que eliminándola.

Una pregunta relacionada con la anterior es si las variables que contribuyen a la predicción pueden ordenarse de alguna forma en función de su importancia. ¿Qué variables son las que más contribuyen y cuáles las que menos a la separación de los grupos? Naturalmente este enfoque de la importancia no es único (Huberty y Wisenbaker, 1992), pero sí es el más relevante para aproximarse a otro problema fundamental del análisis: ¿cómo interpretar las funciones discriminantes significativas? La respuesta a esta cuestión dependerá de la contribución de cada variable a la función. Por ejemplo, si nuestra escala 1 mide delirio y no contribuye de forma significativa a ninguna función, nuestra interpretación será muy diferente de si contribuye de forma relevante a una función discriminante que separa a los paranoides del resto de los grupos.

## 2.2. CUESTIONES QUE AYUDA A RESOLVER EL ANÁLISIS DISCRIMINANTE PREDICTIVO

Como hemos señalado antes, en muchas ocasiones estamos más interesados en clasificar simplemente a los sujetos en grupos que en las variables que podemos emplear para ello. Por ejemplo, nos interesa saber cómo asignar individuos a uno de los tres grupos porque pretendemos usar nuestro test como un instrumento diagnóstico, que debe ser aplicado a personas de las cuales no sabemos si pertenecen a un grupo u otro. En este caso las cuestiones relevantes son diferentes.

En primer lugar lo que nos interesa es encontrar un conjunto de funciones (de hecho tendremos una por grupo incluido en el estudio) que nos permitan clasificar nuevos casos en uno de los grupos. La idea es cómo combinar la información proporcionada por las diferentes escalas para acertar lo máximo posible en la clasificación de nuevos individuos. Naturalmente lo relevante a continuación es preguntarse acerca de la bondad de la clasificación. ¿Qué proporciones de casos antiguos (estudiados en la investigación) son clasificados correctamente en su grupo de pertenencia? Esto es, ¿qué proporción de paranoides son clasificados correctamente como paranoides? ¿Es significativo nuestro acierto en la clasificación? En el caso de cometer errores ¿qué patrón siguen? Los paranoides incorrectamente asignados, ¿son adjudicados más al grupo de alcohólicos o al de personalidad límite?

En las páginas siguientes presentaremos primero los conceptos fundamentales del cálculo del análisis discriminante descriptivo y a continuación del

predictivo. Después haremos referencia a las limitaciones del análisis y a sugerencias para el diseño de investigaciones que pretendan usar esta técnica de análisis.

## 3. ANÁLISIS DISCRIMINANTE DESCRIPTIVO

El análisis discriminante descriptivo se divide en dos partes fundamentales que deben ordenarse en su aplicación. En primer lugar es conveniente determinar si los grupos difieren significativamente entre sí considerando el conjunto de medidas que hemos tomado de los sujetos. Este análisis es seguido por otro encaminado a la obtención de las funciones discriminantes lineales y a la evaluación de la importancia de cada variable en cada función discriminante.

### 3.1. DETERMINACIÓN DE LAS DIFERENCIAS SIGNIFICATIVAS ENTRE LOS GRUPOS

En este sentido MANOVA y análisis discriminante son idénticos. De hecho, MANOVA es la técnica recomendable para decidir si los centroides de grupo son iguales entre sí o no. Puesto que hemos dedicado el Capítulo IX a esta cuestión, la resumiremos brevemente. Recordemos que la ecuación fundamental de MANOVA para diseños de un solo factor manipulado entre grupos es la siguiente:

$$T = B + W$$

donde  $T$  es la matriz que contiene las sumas de cuadrados y productos cruzados totales,  $B$  es la matriz que contiene las sumas de cuadrados y productos cruzados de tratamiento y  $W$  es la matriz de sumas de cuadrados y productos cruzados intragrupo o error.

Un estadístico ampliamente utilizado en este contexto es la lambda de Wilks, que, como se recordará, es un cociente entre los determinantes de  $W$  y  $T$ , y puede interpretarse como la proporción de variabilidad total que es explicada por la variabilidad error. Una forma sencilla de obtener este estadístico consiste en utilizar el programa de análisis multivariado de SPSS. La secuencia de comandos aplicada a nuestro ejemplo sería la siguiente:

Analizar → Modelo Lineal General → Multivariante → Dependientes: E1, E2, E3, E4 → Factores fijos: Grupo → Aceptar (análisis)

Los resultados proporcionados por SPSS aparecen en el Cuadro 10.1. Lo que nos interesa en este punto es constatar simplemente si podemos rechazar o no la hipótesis nula que establece que los grupos no difieren entre sí (que sus

centroides son iguales). Para ello, consultamos la columna Sig. Observemos que efectivamente la probabilidad de cualquiera de los estadísticos está por debajo del límite estándar (0,05) y que, por tanto, podemos rechazar la hipótesis nula. Al menos debe haber una forma de combinar las variables predictivas (las escalas en nuestro ejemplo) que nos permita predecir el grupo de pertenencia. El paso siguiente del análisis es, pues, determinar las funciones discriminantes lineales.

Cuadro 10.1. Análisis discriminante descriptivo. Significación de las diferencias entre grupos

Contrastes multivariados <sup>c</sup>						
Efecto		Valor	F	Gl de la hipótesis	Gl del error gl	Sig.
Intercept	Traza de Pillai	,992	276,841 <sup>a</sup>	4,000	9,000	,000
	Lambda de Wilks	,008	276,841 <sup>a</sup>	4,000	9,000	,000
	Traza de Hotelling	123,041	276,841 <sup>a</sup>	4,000	9,000	,000
	Raíz mayor de Roy	123,041	276,841 <sup>a</sup>	4,000	9,000	,000
GRUPO	Traza de Pillai	1,855	31,948	8,000	20,000	,000
	Lambda de Wilks	,004	35,499 <sup>a</sup>	8,000	18,000	,000
	Traza de Hotelling	38,856	38,856	8,000	16,000	,000
	Raíz mayor de Roy	31,082	77,705 <sup>b</sup>	4,000	10,000	,000

<sup>a</sup> Estadístico exacto.  
<sup>b</sup> El estadístico es un límite superior para la F el cual ofrece un límite inferior para el nivel de significación.  
<sup>c</sup> Diseño: Intercept+GRUPO

### 3.2. FUNCIONES DISCRIMINANTES LINEALES

La ecuación 10.1. indica claramente que una función discriminante lineal puede interpretarse como una ecuación de regresión múltiple en el sentido de que se trata de predecir una puntuación a partir de una suma ponderada de predictores. Sin embargo, las sumas ponderadas que podemos realizar son virtualmente infinitas, puesto que, en principio, cualquier conjunto de pesos puede ser empleado legítimamente. Sin embargo, nos interesan tres cosas fundamentales. Primero, el número de funciones debe ser el mínimo posible. Esto es lógico, si con una función podemos separar los grupos, el resto de funciones dificultan la interpretación de los resultados y no añaden nada nuevo a la que ya tenemos. Segundo, las funciones discriminantes no deben compartir varianza, deben ser ortogonales, no correlacionadas entre sí. Este requisito también es obvio, puesto que si una función ya discrimina entre dos grupos ¿qué necesidad tenemos de otra función que también discrimine entre ellos?

En tercer lugar, es fundamental que los pesos sean seleccionados en función de un objetivo prioritario: separar lo máximo posible a los grupos. Para comprender cómo se logra este objetivo debemos recordar dos cuestiones importantes. Primero,

### Análisis discriminante

cuando combinamos variables obtenemos variados, que en este contexto se llaman variados canónicos o funciones discriminantes canónicas. Segundo, el cociente entre la variabilidad entre grupos ( $SC_{EG}$ ) y la variabilidad error ( $SC_{IG}$ ) en un variado es un valor propio de la matriz  $W^{-1}B$ . Tengamos en cuenta que el producto de estas dos matrices es equivalente a dividir la segunda por la primera (Entre grupos por intragrupos). Así, el objetivo se reduce a maximizar los valores propios de la matriz  $W^{-1}B$ . Para conseguir esto es preciso seleccionar un conjunto de pesos que separen máximamente los grupos. Puesto que se trata de un conjunto ordenado de pesos, podemos hablar de un vector de pesos, llamémosle  $b$ , y recordemos que  $\lambda$  es el conjunto de valores propios. Entonces, nuestro objetivo fundamental puede lograrse resolviendo la siguiente ecuación:

$$(W^{-1}B - \lambda I) b = 0 \quad (\text{Ec. 10.2})$$

Cuadro 10.2. Obtención de valores propios y de los pesos no estandarizados correspondientes a cada función discriminante

1) La secuencia de comandos fue:				
Analizar → Clasificar → Discriminante → Variable de agrupamiento: Grupo → Definir rango → Mínimo: 1; Máximo: 3 → Continuar → Estadísticos → Coeficientes de la función: No tipificados → Continuar → Aceptar (análisis)				
2) Valores propios para cada función discriminante				
Autovalores				
Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	31,082a	80,0	80,0	,984
2	7,774a	20,0	100,0	,941
3) Coeficientes no estandarizados para cada función discriminante				
Coeficientes de las funciones canónicas discriminantes				
	Función		1	2
E1			,127	-,278
E2			-1,303	-,077
E3			,246	1,168
E4			1,162	,338
(Constante)			-2,114	-5,610
Coeficientes no tipificados				

Aunque pueda parecer compleja, en realidad el cálculo que la ecuación implica es sencillo. Puede examinarse el detalle en el Cuadro 10.2, donde se detalla primero cómo obtener mediante SPSS los valores propios (véase Capítulo IX) y el

conjunto de pesos que corresponde a cada valor propio, esto es, para cada función discriminante. A continuación podríamos obtener la puntuación de cada sujeto en cada función discriminante, sin embargo, debemos caer en la cuenta de que estos coeficientes no están tipificados y, por tanto, su valor está reflejando en parte la escala en la cual hemos medido la variable. Es preferible estandarizarlos, de manera que su valor refleje más la influencia real de la variable predictora para definir la función discriminante. SPSS nos proporciona por defecto los coeficientes discriminantes canónicos estandarizados que aparecen en el Cuadro 10.3. para los datos de nuestro ejemplo. Obsérvese que los coeficientes cambian (no mucho) a pesar de que todas las variables aparentan estar en la misma escala.

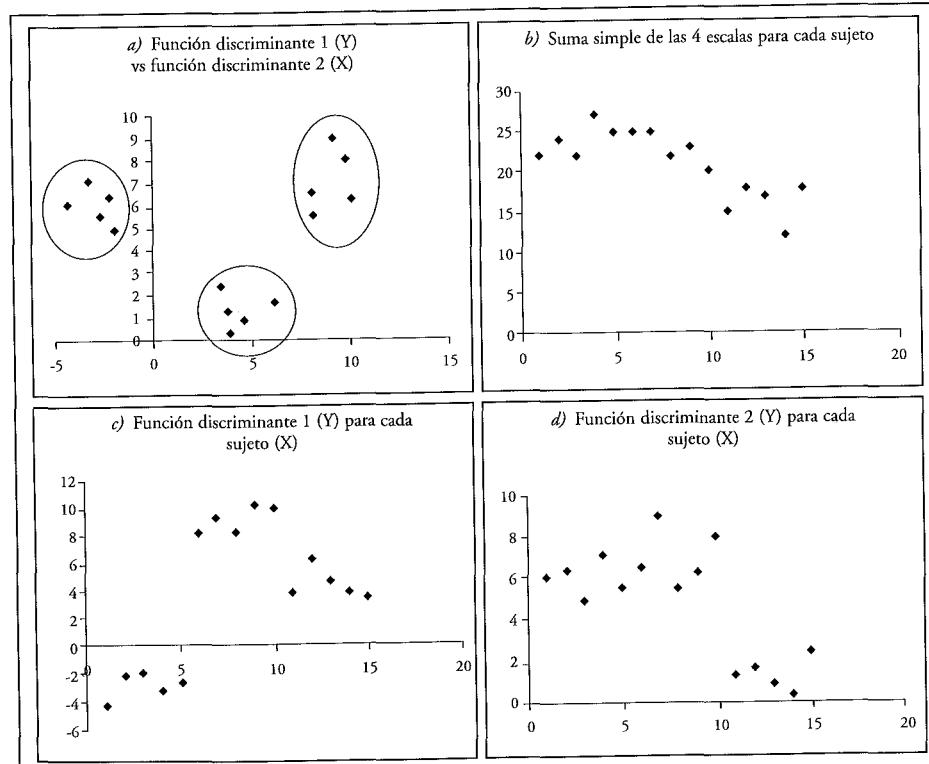
Las funciones discriminantes serían las siguientes:

$$Y_{11} = 0,201 E1 - 1,189 E2 + 0,246 E3 + 1,256 E4 \text{ (la primera función)}$$

y

$$Y_{12} = -0,44 E1 - 0,070 E2 + 1,168 E3 + 0,365 E4 \text{ (la segunda función)}$$

Figura 10.2. Representación gráfica del poder de las funciones discriminantes lineales respecto a la suma simple



Si las aplicamos para cada sujeto de cada uno de los grupos tendríamos un conjunto de puntuaciones que nos podrían permitir discriminar entre los grupos. La Figura 10.2. presenta los datos de los 15 sujetos en cada función discri-

minante. Con fines de comparación se ha incluido también la representación de datos siguiendo una estrategia bastante estándar de sumar simplemente los datos de cada sujeto correspondientes a cada escala. Obsérvese cómo en la suma simple no parece posible discriminar entre los grupos paranoide y alcohólico, mientras que sí a estos dos respecto del tercero, sin embargo, la función discriminante primera parece separar a los paranoides del resto mientras que la segunda separa a los sujetos de personalidad múltiple del resto. También se ha incluido un gráfico de dispersión (Figura 10.2.a) para que el lector pueda valorar cómo las dos funciones discriminantes conjuntamente contribuyen a separar los grupos de manera completa. Nótese cómo las nubes de puntos que representan a cada grupo están localizadas en posiciones diferentes del plano xy. Una evaluación visual no es, sin embargo suficiente para establecer con suficiente confianza si cada función discrimina o no de manera significativa entre los grupos. Podemos tomar la decisión basándonos en dos pruebas diferentes. En primer lugar rea- lizando un análisis de varianza entre grupos para cada función discriminante.

Cuadro 10.3. Coeficientes estandarizados de las funciones discriminantes

	Coeficientes estandarizados de las funciones discriminantes canónicas	
	1	2
E1	.201	-.440
E2	-1.189	-.070
E3	.246	1.168
E4	1.256	.365

Este análisis puede ser continuado mediante comparaciones dentro del marco de ANOVA (véase Cuadro 10.4., en página siguiente). SPSS proporciona por defecto una forma alternativa de evaluar si cada función contribuye de forma significativa a separar los grupos mediante la lambda de Wilks (véase Cuadro 10.5., en página siguiente). En este caso se adopta una estrategia por etapas, primero se evalúa la primera función, para evaluar la segunda se descarta la aportación de la primera y así sucesivamente. Lo importante es que el número de funciones discriminantes que debemos retener para interpretar los resultados depende de su significación. Sólo retendremos las que hacen una aportación fiable a la discriminación entre grupos. Las dos formas producen resultados equi- valentes.

Cuadro 10.4. Análisis de las funciones discriminantes

1) ANOVA					
	Suma de Cuadrados	gl	Media cuadrática	F	Sig.
Y1	Intergrupos 12,022 Intragrupos 375,867	2 12 14	181,923 1,002	181,597	,000
Y2	Intergrupos 13,108 Intragrupos 107,991	2 12 14	47,422 1,092	43,432	,000

2) Comparaciones entre grupos

2.1) Coeficientes de las comparaciones (1: paranoides vs alcohólicos 2: paranoides + alcohólicos frente a personalidad límite).

Contraste	GRUPO		
	Paran.	Alcoh.	Pers Lim.
1	1	-1	0
2	-1	-1	2

2.2) Comparaciones planeadas

Pruebas de contraste

	Contraste	Valor de contraste	Error típico	t	gl	Sig. (bilateral)
Y1	Suponer igualdad de varianzas	1 -11,97900 2 2,4750000	,63302296 1,0964279	-18,923 2,257	12 12	,000 ,043
	No asume igualdad de varianzas	1 -11,97900 2 2,4750000	,59647154 1,1562757	-20,083 2,140	7,996 6,968	,000 ,070
Y2	Suponer igualdad de varianzas	1 -1,007000 2 -10,52700	,66100575 1,1448955	-1,523 -9,195	12 12	,154 ,000
	No asume igualdad de varianzas	1 -1,007000 2 -10,52700	,72573797 1,0205316	-1,388 -10,315	6,497 9,956	,211 ,000

Cuadro 10.5. Evaluación de la significación de las funciones discriminantes lineales

Lambda de Wilks				
Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1 a la 2	,004	59,221	8	,000
2	,114	22,804	3	,000

### 3.3. LA IMPORTANCIA DE CADA VARIABLE PREDICTORA: LA MATRIZ DE ESTRUCTURA

A menudo nuestro interés recae en determinar qué variables son realmente útiles para discriminar unos grupos de otros. Esto es importante en varios sentidos. Primero, cuando se desea determinar la contribución de la variable a la definición del constructo teórico (esto es, etiquetar la función discriminante). Segundo se pretende estudiar la contribución relativa de cada variable a la función. Y tercero cuando se pretende determinar si una variable es necesaria para los efectos entre grupos que se han observado en la investigación. Una solución general consistente en realizar un ANOVA entre grupos para cada variable medida y ordenar la importancia en función del valor de F es claramente deseable, puesto que no se tienen en cuenta las correlaciones entre variables y además se está suponiendo que los valores de F son absolutamente estables. En pocas palabras, si la F en una variable es 6,14 y en otra es 6,04, no estamos legitimados para mantener que la primera es más importante que la segunda.

Mejores alternativas son, con mucho las siguientes. La contribución de una variable al constructo teórico subyacente puede evaluarse fácilmente examinando las cargas canónicas o matriz de estructura. La carga canónica es la correlación entre una variable predictora y la función discriminante. Para calcular las cargas es necesario conocer las correlaciones entre las variables predictoras y también los pesos de cada variable. La matriz de correlaciones se obtiene a partir de la de suma de cuadrados y productos cruzados error (W), dividiendo cada elemento fuera de la diagonal por la raíz del producto de los elementos diagonales correspondientes a cada variable. Las operaciones necesarias aparecen ejemplificadas en el Cuadro 10.6. (pág. siguiente), junto con los resultados que SPSS proporciona por defecto. Para obtenerla basta multiplicar la matriz de correlaciones intragrupo (Rxx) por el vector de pesos estandarizados de la función discriminante. Puesto que tenemos dos funciones obtendremos dos cargas para cada variable, una por función.

Es importante destacar que si usamos directamente los pesos estandarizados para evaluar la importancia de la variable podríamos pensar que ésta tiene un valor muy semejante a la escala 3 en la función 1 y mayor que la escala 2 en la función 2. Sin embargo, su carga es cero en las dos funciones mientras que las escalas 2 y 3 tienen una carga considerablemente mayor. No obstante no hay acuerdo sobre el tamaño que una carga debe tener para decidir si la variable entra o no en la función. Una norma que produce bastante consenso es asumir que cargas por encima de 0,33 indican que la variable debe ser incluida, y excluida si la carga es inferior.

Cuando nuestro interés reside en determinar hasta qué punto una variable es útil para separar los grupos podemos utilizar la carga. En este caso claramente podríamos eliminar la variable E1. Pero la carga no permite tomar una decisión basada en un criterio estadístico. ¿Qué alternativas de análisis podemos emplear? La primera es simple, consiste en repetir el análisis suprimiendo selectivamente variables. Podemos comenzar, por ejemplo, suprimiendo la

escala E1 y evaluando si la capacidad de separar los grupos se deteriora significativamente. A continuación comparamos el cambio en la lambda de Wilks obtenido con las tres escalas que se han mantenido en el análisis con respecto al obtenido con todas las variables. Si el cambio es importante, por ejemplo, una función deja de ser significativa, entonces se considera que la variable eliminada hacia una aportación fundamental. La otra alternativa consiste en realizar un análisis discriminante por etapas, que expondremos a continuación.

Cuadro 10.6. Obtención de la carga canónica de cada predictor en cada función discriminante

1) Cálculo manual
W
$\begin{bmatrix} 30 & -5 & 10 & -11 \\ -5 & 10 & 1 & 7 \\ 10 & 1 & 12 & -4 \\ -11 & 7 & -4 & 14 \end{bmatrix}$
Función discriminante 1
$\begin{array}{c c c c} R_{xx} & bs_1 & \text{Carga} & \\ \hline 1 & 0,201 & -0,0003 & E_1 \\ -0,29 & -1,19 & -0,4815 & E_2 \\ 0,53 & 0,246 & -0,1442 & E_3 \\ -0,54 & 1,256 & 0,3688 & E_4 \end{array}$
Función discriminante 2
$\begin{array}{c c c c} R_{xx} & bs_2 & \text{Carga} & \\ \hline 1 & -0,44 & -0,0001 & E_1 \\ -0,29 & -0,07 & 0,3796 & E_2 \\ 0,53 & 1,168 & 0,8171 & E_3 \\ -0,54 & 0,365 & 0,1993 & E_4 \end{array}$
2) SPSS
Matriz de estructura
$\begin{array}{c c c} & \text{Función} & \\ \hline & 1 & 2 \\ \hline E_2 & -,482 & ,379 \\ E_4 & ,368 & ,199 \\ E_1 & ,000 & ,000 \\ E_3 & -,144 & ,817 \end{array}$
Nota: SPSS ordena las variables según la correlación, no según la carga.

### 3.4. ANÁLISIS DISCRIMINANTE POR ETAPAS

Hasta aquí hemos considerado una técnica de análisis discriminante en la que todos los predictores son introducidos simultáneamente en la ecuación. Sabemos ya (véase Capítulo VIII) que no es ésta la única forma de introducir variables en una ecuación de predicción, sino que hay procedimientos que nos permiten introducir y eliminar variables en la ecuación según la aportación que realicen a la predicción. El análisis discriminante por etapas es una técnica que nos permite introducir de forma sucesiva variables de las funciones discriminantes. El análisis procede del siguiente modo. En primer lugar se trata a cada variable como si fuese la única predictora. La decisión de cuál es la primera que entra en la función discriminante depende del estadístico que se desee emplear. En SPSS pueden emplearse cinco diferentes: lambda de Wilks, varianza no explicada, Distancia de Mahalanobis, razón F y V de Rao. Los resultados pueden variar según el estadístico que se emplee.

¿Cómo se procede en este primer paso? Cuando se emplea la lambda de Wilks es obvio que si no hemos introducido todavía ninguna variable, su valor será 1. A continuación se realizan los cálculos para determinar cuál sería la lambda si entrase una de las variables. En nuestro caso, puesto que tenemos cuatro predictores, obtendríamos cuatro valores de lambda. El menor de ellos es el que implica mayor reducción de lambda. Por tanto se selecciona la variable que menor lambda tiene. En los pasos siguientes se procede de manera similar, pero ahora teniendo en cuenta que ya tenemos variables en la ecuación. Las variables que no reducen la lambda de Wilks no son introducidas en la ecuación. Recuérdese que cuanto menor sea lambda menor proporción de variabilidad total se debe al error. Una vez que tenemos las variables en la ecuación el resto del análisis puede completarse como hemos visto en el caso simultáneo. En el Cuadro 10.7. se presenta el análisis mediante SPSS de nuestro ejemplo. La secuencia de comandos empleada ha sido la siguiente:

Analizar → Clasificar → Discriminante → Variable de agrupamiento: Grupo → Definir rango → Mínimo: 1, Máximo: 3 → Continuar → Usar método de inclusión por pasos → Método → Lambda de Wilks → Resumen de los pasos → Continuar → Aceptar (análisis)

Si en lugar de la lambda utilizamos la reducción en la varianza no explicada, la estrategia es semejante, aunque ahora se introduce una variable en la ecuación y se computa la varianza error. Se procede igual con cada variable. Es seleccionada la variable con la que la varianza error sea mínima, es decir, la que más reduzca el error. Los pasos siguientes van introduciendo sucesivamente variables que contribuyan más a reducir la varianza error. Las variables que no contribuyen a reducir la varianza no son incluidas en la ecuación.

Cuadro 10.7. Análisis discriminante por etapas usando como criterio de inclusión la lambda de Wilks

1) Variables no incluidas aún en el análisis según la etapa					
Variables no incluidas en el análisis					
Paso		Tolerancia	Tolerancia Min.	F que introducir	Lambda de Wilks
0	E1	1,000	1,000	,000	1,000
	E2	1,000	1,000	50,000	,107
	E3	1,000	1,000	35,000	,146
	E4	1,000	1,000	27,143	,181
1	E1	,917	,917	,446	,099
	E3	,992	,992	19,519	,024
	E4	,650	,650	24,176	,020
2	E1	,711	,504	1,070	,016
	E3	,789	,517	19,672	,004
3	E1	,556	,473	,597	,004

Nota: la tolerancia es  $1-R^2$ . La F que introducir se calcula como en un análisis de covarianza, donde los covariados son las otras variables que ya están en la ecuación. Obsérvese que en el paso 0, E2 es la que mayor F para entrar y menor lambda de Wilks tiene, por tanto debe ser incluida la primera en la ecuación. En el paso 1, la de mayor F y menor lambda es E4, por tanto debe ser la segunda. En el paso 2, E3 es la variable idónea. Finalmente E1 no debe ser incluida, puesto que no consigue reducir lambda y su F para entrar es menor que 1.

2) Variables incluidas en el análisis según el paso				
Variables en el análisis				
Paso		Tolerancia	F que eliminar	Lambda de Wilks
1	E2	1,000	50,000	
2	E2	,650	44,642	,181
	E4	,650	24,176	,107
3	E2	,567	25,167	,024
	E4	,517	24,264	,024
	E3	,789	19,672	,020

Nota: la F que eliminar se calcula también como en un análisis de covarianza, pero cada una de las variables de la ecuación. Aquellas que tengan un valor alto no deben ser eliminadas.

3) Variables que quedan finalmente en la ecuación.									
Variables introducidas/eliminadas <sup>b,c</sup>									
Paso	Introducidas	Estadístico	Lambda de Wilks						
			gl1	gl2	gl3	Estadístico	gl1	gl2	Sig.
1	E2	,107	1	2	12,000	50,000	2	12,000	,000
2	E4	,020	2	2	12,000	33,530	4	22,000	,000
3	E3	,004	3	2	12,000	49,212	6	20,000	,000

El método basado en la distancia (D) de Mahalanobis utiliza como estrategia la maximización de la varianza entre grupos (estandarizada) considerando los grupos más cercanos entre sí. En cada paso se introduce la variable que hace máxima D. Las variables que no contribuyen no son incluidas en la ecuación. Algo semejante ocurre cuando se emplea la razón F, que se basa en la distancia de Mahalanobis. Finalmente, la V de Rao (que es la traza de Hotelling-Lawley, véase Capítulo IX) se basa también en las distancias entre las medias de grupo. La estrategia es similar a los casos anteriores.

¿Cuál de los métodos elegir para nuestro análisis? La respuesta que podemos dar es sencilla, los paquetes estadísticos como SPSS son extraordinariamente rápidos para realizar los cálculos, de manera que merece la pena examinar la consistencia de las soluciones a través de distintos estadísticos. Seleccionar las variables que estos diferentes estadísticos introduzcan en la ecuación de forma consistente proporcionará mayor confianza en que la solución alcanzada es satisfactoria. En cualquier caso, en general, no deben esperarse grandes diferencias entre unos y otros, y todos son estadísticamente legítimos, de manera que el investigador debe sentirse libre de seleccionar el que mejor convenga a sus objetivos de investigación.

#### 4. ANÁLISIS DISCRIMINANTE PREDICTIVO

##### 4.1. OBTENCIÓN DE LOS COEFICIENTES DE CLASIFICACIÓN LINEAL

En bastantes ocasiones nuestro interés reside en clasificar individuos en grupos, particularmente personas de las cuales desconocemos su grupo de pertenencia, pero de las que tenemos medidas las variables predictoras. Por ejemplo, a los analistas de encuestas les interesa más poder clasificar a los encuestados que no declaran su voto, que clasificar a los que lo han declarado. En estos casos no se utilizan las funciones discriminantes lineales, sino las funciones de clasificación lineal (LCF). Estas funciones son también combinaciones lineales de las variables predictoras, pero difieren de las discriminantes en dos aspectos básicos. Primero, hay una función de clasificación por cada grupo. Así, en nuestro ejemplo tendremos que construir tres funciones diferentes. Segundo, para cada sujeto se calcula la función correspondiente a cada uno de los grupos. Esto es, en nuestro ejemplo obtendremos tres valores para cada sujeto, uno por cada uno de los tres grupos que tenemos. Tercero, los sujetos son asignados a los grupos siguiendo una regla de clasificación sencilla: el sujeto es miembro del grupo en el que tenga mayor función de clasificación. Si, por ejemplo, tenemos un sujeto que tiene valores 3, 4 y 2,9, respectivamente para las funciones correspondientes a los grupos paranoide, alcohólico y personalidad múltiple, será considerado miembro del grupo de alcohólicos (aunque de hecho sea miembro del grupo paranoide). Por último, los coeficientes de las funciones de clasificación se calculan de manera diferente a los de las funciones discriminantes. Teniendo en cuenta que un valor LCF para un grupo j es obtenido como:

$$LCF_j = c_{j0} + c_{j1}X_1 + c_{j2}X_2 + \dots + c_{jp}X_p \quad (\text{Ec. 10.3})$$

Los coeficientes serán obtenidos multiplicando la inversa de varianzas-covarianzas error ( $S^{-1}$ ) por la matriz que contiene las medias de grupo (M). Recorremos que la matriz de varianzas-covarianzas se obtiene dividiendo cada elemento de la matriz de sumas de cuadrados y productos cruzados error (W) por los grados de libertad intragrupo (número total de sujetos menos número de grupos ( $N-g$ ), o bien  $(n-1)g$ , donde N es el total de sujetos y n los de un grupo). Por tanto, en términos matriciales tendremos:

$$C_j = S^{-1}M_j, \quad (\text{Ec. 10.4})$$

donde  $C_j$  es el vector de coeficientes de la función de clasificación lineal del grupo j. Estos coeficientes corresponden cada uno a una de las variables predictoras. Para hallar la constante  $c_{j0}$  utilizaremos la expresión siguiente:

$$c_{j0} = (-0,5) C_j M_j, \quad (\text{Ec. 10.5})$$

es decir, multiplicamos los coeficientes correspondientes a las variables por las medias de esas variables. Es importante tener presente que la clasificación puede hacerse sólo con las variables seleccionadas tras el análisis discriminante descriptivo. No obstante, la inclusión de todas las predictoras no perjudicará, en general la clasificación.

El Cuadro 10.8. presenta estos cálculos utilizando los datos del ejemplo que venimos considerando a lo largo de este capítulo. Este cálculo se basa en el supuesto de que todos los grupos tienen el mismo número de sujetos, o, lo que es lo mismo, que las probabilidades *a priori* de los grupos son iguales. Cuando esto no ocurre, por ejemplo, puede haber más votantes declarados de un partido que de otros, es recomendable realizar un ajuste de la regla de clasificación basándose en las probabilidades *a priori*. El ajuste consiste en añadir un término que corrige la desigualdad del número de sujetos. Considerando el que utilizan los paquetes estadísticos, la función de clasificación quedaría como:

$$LCF_j = c_{j0} + c_{j1}X_1 + c_{j2}X_2 + \dots + c_{jp}X_p + \ln(n_j/N) \quad (\text{Ec. 10.6})$$

donde  $n_j$  es el número de sujetos del grupo j para el cual estamos calculando la función.

Cuadro 10.8. Cómputo de los coeficientes de las funciones de clasificación lineal

1) Obtención de la matriz de varianzas-covarianzas (S) dividiendo W por  $(N-g) = 15-3 = 12$ 

W				S			
30	-5	10	-11	2,5	-0,4	0,83	-0,9
-5	10	1	7	-0,4	0,83	0,08	0,58
10	1	12	-4	0,83	0,08	1	-0,3
-11	7	-4	14	-0,9	0,58	-0,3	1,17

2) Obtención de los coeficientes de clasificación multiplicando la inversa de S por M (medias de grupo)

$S^{-1}$				$M_1$	$M_2$	$M_3$	$C_1$	$C_2$	$C_3$	
0,72	0,18	-0,5	0,33	6	6	6	3,21	4,49	5,5	E1
0,18	2,16	-0,7	-1,1	8	3	3	9,95	-5,9	0,41	E2
-0,5	-0,7	1,62	0,42	7	6	2	3,91	7,94	0,2	E3
0,33	-1,1	0,42	1,81	3	8	5	1,23	15,6	8,46	E4

3) Obtención de la constante para el grupo 1 multiplicando sus medias por coeficientes respectivos y por -0,5

$$C_{10} = (-1/2)^* \begin{matrix} 6 & 8 & 7 & 3 \end{matrix} = \begin{matrix} 3,21 \\ 9,95 \\ 3,91 \\ 1,23 \end{matrix} = -65$$

En SPSS podemos obtener los coeficientes incluyendo en los comandos de análisis la opción siguiente:

Estadísticos → Coeficientes de la función: De Fisher

El resultado de esta instrucción aparece en el Cuadro 10.9.

Cuadro 10.9. Coeficientes de la función de clasificación lineal en SPSS

	Coeficientes de la función de clasificación		
	GRUPO	1,000000	2,000000
E1	3,208	4,491	5,505
E2	9,949	-5,856	,407
E3	3,909	7,940	,200
E4	1,234	15,582	8,465
(Constante)	-66,051	-91,935	-39,585

Funciones discriminantes lineales de Fisher

#### 4.2. CLASIFICACIÓN DE INDIVIDUOS

Una vez que tenemos los coeficientes lo relevante es clasificar en los grupos a los sujetos cuya pertenencia a grupo es conocida y evaluar el ajuste de la predicción del grupo. Como ya hemos indicado, para clasificar a un sujeto en un grupo es necesario calcular su puntuación en cada función de clasificación. A continuación se asigna el sujeto al grupo en cuya función ha obtenido mayor valor. El Cuadro 10.10. presenta los cálculos para cada uno de los 15 sujetos valor. El Cuadro 10.10. presenta los cálculos para cada uno de los 15 sujetos de nuestro ejemplo. Se incluye también el grupo original y el grupo pronosticado. Finalmente se predice lo que ocurriría con un sujeto del que no supiésemos a qué grupo pertenece y que hubiese obtenido las puntuaciones (4, 6, 8, 9), respectivamente para E1, E2, E3 y E4.

Cuadro 10.10. Clasificación de sujetos en grupos

1) Puntuaciones en las funciones de clasificación de cada grupo para cada sujeto y asignación al grupo en que más puntúa.				
GO	E1	E2	E3	E4
1	5	8	7	2
1	4	9	6	5
1	6	7	6	3
1	8	8	9	2
1	7	8	7	3
2	7	4	6	8
2	5	4	7	9
2	8	2	6	6
2	6	3	5	9
2	4	2	6	8
3	4	4	1	6
3	8	2	3	5
3	7	3	2	5
3	5	2	1	4
3	6	4	3	5

GO	LDF1	LDF2	LDF3	GP
59,4	-30	9,53	1	
65,9	-1,1	29,6	1	
50	-12	22,9	1	
76,9	-0,2	26,4	1	
67,1	-5	29	1	
29,5	88,4	69,5	2	
28,3	103	67,1	2	
10,4	73,4	57,3	2	
13,7	97,4	71,8	2	
0,01	86,6	52,2	2	
-2,1	4,04	35,1	3	
-2,6	34	48,2	3	
0,24	15,7	42,9	3	
-21	-11	22,8	3	
10,9	13,3	38	3	

Nota: GO es grupo original, GP es grupo predicho.

2) Clasificación de un sujeto del que no se conoce su grupo

E1	E2	E3	E4	LCF1	LCF2	LCF3	GP
4	6	8	9	48,9	94,6	62,66	2

¿Cómo evaluar si la clasificación ha sido suficientemente correcta? Suelen emplearse dos formas. Por un lado, se obtiene una tabla de contingencia de dos entradas en las que en la filas se ubica el grupo original de pertenencia y en las columnas el grupo pronosticado. De este modo tendremos las frecuencias con que los sujetos han sido correctamente asignados a sus grupos de pertenencia (diagonal de la tabla) y los errores de asignación (los sujetos han sido asignados

a un grupo diferente del suyo) fuera de la diagonal. Este conteo de frecuencias puede realizarse a mano, sin embargo, es preferible realizarlo de manera automática. Podemos lograr esto en SPSS indicando en los comandos de análisis discriminante que se quieren guardar los grupos pronosticados. La secuencia de instrucciones (además de las indicadas anteriormente) sería:

... → Guardar → Grupo de pertenencia pronosticado → Continuar → ...

Cuadro 10.11. Evaluación de la clasificación

#### 1) Frecuencias de aciertos y errores de clasificación

Tabla de contingencia GRUPO\*Grupo pronosticado para el análisis 1

Recuento

GRUPO	Predicho			
	1,000000	2,000000	3,000000	Total
1,000000	5			5
2,000000		5		5
3,000000			5	5
Total	5	5	5	15

#### 2) Evaluación del ajuste

Pruebas de chi-cuadrado

	Valor	gl	Sig. asint. (bilateral)
Chi-cuadrado de Pearson	30,000 <sup>a</sup>	4	,000
Razón de verosimilitud	32,958	4	,000
Asociación lineal por lineal	14,000	1	,000
Núm. de casos válidos	15		

<sup>a</sup> 9 casillas (100,0%) tienen una frecuencia esperada inferior a 5.  
La frecuencia mínima esperada es 1,67

De este modo SPSS crea una nueva variable en nuestro fichero de datos que etiqueta como dis\_1. Podemos cambiar ese nombre por uno más apropiado, como «Predicho». El paso siguiente consiste en utilizar la utilidad de tablas de contingencia incluida en los estadísticos descriptivos para construir la tabla y determinar si el ajuste de la predicción es significativo (por ejemplo mediante chi-cuadrado). Los comandos serían los siguientes:

Analizar → Estadísticos descriptivos → Tablas de contingencia → Filas: Grupo, Columnas: Predicho → Estadísticos → Chi-cuadrado → Continuar → Aceptar (análisis)

La salida que ofrece SPSS aparece en el Cuadro 10.11. (véase página anterior). Obsérvese que la clasificación es perfecta. Como el lector puede comprobar fácilmente, también lo hubiese sido si hubiésemos excluido la escala E1, cuya carga canónica es despreciable y que en el análisis por etapas fue excluida de las funciones discriminantes.

El segundo procedimiento recomendable consiste en calcular los coeficientes de las funciones excluyendo algunos de los casos, lo que se llama genéricamente validación cruzada. La forma estándar de aplicarlo requiere la obtención de una nueva muestra de sujetos de los cuales se conoce también su pertenencia a grupo. Los coeficientes se calculan en una de las muestras y se usan para clasificar los sujetos en la segunda muestra. Aunque esta técnica tiene ventajas evidentes en cuanto que permite evaluar la generalización de los coeficientes obtenidos, es difícil de aplicar en muchos contextos de investigación. Por ello, es frecuente emplear una segunda aproximación, conocida como método de la navaja (*jackknifed classification*), en la que los coeficientes son calculados excluyendo uno por uno a todos los sujetos de la muestra. Es decir, si tenemos 15 sujetos, se calculan 15 conjuntos de coeficientes empleando sólo 14 sujetos en cada cómputo. De este modo cada sujeto tiene un conjunto de coeficientes que se basan en los 14 sujetos restantes. Este método reduce el sesgo clasificadorio que se produce cuando un sujeto es utilizado para computar coeficientes que luego serán empleados para clasificarlo. En SPSS se le llama validación cruzada. La secuencia de comandos, además de los ya mencionados, sería:

→ Clasificar ... → Clasificación dejando uno fuera → Continuar →

Los resultados empleando este método aparecen en el Cuadro 10.12. (véase la página siguiente) junto con el método estándar. Cuando hay discrepancias entre ambos es preferible emplear como resumen de la clasificación los resultados proporcionados por la validación cruzada.

Cuadro 10.12. Validación cruzada

		Resultados de la clasificación <sup>b,c</sup>			Total
		GRUPO	1,000000	2,000000	
Original	Recuento	1,000000	5	0	5
		2,000000	0	5	5
		3,000000	0	0	5
	%	1,000000	100,0	,0	100,0
		2,000000	,0	100,0	100,0
		3,000000	,0	,0	100,0
Validación cruzada <sup>a</sup>	Recuento	1,000000	5	0	5
		2,000000	0	5	5
		3,000000	0	0	5
	%	1,000000	100,0	,0	100,0
		2,000000	,0	100,0	100,0
		3,000000	,0	,0	100,0

<sup>a</sup> La validación cruzada sólo se aplica a los casos del análisis. En la variación cruzada, cada caso se clasifica mediante las funciones derivadas a partir de los casos.

<sup>b</sup> Clasificados correctamente el 100,0% de los casos agrupados originales.

<sup>c</sup> Clasificados correctamente el 100,0% de los casos agrupados mediante validación cruzada.

## 5. LIMITACIONES Y SUPUESTOS DEL ANÁLISIS DISCRIMINANTE

La investigación científica trata de establecer afirmaciones de naturaleza causal cuando se manipula una variable y se miden sus efectos sobre otra. Que estas afirmaciones puedan realizarse no depende de la técnica estadística que se emplee en el análisis de los resultados, sino de cómo se hayan manipulado las variables y del grado de control experimental ejercido. Cuando las variables se manipulan de forma indirecta, como en nuestro ejemplo, difícilmente podemos llegar más allá del establecimiento de afirmaciones de carácter correlacional. Así, no podemos establecer que la patología (paranoide, alcohólico, personalidad límite) es causante de las puntuaciones en las escalas. Sin embargo, si el control experimental ha sido cuidadoso el investigador puede ganar confianza en una creencia de ese tipo. Por otro lado, cuando la variable se ha manipulado de forma directa y los sujetos han sido asignados al azar a los grupos, la garantía de causalidad es sustancialmente mayor. El análisis discriminante es un MANOVA vuelto del revés, por lo que sus limitaciones en este sentido son las mismas que las de MANOVÁ (o que ANOVA). Algo semejante ocurre con sus limitaciones de carácter estadístico, por lo que las discutiremos en este contexto de manera somera. Las más importantes son normalidad multivariada, homogeneidad de varianzas-covarianzas, linealidad, multicolinealidad, puntos extremos y tamaño de las muestras.

### 5.1. NORMALIDAD MULTIVARIADA

Como ya hemos mencionado las puntuaciones se suponen una muestra aleatoria de una distribución normal multivariada. Esto implica, entre otras cosas, que cualquier combinación lineal (las discriminantes lo son) de las variables predictoras debe distribuirse normalmente. Este problema es poco importante si no es producido por puntuaciones extremas.

### 5.2. HOMOGENEIDAD DE VARIANZAS-COVARIANZAS

Si las matrices de varianzas-covarianzas no son homogéneas la tasa de error tipo I real puede ser diferente de la teórica, la que suponemos que estamos cometiendo. Esto es especialmente importante cuando los grupos son pequeños y desiguales. Como ya hemos dicho MANOVA adolece del mismo defecto. Sin embargo, hay un aspecto diferencial en el análisis discriminante. La clasificación de individuos se basa en la matriz de varianzas-covarianzas error (véase Ec. 10.4). El efecto más severo del no cumplimiento del supuesto es que los sujetos tienden a ser clasificados con preferencia en el grupo que mayor varianza tiene, incrementando así los errores de predicción de forma espuria. Si se comprueba la falta de homogeneidad una buena alternativa puede consistir en emplear matrices de covarianzas de los grupos separados.

### 5.3. LINEALIDAD Y MULTICOLINEALIDAD

El supuesto de linealidad implica que los predictores están relacionados linealmente entre sí en cada grupo (igual que en MANOVA). Su incumplimiento reduce la potencia. Sin embargo, si los predictores están altamente relacionados puede producirse multicolinealidad e impedir el análisis puesto que la matriz de sumas de cuadrados y productos cruzados error carecerá de inversa. Cuando esto ocurre lo mejor es eliminar las variables que producen la multicolinealidad.

### 5.4. MUESTRAS DESIGUALES

El mayor problema que se presenta cuando el número de sujetos por grupo es desigual radica en la desaparición de la ortogonalidad de las funciones discriminantes. Por lo demás, al tratarse siempre de diseños de un solo factor entre grupos no se producen grandes alteraciones de potencia o de error tipo I. Sin embargo, es preciso tener cautela a la hora de clasificar a los individuos, puesto que las probabilidades *a priori* no son iguales y, como hemos señalado antes, es

preciso corregir el sesgo que eso supone. En SPSS puede corregirse indicando que la clasificación debe hacerse teniendo en cuenta el tamaño de los grupos.

### 5.5. PUNTOS EXTREMOS

Los puntos extremos pueden afectar en gran medida tanto a las funciones discriminantes como a las de clasificación. Por tanto es importante cerciorarse de si existen y eliminarlos en su caso antes de analizar los datos.

## 6. LA SECUENCIA DE INVESTIGACIÓN ORIENTADA AL ANÁLISIS DISCRIMINANTE

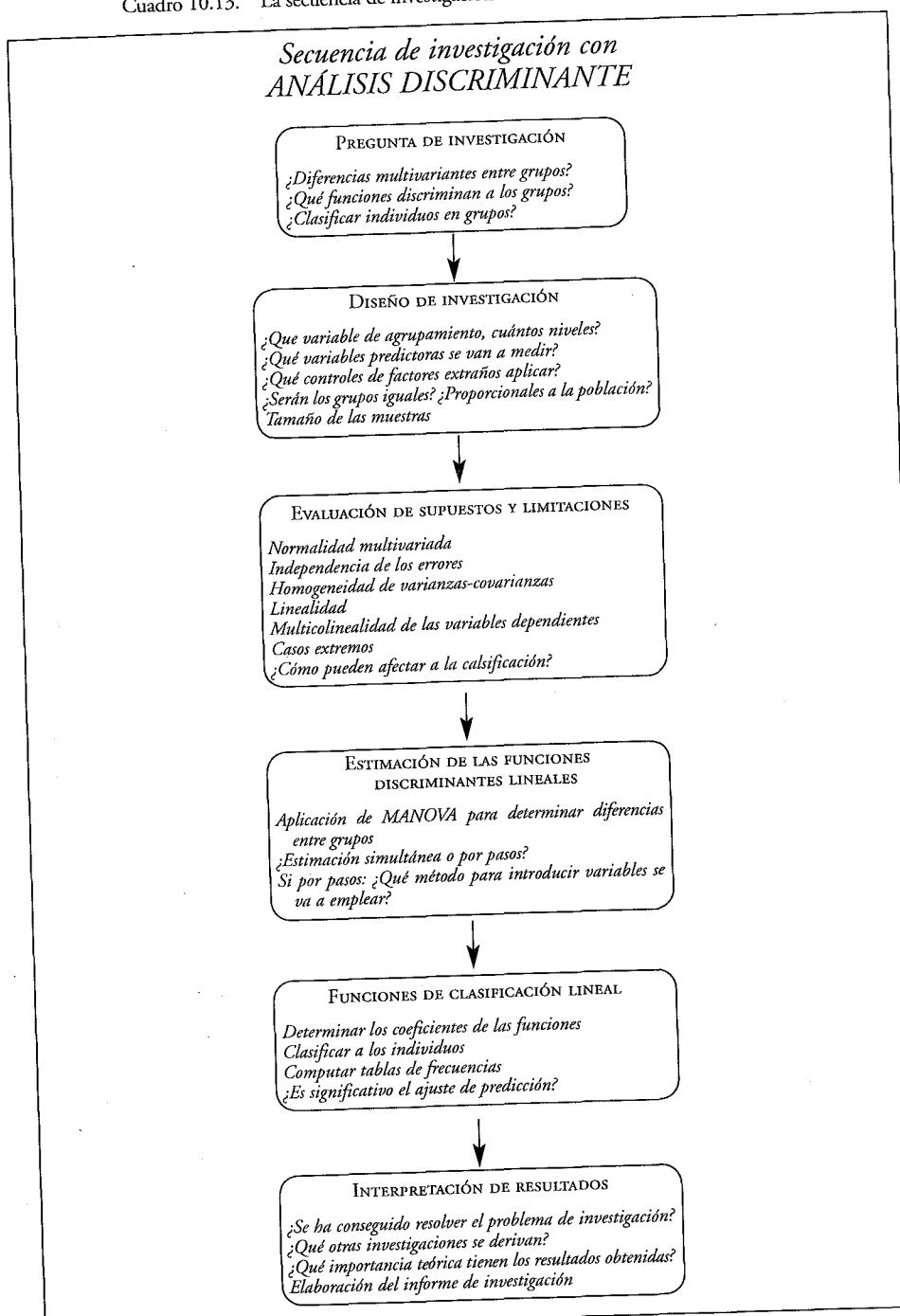
Las etapas del análisis discriminante son en gran parte semejantes a las de MANOVA. Podemos concretarlas en siete fundamentales (véase Cuadro 10.13., en página siguiente). En primer lugar es conveniente especificar cuidadosamente el problema de investigación. Como ya hemos señalado este paso resulta crucial para que el conjunto del proceso permita llegar a conclusiones válidas desde un punto de vista estadístico y teórico. En segundo lugar, es preciso preparar la investigación resolviendo las cuestiones de diseño relativas a la variable que se va a manipular, cómo se la va a manipular (de forma directa o indirecta), qué variables se van a medir, cómo se las medirá, qué variables podrían afectar a la investigación y deben ser controladas y qué tamaño van a tener los grupos. En este sentido debe recordarse que la clasificación es afectada por el número desigual de sujetos por grupo.

En tercer lugar debe procederse al análisis discriminante, que se realiza en tres etapas. Primero es importante comprobar el cumplimiento de los supuestos y limitaciones, particularmente aquellos (como la existencia de puntos extremos o la homogeneidad de varianzas-covarianzas) que pueden afectar a los estimados de los efectos. A continuación es importante realizar el análisis discriminante descriptivo, en el que primero se comprobará si efectivamente hay diferencias multivariadas entre los grupos. Si la respuesta es negativa, lógicamente hay que concluir el análisis. Pero si es positiva, el siguiente paso es computar las funciones discriminantes lineales. Para ello hemos de decidir si el método va a ser simultáneo, introduciendo todos los predictores a la vez, o sucesivo. La elección de un procedimiento sucesivo conlleva la necesidad de decidir cuáles de los métodos de inclusión/exclusión de variables se va a utilizar. Recuérdese que métodos diferentes pueden producir resultados diferentes.

Una vez obtenidas las funciones discriminantes es importante interpretarlas. Ello implica determinar qué variables quedan incluidas. La carga canónica es un procedimiento alternativo al análisis sucesivo que indica, en general, la importancia que la variable tiene para cada función discriminante. Las representaciones gráficas de las funciones discriminantes pueden ayudar a comprender cómo se produce la separación entre los grupos.



Cuadro 10.13. La secuencia de investigación orientada al análisis discriminante



Cuando se ha decidido qué variables son las que contribuyen a discriminar el paso siguiente es clasificar a los individuos en los grupos. Las funciones de clasificación lineal deben evaluarse en cuanto a su eficacia predictora. La evaluación debe ser doble, es importante que el ajuste sea bueno cuando todos los sujetos han sido incluidos en los cálculos, pero no hay que olvidar que una intención clara es clasificar nuevos individuos, por lo que la validación cruzada es siempre recomendable como método para determinar la generalidad de las funciones de clasificación obtenidas.

Finalmente, los resultados deben contextualizarse teóricamente, lo que puede dar lugar a nuevos problemas de investigación, que pueden requerir o no del análisis discriminante. La preparación del informe de investigación es un aspecto central de la elaboración que el investigador debe realizar de sus datos.

## 7. LECTURAS RECOMENDADAS

- COOLEY, W. W. y LOHNEs, P. R., *Multivariate data analysis*, Nueva York, Wiley, 1971.  
 GIL, J.; GARCÍA, E. y RODRÍGUEZ, G., *Análisis discriminante*, Madrid, La Muralla, 2001.  
 HUBERTY, C. J., «Issues in the use and interpretation of discriminant analysis», *Psychological Bulletin*, 95, 1984, págs. 156-171.  
 — *Applied Discriminant Analysis*, Nueva York, Wiley, 1994.  
 RIVAS, M.ª T.; MARTÍNEZ-ARIAS, M.ª R. y RIUS, F., *Análisis discriminante: una aplicación del método «stepwise»*, Málaga, Universidad de Málaga, Departamento de Didáctica y Organización Escolar, 1990.

## LECTURAS COMPLEMENTARIAS

- HUBERTY, C. J. y WIENBAKER, J. M., «Variable importance in Multivariate group comparisons», *Journal of Educational Statistics*, 17, 1992, págs. 75-91.  
 KLECKA, W. R., *Discriminant analysis*, Beverly Hills, CA, Sage, 1982.  
 STEVENS, J., *Applied multivariate statistics for the social sciences (3.ª ed.)*, Mahwah, NJ, Lawrence Erlbaum Associates, 1996.

## 8. EJERCICIOS

1. A partir de un conjunto de síntomas se quiere diagnosticar si las personas sufren alteraciones cognitivas o emocionales. ¿Qué técnica se emplearía para analizar sus resultados?
2. ¿Para qué sirven las funciones discriminantes lineales?

3. ¿Qué indica la carga canónica?
4. En una investigación con variables de memoria a corto plazo, a largo plazo, de orientación atencional y de vigilancia, usted tiene la siguiente matriz de estructura

	FD1	FD2
mcp	-0,45	0,24
mlp	-0,63	0,28
oa	0,28	0,65
vigil	0,15	0,34

- a) ¿Cuántos grupos tenía en el estudio?  
 b) ¿Cómo interpretaría las dos funciones discriminantes?

#### RESPUESTAS A LOS EJERCICIOS

1. Análisis discriminante, puesto que la variable dependiente es categórica. Alternativamente podría usar regresión logística.
2. Determinar si los grupos difieren en alguna dimensión multivariante.
3. La importancia de las variables predictoras
4. a) Tres grupos.  
     b) En la primera función discriminante tienen mayor peso las variables de memoria, mientras que en la segunda las atencionales.

#### CAPÍTULO XI

### Análisis de regresión logística

#### OBJETIVOS DE APRENDIZAJE

- 1) Diferenciar la regresión logística de otras técnicas de predicción como la regresión múltiple o el análisis discriminante.
- 2) Conocer en qué situaciones es aplicable la regresión logística con preferencia a otras técnicas multivariadas de predicción.
- 3) Conocer qué problemas de investigación pueden estudiarse mediante regresión logística.
- 4) Conocer la lógica de la predicción en la regresión logística.
- 5) Conocer cómo se construye la ecuación de predicción.
- 6) Conocer los tipos de regresión logística.
- 7) Saber interpretar los resultados del análisis.
- 8) Conocer las limitaciones y condiciones de aplicación del análisis.

#### 1. REGRESIÓN LOGÍSTICA: REGRESIÓN CON UNA VARIABLE DEPENDIENTE NO MÉTRICA

La predicción de variables no métricas no es adecuada desde una perspectiva de regresión múltiple, pero sí desde el análisis discriminante. Sin embargo, hemos podido percatarnos de que el análisis discriminante tiene algunas limitaciones importantes como la normalidad multivariada, la relación lineal entre predictores o la homogeneidad de las varianzas-covarianzas entre grupos. Estos supuestos no se cumplen muy frecuentemente. Por ejemplo, si uno de nuestros predictores es una variable categórica (digamos dicotómica, como la presen-

cía/ausencia de una característica) difícilmente podemos pensar que se distribuya normalmente, o que la relación con un predictor métrico continuo sea lineal. La regresión logística es una técnica que tiene los mismos objetivos que el análisis discriminante, predecir la pertenencia a una categoría o grupo y prácticamente de la misma manera, mediante una combinación lineal de los predictores, pero su gran ventaja es que no plantea exigencias tan estrictas sobre las características de esos predictores. Concretamente, no asume la relación lineal entre ellos, no asume homogeneidad de varianzas-covarianzas y no asume que se distribuyan según una normal multivariada. Además, la semejanza entre la regresión logística y la regresión múltiple es evidente, lo que hace que sus resultados sean más fácilmente interpretables que los del análisis discriminante, y sobre todo, permite que las variables predictoras interactúen entre sí. Estas ventajas han hecho que la regresión logística sea una técnica cada vez más empleada cuando se pretende clasificar a individuos en categorías.

Inicialmente fue diseñada para la predicción de variables dicotómicas (por ejemplo, enfermo/sano), pero puede emplearse también para predecir variables políticas (por ejemplo, bajo, medio, alto rendimiento). La lógica de la predicción es la siguiente: se evalúa la probabilidad de que una persona tenga un rendimiento bajo, puesto que tiene un determinado patrón de valores en las variables predictoras. La persona será asignada al grupo al que mayor probabilidad tenga de pertenecer. En otras palabras, la variable dependiente (lo que predecimos) no es el grupo en sí, la categoría de la variable independiente, sino la probabilidad de que algo ocurra dadas ciertas circunstancias (ciertos valores). Esto implica que el análisis es adecuado cuando la relación entre la variable dependiente (la probabilidad) y las variables independientes (los predictores) es no lineal. Por ejemplo, es posible que cuando nos situamos en la parte alta de la escala de CI, grandes cambios (de 130 a 150) no impliquen un cambio importante de rendimiento en tareas de razonamiento, pero cuando nos situamos en el centro de la escala (90-110) los cambios de rendimiento sean más apreciables.

### 1.1. LAS ECUACIONES BÁSICAS DE LA REGRESIÓN LOGÍSTICA

Cuando se trata de predecir una variable dicotómica, que adopta valores 0 ó 1 (por ejemplo, fracaso y éxito, respectivamente), su relación con los predictores es no lineal. En este caso la ecuación de regresión tiene que ser diferente de la que se emplea en regresión múltiple, puesto que de usarlas correríamos el riesgo cierto de predecir valores fuera del rango de la variable (mayores que 1 o menores que 0). Por ello, lo que se predice no es directamente la variable sino la probabilidad de que la variable adopte un cierto valor, por ejemplo, la probabilidad de que se produzca éxito escolar ( $p$ ). La variable dependiente es pues un probabilidad. Para predecir una probabilidad pueden utilizarse diferentes funciones, entre las que destaca la logística. Esta función es la base del cálculo

de la probabilidad  $p$  que queremos predecir. Si llamamos  $X_j$  a los predictores, la ecuación se escribe de la siguiente manera:

$$p = \frac{e^u}{1 + e^u} \quad (\text{Ec. 11.1})$$

donde  $u$  es

$$u = a + b_1X_1 + b_2X_2 + \dots + b_pX_p \quad (\text{Ec. 11.2})$$

Nótese que la Ec. 11.2 tiene la forma de la ecuación de regresión múltiple, donde  $a$  es la constante y los  $b_j$  son los coeficientes de los predictores  $X_j$  correspondientes. Esta expresión es conocida como *logit*, o logaritmo de las verosimilitudes. La razón es la siguiente. La estimación de los parámetros implicados en el cómputo de  $u$  se simplifica bastante si dividimos 11.1 por  $(1-p)$ , esto es, por la probabilidad de que ocurra el otro resultado (por ejemplo, fracaso escolar). Teniendo en cuenta que  $(1-p)$  es

$$1 - p = 1 - \frac{e^u}{1 + e^u} = \frac{1}{1 + e^u} \quad (\text{Ec. 11.3})$$

tendremos que el cociente de probabilidades, conocido como *odds ratio* será

$$\frac{p}{1 - p} = e^u \quad (\text{Ec. 11.4})$$

y por tanto, tomando logaritmos obtendremos

$$\ln\left(\frac{p}{1 - p}\right) = u = a + b_1X_1 + b_2X_2 + \dots + b_pX_p \quad (\text{Ec. 11.5})$$

lo que indica que hay una relación lineal entre el cociente de probabilidades (la probabilidad de pertenecer a un grupo (éxito escolar) partido por la probabilidad de pertenecer al otro (fracaso escolar)) y los predictores.

Ahora todo el problema se reduce a estimar los parámetros. En principio pudiera pensarse que podríamos ya resolver el problema como si se tratase de estimar parámetros en regresión múltiple, sin embargo, desconocemos la probabilidad de cada sujeto de ser miembro de uno u otro grupo. Por ello, el problema reside en obtener la combinación lineal de predictores que hace máxima la probabilidad de obtener los resultados observados en la variable de agrupamiento (el que los sujetos tengan éxito o fracaso escolar). El procedimiento de

estimación que se emplea es, por tanto, el de máxima verosimilitud. Se trata de un procedimiento de carácter iterativo en el que se comienza utilizando un conjunto de coeficientes, y se determina el ajuste de las predicciones con respecto a la variable de agrupamiento (se computan los residuos o errores de predicción). A continuación se modifican los coeficientes y se vuelven a analizar los residuos. El proceso finaliza cuando no se puede mejorar el ajuste.

## 1.2. UTILIDAD DE LA REGRESIÓN LOGÍSTICA

¿Qué tipo de cuestiones de investigación ayuda a resolver la regresión logística? El objetivo fundamental de la regresión logística es el mismo que el del análisis de regresión múltiple o el análisis discriminante, determinar si hay relación entre una variable predicha (usualmente la pertenencia a un grupo de sujetos, también llamada resultado) y un conjunto de predictores. Por ejemplo, ¿podemos predecir el éxito en el curso escolar a partir del tipo de colegio donde se estudia (público, privado, concertado), el nivel socioeconómico de los padres (bajo, medio, alto) y el número de horas de estudio diarias (1, 2, 3 ó 4)? Nótese que el éxito escolar puede definirse como una variable binaria (se aprueba o se fracasa), y que algunos de los predictores pueden considerarse variables métricas (horas de estudio) y otras no (tipo de centro, nivel socioeconómico). Si la relación existe, es importante determinar si todos los predictores son necesarios o, por el contrario, algunos pueden excluirse sin que la predicción se deteriore de modo significativo. Así, el objetivo del análisis se concreta en hallar un modelo que incluya un conjunto de predictores más, como en el resto de análisis, un término constante que sea significativamente superior a un modelo de referencia, que habitualmente contendrá sólo la constante, aunque también puede emplearse un modelo con menos predictores (o menos términos) que el primero. Obviamente si el primero es superior al segundo, ello implica que es necesario incluir los predictores añadidos al primero para predecir el resultado.

Una diferencia importante de la regresión logística con respecto a otras técnicas de clasificación como el análisis discriminante estriba en que aquí es posible introducir términos de interacción de predictores en la ecuación. Por ejemplo, es posible que el nivel socioeconómico esté más asociado con el éxito escolar en colegios concertados, que en privados o públicos. La inclusión de las interacciones hace más difícil la interpretación del modelo, pero, en cualquier caso estos términos son tratados como si se tratase de predictores normales.

El modelo especifica la combinación lineal de predictores que es necesario realizar para predecir el resultado. Por tanto, hay dos cuestiones importantes relativas a la contribución de cada predictor. Primera, lógicamente es necesario conocer los pesos (parámetros) de cada uno y cómo interpretarlos. La interpretación está relacionada con el cambio de resultado cuando cambian los parámetros. Segundo, cuestiones como ¿el nivel socioeconómico incrementa,

disminuye o no afecta a la probabilidad de obtener un resultado determinado? son fundamentales para comprender la manera en que el modelo predice el resultado. En el Capítulo IX hemos visto que podemos evaluarlas de dos formas diferentes. Por un lado puede construirse un test estadístico para contrastar la significación de los coeficientes (por ejemplo, el test de Wald), pero también puede optarse para realizar un análisis por etapas que permite introducir predictores de uno en uno y evaluar, por tanto, cómo cambia la predicción tras la inclusión de un predictor. En este aspecto la regresión logística también difiere de otras técnicas de predicción, puesto que cuando se comparan modelos (por ejemplo uno con todos los predictores, completo, frente a uno que incluya menos predictores) lo que interesa es conocer si la relación entre el conjunto de predictores incluidos en el modelo y el resultado cambia de manera significativa o no de un modelo a otro, si la capacidad predictiva cambia, si la variabilidad explicada cambia. En este caso se emplean medidas análogas al coeficiente de correlación múltiple (o a su cuadrado, el de determinación).

Finalmente, en cualquier técnica de clasificación de individuos nos interesa conocer cómo de eficaz es para asignar correctamente los individuos a sus categorías de pertenencia. Puesto que lo que se predice es la probabilidad de un resultado (la probabilidad de éxito), la regla de asignación es sencilla, se fija un punto de corte en la probabilidad (por ejemplo, 0,50) y se asignan a una categoría (por ejemplo, éxito escolar) aquellos cuya probabilidad sea superior y a la otra categoría (fracaso escolar) a los que estén por debajo. Contando las frecuencias de aciertos y errores de clasificación (¿cuántos individuos que han tenido éxito han sido asignados al grupo de éxito? ¿Cuántos individuos que han fracasado han sido asignados al grupo de fracaso? ¿cuántos errores de clasificación se producen?) podemos evaluar el ajuste global del modelo.

En las páginas que siguen expondremos en primer lugar los cálculos fundamentales de la regresión logística, incluyendo los relacionados con la interpretación de los coeficientes y la evaluación del ajuste y comparación de modelos. A continuación presentaremos los tipos de regresión logística, la clasificación de individuos y su evaluación y las limitaciones y supuestos. Como en capítulos precedentes emplearemos en nuestra exposición los datos de un ejemplo con un pequeño número de sujetos para simplificar. Sea el siguiente. Consideremos que estamos interesados en determinar si puede predecirse el éxito o fracaso escolar (de aquí en adelante RE) a partir de tres variables: nivel socioeconómico (CE de aquí en adelante), con tres valores, alto medio y bajo, tipo de colegio (TC) con tres valores: privado, concertado y público y el número de horas semanales de estudio. Supongamos que los datos registrados tras la evaluación final fueron los que aparecen en la tabla siguiente:

Sujeto	RE	CE	TC	Horas
1	1	1	1	8
2	1	2	2	5
3	1	2	3	12
4	1	3	3	9
5	1	2	2	13
6	1	3	3	10
7	1	2	1	12
8	1	1	2	7
9	1	2	2	6
10	1	2	1	4
11	0	1	1	5
12	0	3	3	8
13	0	1	2	6
14	0	1	1	7
15	0	1	1	4
16	0	1	2	6
17	0	1	2	5
18	0	1	2	7
19	0	2	2	4
20	0	1	1	6

## 2. FUNDAMENTOS DEL ANÁLISIS DE REGRESIÓN LOGÍSTICA

En nuestro ejemplo tenemos tres variables predictoras, de las cuales dos son categóricas y una métrica. Además tenemos una variable de grupo a predecir, RE, que es obligatoriamente categórica. Antes de comenzar el análisis es conveniente caer en la cuenta de que las variables categóricas tienen que ser codificadas de la forma apropiada. Concretamente, muchos paquetes asumen que la variable de agrupamiento adopta valores bien definidos, 1 para un nivel (éxito escolar en nuestro caso) y 0 para el otro (fracaso escolar). Esta codificación puede cambiarse (por ejemplo, 0 para éxito y 1 para fracaso), pero debe tenerse en cuenta que los coeficientes de la ecuación tendrán un signo diferente si se usa un código u otro. Respecto de los predictores categóricos puede decirse algo semejante. Es necesario codificarlos para que puedan entrar en el análisis. La forma más común de código consiste en crear tantas variables ficticias (*dummy*) como categorías tenga la variable menos 1. En cada una de las variables ficticias se codifica una categoría (asignándole 1 a los sujetos que la poseen y 0 a los que no), y en el conjunto quedan codificadas todas. En la siguiente tabla aparece una porción de la codificación que sería necesario aplicar a las dos variables categóricas de nuestro ejemplo.

CE	TC	CE(1)	CE(2)	TC(1)	TC(2)
1	1	1	0	1	0
2	2	0	1	0	1
2	3	0	1	0	0
3	3	0	0	0	0
...	—	—	—	—	—
1	2	1	0	0	1
1	2	1	0	0	1
2	2	0	1	0	1
1	1	1	0	1	0

Obsérvese que la categoría 3 es codificada con el par (0,0) tanto en CE como en TC, mientras que la categoría 1 lo es con (1,0) y la 2 con (0,1). SPSS utiliza este esquema de codificación por defecto, pero podemos especificar uno alternativo. En general los alternativos pueden entenderse como comparaciones planeadas entre las categorías. Los nombres de las variables ficticias son los de las originales más un indicador del número de variable.

El análisis de regresión logística es incluido en SPSS en el grupo de programas de regresión, bajo las etiquetas de logística binaria (el adecuado cuando la variable de agrupamiento tiene dos niveles) y logística multinomial (adecuado cuando la variable de agrupamiento tiene más de dos niveles). Otros paquetes estadísticos, como STATISTICA (6.0) incluyen ambas en un subprograma de estimación no lineal. Los comandos necesarios para aplicar el análisis en SPSS son los siguientes:

Analizar → Regresión → Logística binaria → Dependiente: RE → Covariables: CE, TC, Horas → Covariables categóricas: CE, TC → Continuar → Método: Introducir → Aceptar (análisis)

Es importante precisar que las variables independientes son llamadas covariables por SPSS y que la indicación de qué covariables son categóricas hace que SPSS codifique automáticamente las mismas de la forma en que hemos indicado más arriba. Cuando, como ahora, se utiliza el método introducir y se han incluido todos los predictores como covariables, el resultado de este conjunto de comandos es una estimación de los coeficientes de cada una de ellas. El Cuadro 11.1. (pág. siguiente) presenta los coeficientes (B) para las variables de nuestro ejemplo.

Así, pues, la probabilidad de éxito podemos escribirla como:

$$P = \frac{e^{-162,125 - 110,32CE_1 - 52,51CE_2 + 138,66TC_1 + 138,05TC_2 + 19,07\text{Horas}}}{1 + e^{-162,125 - 110,32CE_1 - 52,51CE_2 + 138,66TC_1 + 138,05TC_2 + 19,07\text{Horas}}}$$

puesto que hemos adjudicado el código 1 a éxito. ¿Cómo podemos aplicar esta ecuación para predecir la probabilidad de que cualquiera de nuestros sujetos

tenga éxito? Obviamente basta sustituir las puntuaciones en los predictores y operar. Por ejemplo, el sujeto 3 trabaja durante 12 horas a la semana, y sus valores en CE han sido CE(1)=0, CE(2)=1, y en TC: TC(1)=0 y TC(2)=0. Por tanto tendremos que

$$p = \frac{e^{-162,125-52,51+19,07(12)}}{1 + e^{-162,125-52,51+19,07(12)}} = 1$$

Cuadro 11.1. Coeficientes de la función *logit*

Variable	B	S.E.	Wald	df	Sig	R	Exp (B)
CE							
CE (1)	-110,320	1625,7015	,0136	2	,9932	,0000	
CE (2)	-52,5066	1347,4768	,0046	1	,9459	,0000	
			,0015	1	,9689	,0000	,0000
TC							
TC (1)	138,6556	1801,0607	,1066	2	,9481	,0000	
TC (2)	138,0506	1801,0607	,0059	1	,9386	,0000	1,650E+60
			,0059	1	,9389	,0000	9,008E+59
HORAS	19,0695	166,3343	,0131	1	,9087	,0000	131328444
Constant	-162,125	1419,0894	,0131	1	,9090		

Está claro que la predicción es bastante aceptable, puesto que este sujeto es realmente miembro del grupo de los que han tenido éxito.

En el cálculo anterior hemos considerado todas las variables para predecir la probabilidad, sin embargo, es posible que algunas no realicen una contribución significativa, y, por tanto no debieran ser incorporadas en la ecuación de predicción. Naturalmente, una forma de evaluar la contribución de las variables consiste en evaluar sus coeficientes. Uno de los test más empleados en este sentido es el de Wald, que es un cociente entre el coeficiente y su error típico. La interpretación del cociente de Wald es sencilla, puesto que se trata de una puntuación z. Este test aparece en el Cuadro 11.1, donde puede comprobarse que ninguno de los coeficientes es significativo. Por tanto, no deberían emplearse ninguna de las variables independientes para predecir el criterio. Sin embargo, como veremos, es posible obtener clasificaciones significativas usando, por lo menos alguno de ellos.

Si el modelo produce o no un ajuste significativo en la predicción es algo que no puede conocerse considerando los coeficientes de las variables. Para evaluar el ajuste lo mejor sería realizar las predicciones y obtener un medida de la precisión alcanzada. Esta idea es la que cuantifica el logaritmo de la verosimilitud (*log-likelihood*), cuya expresión de cálculo es la siguiente:

$$\text{log-likelihood} = \sum_{i=1}^N [Y_i \ln(p_i) + (1 - Y_i) \ln(1 - p_i)] \quad (\text{Ec. 11.6})$$

Primero se computan las predicciones y a continuación se opera para obtener la suma. Sin embargo, el ajuste de un modelo debe computarse en relación al ajuste que produce otro. Cuando todas las variables han sido incluidas en la ecuación, el modelo de referencia debe ser uno que no incluya ningún predictor, sólo la constante. SPSS nos proporciona la *log-likelihood* asociada a la constante al inicio del proceso de datos:

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 27,725887

\* Constant is included in the model.

La comparación entre los dos modelos se realiza computando las diferencias entre el modelo que más variables incluye y el que menos. La diferencia es una chi-cuadrado que se distribuye según la diferencia entre los grados de libertad del primer y del segundo modelo. Los grados de libertad del modelo son el número de predictores incluido (teniendo en cuenta la codificación ficticia) más uno (por la constante). De esta forma, en nuestro caso tendremos 6 grados de libertad para el modelo completo (2 por CE, 2 por TC, 1 por Horas y 1 por la constante) y 1 grado de libertad en el modelo que incluye sólo la constante. La chi-cuadrado se calcula como:

$$\chi^2 = 2[(\text{log-likelihood } (F)) - (\text{log-likelihood } (C))] = -6,491 - (-27,726) = 21,235$$

consultando en las tablas de chi-cuadrado observaremos que la probabilidad de un valor como el obtenido suponiendo que las variables no aportan nada a predecir el resultado es de 0,0007, por tanto, rechazamos ese supuesto y concluimos que sí hay un incremento significativo de la predicción cuando introducimos las variables. En SPSS el resultado aparece del siguiente modo:

Variable(s) Entered on Step Number			
1..	CE	TC	HORAS
Estimation terminated at iteration number 13 because Log Likelihood decreased by less than .01 percent.			
-2 Log Likelihood	6,491		
Goodness of Fit	5,099		
Cox & Snell - R <sup>2</sup>	,654		
Nagelkerke - R <sup>2</sup>	,872		
	Chi-Square	df	Significance
Model	21,235	5	,0007
Block	21,235	5	,0007
Step	21,235	5	,0007

donde Block y Step hacen referencia a la regresión por etapas. El coeficiente R<sup>2</sup> de Cox y Snell (pseudo-R cuadrado) es una medida análoga del coeficiente de determinación en la regresión múltiple. Indica la fuerza de la asociación entre los predictores y la variable dependiente. Su expresión de cálculo es la siguiente:

$$R^2_{CS} = 1 - e^{-(2/n)[LL(B)-LL(A)]},$$

donde LL(B) es la *log-likelihood* del modelo con más predictores y LL(A) la del de menos. El coeficiente de Nagelkerke es una pseudo-R cuadrado de Cox y Snell ajustada. Su expresión es:

$$R^2_N = R^2_{CS}/R^2_{MAX},$$

siendo

$$R^2_{MAX} = 1 - e^{(2/n) LL(A)}$$

Una forma alternativa de computar un coeficiente de determinación consiste en calcular la correlación entre la probabilidad predicha y la variable de agrupamiento. La probabilidad predicha puede obtenerse fácilmente indicando al paquete estadístico que quieren almacenarse las predicciones en una nueva variable (SPSS la llamará *pre\_1*). El valor obtenido no debe diferir mucho del que proporcionan los estimados de Cox y Snell y la corrección de Nagelkerke.

El ajuste del modelo también puede realizarse mediante un procedimiento conocido como deciles de riesgo. El procedimiento consiste en dividir la muestra ordenada de probabilidades predichas en 10 grupos o valores de probabilidad, escalados cada 0,1. Así, el primer grupo estaría formado por los sujetos cuya probabilidad predicha es menor de 0,1, el segundo por los que están entre 0,1 y menos de 0,2, y así sucesivamente. A continuación se separan los sujetos

según la variable dependiente, de modo que se forma una matriz de 2 (grupos en la variable dependiente) x 10 (grupos según los deciles). Si el modelo es adecuado para predecir, los sujetos que tienen 1 en la variable dependiente irán a los deciles altos y los que tienen 0 irán a los deciles bajos. El estadístico de Hosmer-Lomeshow usa la chi-cuadrado para comparar las frecuencias observadas con las predichas. Si el modelo es adecuado, la chi-cuadrado no será significativa. SPSS presenta los resultados del siguiente modo:

Hosmer and Lemeshow Goodness-of-Fit Test					
	RE = 0		RE = 1		
Group	Observed	Expected	Observed	Expected	Total
1	3,000	,000	,000	,000	3,000
2	3,000	,000	,000	,000	3,000
3	1,000	,000	,000	,000	1,000
4	2,000	,000	1,000	,000	3,000
5	1,000	,000	1,000	,000	2,000
6	,000	,000	2,000	,000	2,000
7	,000	,000	2,000	,000	2,000
8	,000	,000	1,000	,000	1,000
9	,000	,000	3,000	,000	3,000
	Chi-Square	df	Significance		
Goodness-of-fit test	,0002	7	1,000		

La última cuestión relevante concierne al análisis de los residuos, cuyo interés reside en la posibilidad de detectar puntos extremos que puedan haber sesgado la estimación de los coeficientes de la función logit. La forma más estándar de evaluarlos (SPSS incluye hasta cinco formas diferentes) consiste en tipificarlos. La tipificación se realiza compuntando el residuo (el grupo observado menos la probabilidad pronosticada) y dividirlo por su error típico. SPSS proporciona los errores estandarizados con la siguiente secuencia de comandos, cuyo efecto es crear una nueva variable, *zre\_1*, que contendrá estos residuos:

... → Guardar → Residuos: Tipificados → Continuar → ...

Los residuos que se sitúen a más de un número, determinado por el investigador, de desviaciones típicas por encima o por debajo de la media (cero) pueden ser considerados extremos y quizás deba pensarse en eliminarlos. Una forma adicional de comprobar si su eliminación es aconsejable consiste en examinar el estadístico de influencia de Cook, que computa el cambio que se produciría

en los residuos si se eliminase cada uno de los casos. La instrucción siguiente pide a SPSS que cree una nueva variable, coo\_1, que contendrá los valores de influencia de Cook.

... → Guardar → Influencia: De Cook → Continuar → ...

Un procedimiento semejante al de validación cruzada que hemos considerado en el Capítulo X. Si los residuos se reducen mucho eliminando un caso particular, y además cumple el criterio de exclusión fijado para los residuos estandarizados (por ejemplo, 2 desviaciones típicas), probablemente ese caso deba ser eliminado. El Cuadro 11.2. muestra parte de los resultados que SPSS crea cuando se introduce las dos instrucciones anteriores. Nótese que ningún sujeto cumple los dos criterios de exclusión. Hay algunos sujetos cuya eliminación, según Cook, cambiaría de forma importante el residuo, pero se sitúan claramente por debajo del criterio que hemos fijado de 2 desviaciones típicas en su error estándar.

Cuadro 11.2. Valores de influencia de Cook y residuos estandarizados

RE	coo_1	zre_1
1	0	0,00008
1	0	0,00008
1		0,00082
1	0,25163	0,00865
1		
1	0	0
1		
1	1,773	7,574
1		
1	1,919	0,85964
0	0	0
0	0,21918	-0,00836
0	0	-0,00005
0	2	-0,85964
0	0	0
0	0	-0,00005
0	0	0
0	0,28874	-0,63526
0	1,919	-0,85964
0	0	-0,00006

### 3. TIPOS DE REGRESIÓN LOGÍSTICA

Hemos podido comprobar ya que es frecuente en las técnicas predictivas (regresión múltiple, análisis discriminante) que haya varios modos de introducir las variables en la ecuación. Esencialmente, los predictores pueden incluirse de manera simultánea (el procedimiento que hemos ilustrado en el punto 2), pero también pueden introducirse de modo secuencial.

#### 3.1. REGRESIÓN LOGÍSTICA SIMULTÁNEA

En la regresión logística simultánea los predictores son introducidos todos a la vez. Esto implica que cada predictor es evaluado como si fuese el último en haber sido introducido en la ecuación. Recordemos que esto implica que el predictor añade a la predicción sólo lo que no comparte con los otros, lo que le diferencia de los otros. El único caso de exclusión posible es que la tolerancia de un predictor sea demasiado baja. Recordemos que la tolerancia es el complementario del coeficiente de determinación, y que una baja tolerancia se produce cuando el predictor es redundante con otro u otros. La inclusión simultánea es recomendable siempre que no se tenga claro qué variables pueden ser más importantes para hacer la predicción.

#### 3.2. REGRESIÓN LOGÍSTICA SECUENCIAL

En la regresión logística secuencial se introduce o elimina un predictor en cada paso. Hay dos subtipos básicos, la regresión logística serial y la regresión logística por etapas (*stepwise*). La diferencia entre ambos procedimientos reside en las razones por las cuales los predictores son incluidos/excluidos de la ecuación.

##### 3.2.1. Regresión logística serial

En la regresión jerárquica los predictores son introducidos según un criterio teórico. Por ejemplo, podríamos pensar que el factor predictivo más importante del éxito escolar son las horas semanales de estudio. Esto haría que fuese el primer predictor que introduciríamos en la ecuación. Podríamos introducir a continuación el nivel socioeconómico, bajo la creencia de que este factor es más importante que el tipo de centro en el que estudian los sujetos. Esta jerarquía teórica será la que guiará, por tanto, el orden de introducción de predictores en la ecuación. Es importante caer en la cuenta de que en cada paso puede introducirse más de un predictor, pero una vez que el predictor está dentro de

la ecuación no puede ser eliminado de la misma. En SPSS la secuencia de comandos es muy semejante a la que habíamos empleado en la regresión múltiple:

Analizar → Regresión → Logística binaria → Dependiente: RE → Covariables: Horas → Siguiente (Bloque) → Covariables: CE → Siguiente (Bloque) → Covariables: TC → Categórica → Covariables categóricas: CE, TC → Contraste: Indicador → Continuar → Aceptar (análisis)

En cada etapa se estiman los coeficientes de cada variable y se evalúa el modelo de la misma forma que hemos indicado ya en el punto 2. Solamente una cuestión novedosa. Primero, se compara el modelo hasta la etapa actual con la etapa actual. En SPSS este resultado aparece del siguiente modo:

	Chi-Square	df	Significance
Model	6,380	1	,0115
Block	6,380	1	,0115
Step	6,380	1	,0115

que corresponde al primer predictor (más la constante, que se introduce en el paso cero, a no ser que se excluya del modelo). Obsérvese cómo el modelo y el bloque coinciden exactamente. Sin embargo, cuando introducimos el segundo predictor (CE) obtenemos:

	Chi-Square	df	Significance
Model	21,134	3	,0001
Block	14,754	2	,0006
Step	14,754	2	,0006

lo que indica que este segundo predictor realiza también una aportación significativa a la predicción. Si examinamos los log-likelihood caeremos inmediatamente en la cuenta de lo que esto implica. Cuando tenemos el predictor Horas y la constante, su valor, LL(1), multiplicado por -2, según SPSS, es

-2 Log-Likelihood	21,346
-------------------	--------

Sin embargo, cuando introducimos también CE, su valor, LL(2), es

-2 Log-Likelihood	6,592
-------------------	-------

por tanto, la comparación entre ambos modelos será la que viene a continuación:

$$\chi^2 = \text{LL}(2) - \text{LL}(1) = 14,754$$

que es justamente la chi-cuadrado del bloque y la etapa. Si la chi-cuadrado es significativa ello implica que la variable hace una aportación relevante a la predicción.

Para terminar, observemos que el tercer predictor (TC), que se introduce en el último paso no hace aportación relevante alguna:

	Chi-Square	df	Significance
Model	21,235	5	,0007
Block	,101	2	,9507
Step	,101	2	,9507

y que el modelo incluyéndolo es idéntico en la práctica al modelo sin incluirlo. Esto indica claramente que el predictor TC puede ser excluido de la ecuación sin que eso suponga menoscabo significativo en la predicción.

### 3.2.2. Regresión logística secuencial

En este procedimiento cada predictor es incluido o eliminado de la ecuación según cumpla los criterios estadísticos de inclusión/exclusión prefijados por el investigador. Como se recordará de la discusión sobre este aspecto realizada en el capítulo de regresión múltiple, esta aproximación es útil cuando se sospecha que hay variables más importantes que otras para la predicción, pero no se tienen hipótesis específicas sobre cuáles pueden ser o cuál puede ser el orden. Además, debe tenerse en cuenta que el problema más importante es que puede excluirse alguna variable que tenga una alta relación con la dependiente debido a su correlación con otras predictoras.

Hay varios procedimientos para tomar la decisión de introducir/excluir variables. SPSS implementa tres hacia delante y tres hacia atrás. Como sabemos ya, en los procedimientos hacia delante se comienza con la constante y se van añadiendo variables que cumplan el criterio de inclusión. En los procedimientos hacia atrás se introducen todas las variables en el primer paso y se van eliminando las que cumplan el criterio de exclusión. Hay varios métodos de selección de variables, todos basan la introducción en el estadístico de puntuación, pero difieren en lo que refiere a la exclusión, uno se basa en el cálculo condicional de los parámetros, otro en la razón de verosimilitud y el tercero se basa en el estadístico de Wald. Los resultados pueden diferir de un método a otro.

#### 4. LA CLASIFICACIÓN DE INDIVIDUOS

Una vez que se han seleccionado las variables que deben ser incluidas en la ecuación es importante determinar si el modelo permite asignar los individuos a las categorías de forma significativa. En este sentido, la primera cuestión fundamental es la regla de asignación de los sujetos a los grupos. Por defecto suele emplearse la probabilidad 0,5 como criterio, de modo que un sujeto cuya probabilidad pronosticada sea menor de 0,5 será asignado al grupo 0, mientras que será asignado al grupo 1 en caso contrario. Sin embargo, este criterio no es obligatorio, puede emplearse cualquier otro. Lo importante es que el criterio seleccionado puede determinar el grado de ajuste clasificatorio del modelo. En SPSS podemos seleccionar la probabilidad de corte (por ejemplo, 0,6) con la siguiente secuencia de comandos:

...→ Opciones → Punto de corte para la clasificación: 0,6 → Continuar ...

La forma más simple de proceder consiste en contar las frecuencias de casos en que los sujetos han sido asignados a su grupo original y los errores de asignación. Los paquetes estadísticos proporcionan el conteo de frecuencia en su salida. Por ejemplo, SPSS lo presenta de la siguiente forma:

Classification Table for RE			
The Cut Value is .50			
Observed	Predicted		Percent Correct
	0	1	
0	10	0	100,00%
1	2	8	80,00%
			90,00%
		Overall	

A la hora de valorar la solución hay que tener en cuenta dos cuestiones básicas. Primero, es importante obtener un grado significativo de acierto en la clasificación. Como sabemos, podemos utilizar un estadístico de ajuste como chi-cuadrado para tomar la decisión. Puesto que SPSS no proporciona el estadístico, es preciso salvar los grupos predichos en una nueva variable. La secuencia de instrucciones es la siguiente:

... → Guardar → Valores pronosticados: Grupo de pertenencia → Continuar →

cuyo resultado es la creación de una nueva variable, pgr\_1, que podremos utilizar junto con la de agrupamiento para realizar el test. En nuestro caso  $\chi^2(1)=16,364$ , que es significativa ( $p<0,05$ ).

Segundo, a menudo clasificar individuos de forma errónea tiene asociado un coste. Los errores de clasificación no son todos del mismo tipo. Por ejemplo, diagnosticar a una persona como paranoide cuando no lo es, es un error conocido como falsa alarma o error tipo I. El segundo tipo de errores son los fallos o tipo II, por ejemplo, diagnosticar como no paranoide a alguien que sí lo es. La consideración de los costes de los errores de clasificación es importante tanto a nivel personal, para los individuos mal clasificados (imagínese que lo diagnostican como esquizofrénico), como económico (el tratamiento es costoso en tiempo y dinero) y social. Esta consideración puede llevar a que el punto de corte seleccionado trate de equilibrar la relación entre unos errores y otros. Habitualmente (por lo menos cuando el número de sujetos por grupo es el mismo) los puntos de corte intermedios son los más recomendables para mantener ese equilibrio.

#### 5. EL ANÁLISIS DE REGRESIÓN POLITÓMICO O MULTINOMIAL

El análisis de regresión logística puede aplicarse también cuando la variable de agrupamiento tiene más de dos niveles. Hay dos casos diferentes posibles. En primer lugar la variable de agrupamiento puede ser nominal y entonces las categorías no pueden ordenarse de ninguna forma. En segundo lugar la variable es ordinal y entonces hay un rango claro de sus niveles. En ambos casos se obtienen tantas ecuaciones de predicción como grados de libertad tiene la variable de agrupamiento (por ejemplo, si tenemos tres grupos, habrá dos grados de libertad y dos ecuaciones diferentes). La interpretación de las probabilidades predichas cambia, sin embargo, de un caso al otro. Cuando los grupos no pueden ordenarse (por ejemplo, las psicopatologías: depresión, ansiedad, etcétera), cada ecuación predice la probabilidad de que el individuo sea miembro de un grupo diferente. Esto es,

$$P(Y=j) = \frac{e^u}{1 + e^u} \quad (\text{Ec. 11.7})$$

donde Y es la variable de agrupamiento, j es el grupo y u es la ecuación de regresión. Por el contrario, cuando los grupos pueden ordenarse (por ejemplo, rendimiento escolar: bajo, medio, alto), las ecuaciones predicen la probabilidad de que el sujeto se sitúe en el grupo superior al índice de la ecuación, esto es, la primera ecuación predice la probabilidad de que el sujeto se sitúe en el segundo grupo, la segunda la probabilidad de que se sitúe en el tercer grupo, y así sucesivamente. Formalmente,

$$P(Y>j) = \frac{e^u}{1 + e^u} \quad (\text{Ec. 11.8})$$

El análisis mediante SPSS se realiza a través del programa de regresión multinomial. La secuencia inicial de comandos es:

Analizar → Regresión → Regresión multinomial →

La diferencia más importante con respecto al programa de regresión logística consiste en que los predictores son separados en dos tipos: categóricos y continuos. Los primeros son denominados factores y los segundos covariables. El concepto subyacente es simple si un predictor permite organizar grupos de sujetos, entonces es categórico y debe ser introducido como un factor. SPSS trata a los factores de una forma similar lo que hemos visto para regresión con variables dicotómicas. En caso contrario, el predictor es una covariante. Por lo demás el análisis no difiere en lo sustancial de lo que hemos presentado más arriba.

## 6. LIMITACIONES Y SUPUESTOS

El análisis de regresión logística es una técnica de clasificación de individuos en clara competencia con el análisis discriminante. La elección entre una y otra depende del cumplimiento de los supuestos subyacentes, sobre todo, al análisis discriminante. Claramente cuando hay predictores categóricos parte de esos supuestos no se cumplen y, por ende, la regresión logística es una técnica más apropiada. Sin embargo, si los supuestos se cumplen, el análisis discriminante es más fácil de interpretar y más eficaz en la predicción. También debe tenerse en cuenta que la regresión logística, aunque menos restrictiva, tiene supuestos y limitaciones cuyo no cumplimiento puede acarrear consecuencias importantes en el ajuste del modelo y de la clasificación.

### 6.1. LINEALIDAD DE LA FUNCIÓN LOGIT

Como hemos mencionado más arriba la función logit es el logaritmo del cociente de probabilidades, o, lo que es lo mismo, la combinación lineal de los predictores. El supuesto implica que los predictores continuos y la función logit estén relacionados linealmente. A diferencia del análisis discriminante, la relación entre los propios predictores es irrelevante. La forma más clara y simple de no linealidad es la interacción de factores (recuérdese que la interacción implica que para obtener un resultado no basta con sumar dos factores, sino que hay que añadir también su producto). Pues bien, una forma simple de comprobar este supuesto consiste en introducir términos de interacción en el modelo y comparar con un modelo sin interacciones. Si alguna de las interacciones es significativa, estará claro que el supuesto no se cumple. Cuando esto ocurra debemos

recordar que la forma más sencilla de eliminar una interacción es mediante una transformación, por ejemplo, logarítmica, de los predictores implicados.

### 6.2. INDEPENDENCIA DE LOS ERRORES

El supuesto implica que las puntuaciones de un sujeto no son predecibles a partir de las de otro sujeto, pero más específicamente debe entenderse en el sentido de que el valor de la variable de agrupamiento de un sujeto no pueda ser predicho a partir del valor de otro sujeto. Esto puede ocurrir cuando los sujetos han sido medidos en la variable dependiente de manera secuencial, o cuando los grupos han sido igualados en variables relevantes. La solución puede pasar por cambiar radicalmente la estrategia del análisis, por ejemplo, considerando la variable de agrupamiento como un intrasujeto.

### 6.3. MULTICOLINEALIDAD

La multicolinealidad se produce cuando las variables predictoras categóricas correlacionan mucho entre sí y/o cuando las variables métricas correlacionan entre sí, lo que implica que hay predictores redundantes. La solución puede consistir en eliminar las variables que producen la multicolinealidad. Para contrastarla debe examinarse la correlación (mediante análisis de frecuencias en el caso categórico) entre todos los pares de predictores.

### 6.4. NÚMERO DE VARIABLES Y NÚMERO DE SUJETOS

Cuando el número de sujetos es relativamente bajo en relación con el número de variables pueden presentarse problemas de estimación de parámetros, debidos especialmente a las variables categóricas, puesto que es posible que las condiciones definidas por su mezcla no contengan sujetos. Por ejemplo, la combinación de nuestras variables CE y TC, que tienen tres niveles cada una, produciría 9 condiciones, pero obsérvese en la tabla siguiente que no todas las condiciones contienen sujetos y que tampoco hay igualdad en su número entre el resto de condiciones. Cuando esto ocurre los estimados de los parámetros pueden ser demasiado altos, y también sus errores estándares. Lo cual puede ser una consecuencia de que el algoritmo iterativo de máxima verosimilitud no encuentra una solución convergente (un conjunto de parámetros que estabilicen el error), por tanto, el número de iteraciones incrementa, y con él el valor de los parámetros. Naturalmente la solución no consiste en limitar el número de iteraciones, sino en incrementar el número de sujetos (lo cual garantizará además la generalidad) y/o en eliminar variables (por ejemplo, en nuestro caso podríamos eliminar TC). Una consecuencia adicional del número de sujetos muy

bajo, si la variable es un buen predictor del resultado, es que puede conseguirse una clasificación perfecta. Por ejemplo, si CE predice bien el resultado, pero tenemos muy pocos sujetos es altamente probablemente que los de un grupo (éxito escolar) tengan una puntuación homogénea en CE y los del otro (fracaso) también. Por otro lado, la predicción puede incrementarse de manera artificial aumentando el número de variables sin cambiar el número de sujetos.

Tabla de contingencia TC\*CE

		Recuento			Total
		1	2	3	
TC	1	5	2		7
	2	5	4		9
	3		1	3	4
Total		10	7	3	20

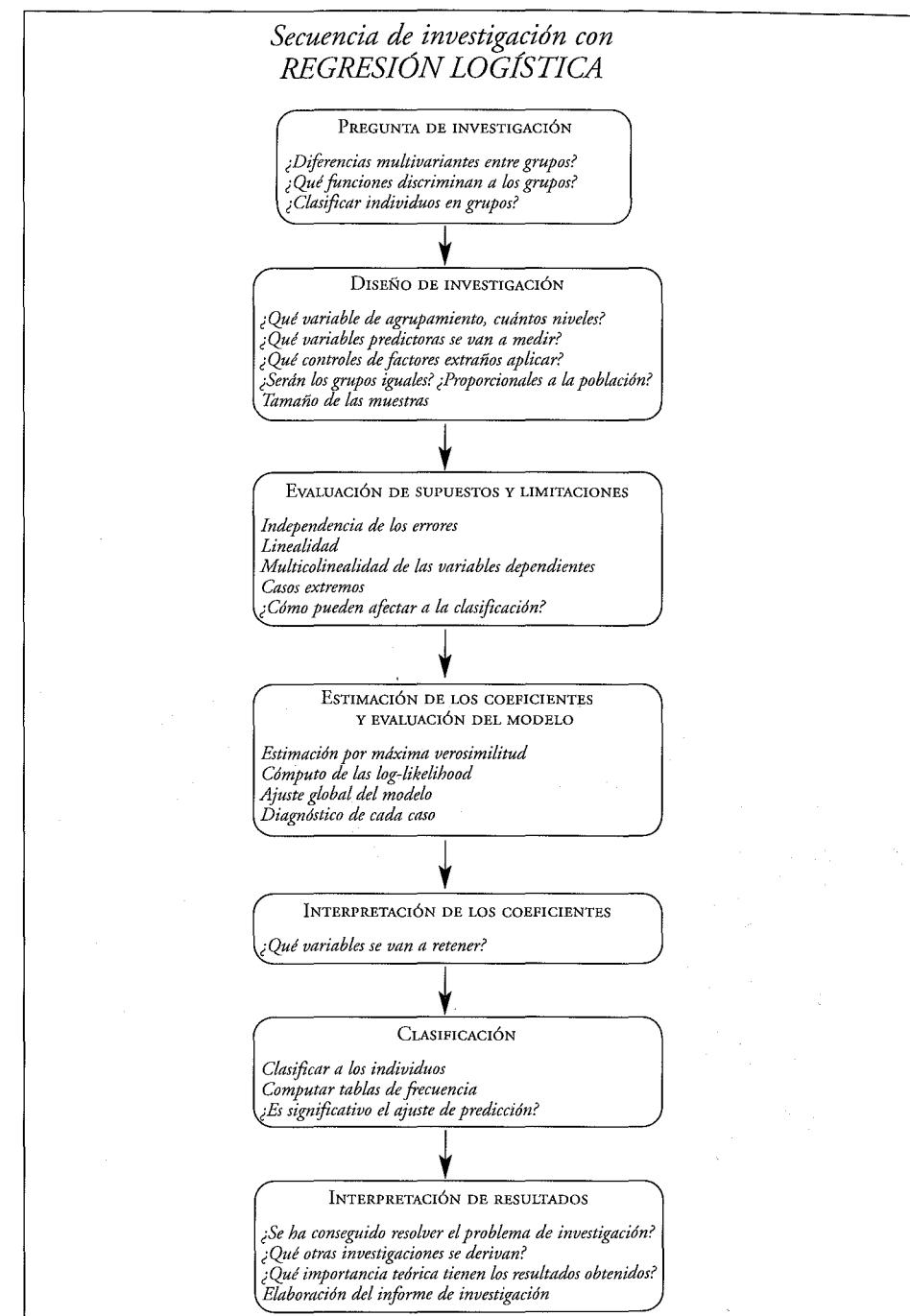
### 6.5. PUNTOS EXTREMOS

La presencia de puntos extremos puede traducirse en una baja capacidad predictiva del modelo. Una solución posible es eliminarlos, pero debe tenerse en cuenta que un criterio de eliminación relajado puede incrementar artificialmente la predicción.

## 7. LA SECUENCIA DE INVESTIGACIÓN ORIENTADA AL ANÁLISIS DE REGRESIÓN LOGÍSTICA

La serie de etapas necesarias para realizar el análisis de regresión logística aparecen en el Cuadro 11.3. Las primeras son, en general, comunes con el análisis discriminante, si exceptuamos la evaluación de las limitaciones y supuestos debido al reducido número de ésta en relación a aquél. Recuérdese que una buena investigación se origina siempre en una buena pregunta y que las buenas preguntas dependen bastante del conocimiento que el investigador tenga del área en la que trabaja. La cuarta etapa consiste en la estimación de los parámetros mediante máxima verosimilitud y en la evaluación del modelo a través de su ajuste. A continuación los parámetros del modelo deben ser revisados en orden a considerar si algunas de las variables son realmente necesarias para realizar una predicción ajustada. El uso del estadístico de Wald y del ajuste del modelo en un análisis en el que se eliminan algunas de las variables son instrumentos útiles en este sentido. Las últimas etapas se diferencian poco, si algo, del análisis discriminante: hay que clasificar a los individuos, evaluar la bondad de la clasificación y, finalmente, se interpretan los resultados según el problema de partida.

Cuadro 11.3. Las etapas del análisis de regresión logística



## 8. LECTURAS RECOMENDADAS

- JOVELL, A. J., *Análisis de regresión logística*, Madrid, Centro de Estudios Políticos y Constitucionales, 1995.
- KLEINBAUM, D. G., *Logistic regression: a self-learning text*, Nueva York, Springer-Verlag, 1994.
- SÁNCHEZ, E., *Regresión logística en salud pública*, Granada, Escuela Andaluza de Salud Pública, 2000.

### LECTURAS COMPLEMENTARIAS

- FOX, J. D., *Linear Statistical Models and Related Methods*, Nueva York, Wiley, 1984.
- HOSMER, D. W. y LEMESHOW, S., *Applied Logistic Regression*, Nueva York, Wiley, 1989.
- SHARMA, S., *Applied multivariate techniques*, Nueva York, Wiley, 1996.
- WALSH, A., «Teaching understanding and interpretation of logit regression», *Teaching Sociology*, 15, 1987, págs. 178-183.

## 9. EJERCICIOS

Imagine una investigación en la que pretende predecir la aparición de síntomas de Alzheimer (SA) a partir de la historia de enfermedades (HE, se han padecido frecuentemente o no) y una subescala del WAIS. Los resultados aparecen en la siguiente tabla.

SA	W	HE	SA	W	HE	SA	W	HE
1	10	1	1	10	1	0	14	0
1	14	1	1	8	1	1	14	1
1	7	1	0	15	0	1	9	1
1	9	1	0	10	0	0	16	0
1	11	0	1	10	1	0	11	0
1	5	1	0	12	1	1	12	1
1	15	1	0	14	0	0	13	0
0	9	0	0	16	0	0	5	1
1	12	0	1	12	0	0	15	0
0	7	1	0	6	0	1	20	1

1. Analice los datos mediante regresión logística.
2. Analice los datos mediante regresión lineal.
3. Compare los resultados obtenidos en 1 y 2.
4. Clasifique a los sujetos y valide la solución obtenida.

## RESPUESTAS A LOS EJERCICIOS

1. La regresión logística introduciendo simultáneamente todos los predictores en la ecuación indica que el ajuste es significativo (índice de Hosmer-Lemeshow = 14,0128,  $p < 0,05$ ), aunque el único predictor que hace una aportación significativa es HE (Wald = 8,41,  $p < 0,05$ ).
2. La regresión lineal es significativa,  $F(2,27)=7,85$ ,  $p < 0,05$ . Sólo realiza una aportación significativa la variable HE (Beta estandarizada=0,627,  $p < 0,05$ ).
3. La solución obtenida es muy semejante con ambas aproximaciones. Esto se confirma examinando los residuos de predicción estandarizados, donde encontramos que el número de errores de predicción significativos es equivalente en ambos análisis.
4. Los resultados de clasificar a los sujetos por grupos, respecto del grupo real de pertenencia son los siguientes:

Tabla de contingencia SA\*Predicted Group

Recuento

	Predicted Group		Total
	,00	1,00	
SA	11	3	14
	3	13	16
Total	14	16	30

El ajuste es significativo considerando cualquiera de los estadísticos ofrecidos por SPSS.

## CAPÍTULO XII

### Modelos de ecuaciones estructurales

#### OBJETIVOS DE APRENDIZAJE

- 1) Conocer en qué contextos de investigación puede emplearse análisis factorial confirmatorio y/o modelos de ecuaciones estructurales (SEM).
- 2) Conocer qué tipos de problemas de investigación puede ayudar a resolver SEM.
- 3) Comprender las diferencias entre el análisis causal clásico y SEM.
- 4) Conocer los aspectos fundamentales del diseño de modelos estructurales.
- 5) Comprender la lógica del análisis SEM.
- 6) Comprender la utilidad de los diferentes estadísticos de ajuste.
- 7) Conocer las limitaciones y supuestos del análisis SEM.
- 8) Interpretar la salida que ofrecen los paquetes estadísticos.

#### 1. CONCEPTOS BÁSICOS DEL ANÁLISIS FACTORIAL CONFIRMATORIO Y LOS MODELOS DE ECUACIONES ESTRUCTURALES

Una preocupación importante de muchos investigadores ha sido la de encontrar formas de realizar inferencias de naturaleza causal a partir de datos obtenidos en investigaciones correlacionales. Como hemos indicado ya, en bastantes ocasiones se diseñan estudios en los cuales se introducen solamente variables observadas, unas manipuladas (las independientes) y las otras simple-

mente medidas (las dependientes). En estos casos, la inferencia causal se produce de manera bastante natural, puesto que los cambios observados en las variables dependientes pueden atribuirse, si las condiciones de inferencia se cumplen, a las variables independientes. La cuestión, sin embargo, no es tan simple cuando en la investigación no se introducen variables manipuladas, sino simplemente variables medidas. En estas situaciones, una de las opciones de análisis es el análisis factorial exploratorio (véase Capítulo IV), en el que se asume que la variabilidad de las variables medidas (indicadoras o variables superficiales) se debe a la existencia de variables latentes (o factores). Los factores no son observables directamente y deben ser inferidos a partir del patrón de correlaciones entre las variables superficiales. Cuando no se tienen hipótesis previas sobre la estructura factorial el investigador puede limitarse simplemente a interpretar los factores inferidos. Sin embargo, en muchas ocasiones el investigador tiene ideas bastante claras sobre cuáles pueden ser las variables latentes y sobre qué relaciones puede haber entre ellas y las variables observadas. En estos casos, dispone de dos técnicas muy relacionadas entre sí, el factorial confirmatorio (CFA) y los modelos de ecuaciones estructurales (SEM). Como veremos más adelante, CFA es un caso particular de SEM, por lo que de aquí en adelante lo que indiquemos sobre SEM puede aplicarse de forma prácticamente automática a CFA. Estas dos técnicas tienen un objetivo fundamental, contrastar un modelo teórico sobre las relaciones entre variables independientes y variables dependientes. Tanto unas como otras pueden ser variables latentes o variables observadas, medidas. Para lograr este objetivo el investigador dispone, como en el análisis factorial exploratorio, de un conjunto de datos observados que se resumen en una matriz de varianzas-covarianzas. Así, el objetivo se reduce a estimar un conjunto de parámetros que indiquen la relación (causal o correlacional) entre las variables, de modo que pueda estimarse una matriz de varianzas-covarianzas poblacional que sea lo más parecida posible a la matriz observada, muestral. La matriz estimada dependerá lógicamente del modelo que el investigador quiera contrastar, esto es, del conjunto de parámetros que suponga necesarios para explicar sus datos.

## 2. LAS CLASES DE CUESTIONES QUE SEM AYUDA A RESOLVER

El análisis factorial confirmatorio y el análisis de ecuaciones son técnicas explicativas que pueden emplearse en una gran variedad de contextos de investigación y que permiten resolver un gran número de cuestiones acerca de la explicación causal de un conjunto de variables. Las cuestiones más importantes que SEM y CFA ayudan a resolver son las siguientes:

### 2.1. COMPROBACIÓN DE TEORÍAS

En muchas ocasiones el investigador realiza sus estudios basándose en supuestos teóricos bien establecidos en el dominio de su interés. Por ejemplo, en el ámbito de la personalidad, o de la inteligencia, hay teorías bien establecidas sobre cuáles pueden ser los factores, las variables latentes, que expliquen la ejecución de los individuos en las pruebas de personalidad o inteligencia. Sin embargo, no es necesario que sea un modelo teórico de alto nivel el que se somete a contraste, sino que puede ser un modelo de carácter personal, un simple conjunto de creencias acerca de cuáles pueden ser los factores explicativos, sobre cuáles pueden ser las relaciones entre unas variables y otras. Lo importante es que cualquier modelo teórico realiza predicciones concretas. Las predicciones no son sobre valores de las variables observadas, sino sobre las relaciones entre ellas, esto es, sobre la matriz de varianzas-covarianzas que debería observarse en la población. La predicción de la matriz poblacional sólo puede realizarse cuando se han estimado todos los parámetros del modelo. Los parámetros hacen referencia a las relaciones (causales o de mera correlación) entre las variables observadas y latentes, incluyendo los errores. Evidentemente, la bondad del modelo se evalúa comparando la matriz poblacional con la matriz de varianzas-covarianzas observada en la muestra. Un buen modelo es aquel en el que la semejanza entre las dos matrices es elevada. Como veremos, buena parte del análisis está orientado a la determinación de la bondad de ajuste (la semejanza) entre ambas matrices. Cuando el ajuste se considera suficiente el modelo es empleado para explicar los datos observados. Si no es suficiente, suele ser conveniente modificarlo con el fin de mejorar su capacidad explicativa.

### 2.2. CANTIDAD DE VARIANZA EXPLICADA

Un buen modelo debe ser capaz de explicar una gran cantidad de la variabilidad de las variables observadas. La razón es simple, si un factor se asume que es causa (entre otras) de una variable superficial, parece razonable suponer que el grado en que los individuos poseen ese factor es un determinante de la puntuación en las variables superficiales conectadas causalmente con el mismo.

### 2.3. MEDIACIÓN

Las variables pueden estar relacionadas de forma directa entre sí, o a través de terceras variables. Por ejemplo, en la Figura 12.2. puede apreciarse que la relación entre dígitos hacia delante y el almacén a corto plazo es directa, a su vez, también se postula una relación directa entre el almacén y el factor g de inteligencia general. Sin embargo, la relación entre dígitos hacia delante y el

factor  $g$  es indirecta, mediada por el factor almacén a corto plazo. Si esa relación indirecta existe o no es algo que SEM nos ayuda a responder de forma clara. Por ejemplo, si el parámetro de relación entre almacén y factor  $g$  es cero, dígitos hacia delante y factor  $g$  no están relacionados en absoluto.

#### 2.4. DIFERENCIAS ENTRE GRUPOS E INTRASUJETOS

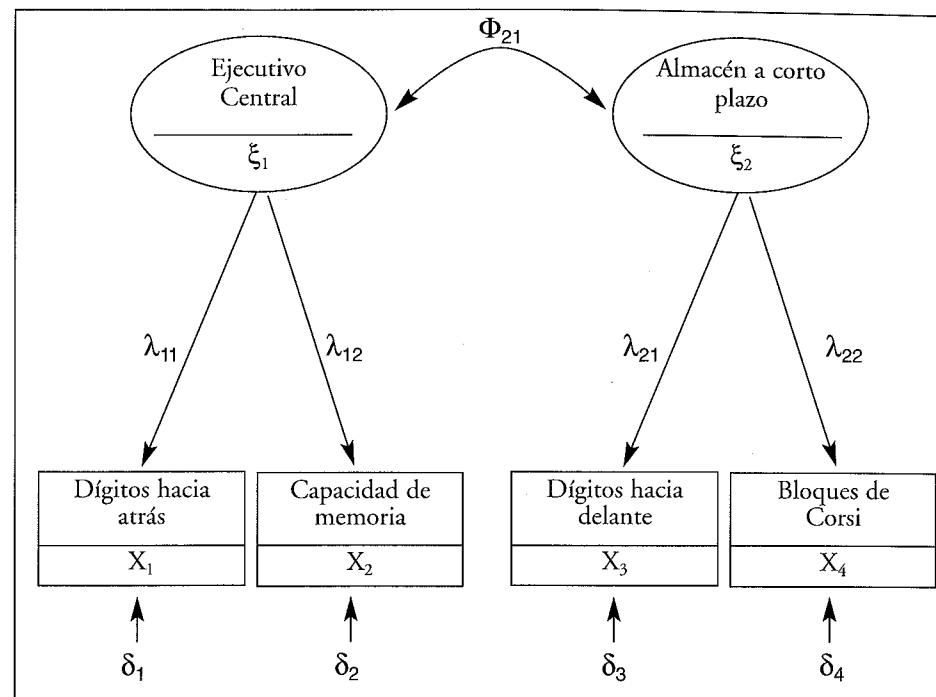
SEM y CFA también ayudan a resolver preguntas relacionadas con diferencias entre distintos grupos o entre los propios sujetos medidos en ocasiones diferentes. La razón es simple, grupos distintos de sujetos pueden diferir en las relaciones entre las variables, o en los pesos de esas relaciones. De la misma forma, las relaciones entre las variables (y/o los pesos de las mismas) pueden cambiar a medida que transcurre el tiempo.

#### 3. CONSTRUCCIÓN DE MODELOS Y ECUACIONES BÁSICAS

En el ejemplo que considerábamos con intenciones meramente ilustrativas en el Capítulo IV, dedicado a análisis factorial exploratorio, se medían cuatro variables superficiales, dígitos hacia delante, dígitos hacia atrás, capacidad de memoria y bloques de Corsi. La literatura sobre memoria operativa parece coincidir en que hay dos factores que pueden explicar la variabilidad observada en esas variables: el ejecutivo central, encargado de la flexibilidad, coordinación y planificación del comportamiento, y el almacén a corto plazo, que se encargaría sólo de retener información de manera temporal. También parece haber acuerdo en que las tareas de dígitos hacia atrás y capacidad de memoria implican más al ejecutivo que al almacén a corto plazo, mientras que las dos tareas restantes implican más al almacén que al ejecutivo. Por otra parte, si la función del ejecutivo es controlar el procesamiento de información, parece razonable pensar que debe haber alguna relación entre él y el almacén. Por lo tanto, parece que el estudio se podría haber orientado en el sentido de contrastar un modelo de memoria operativa, más bien que como meramente exploratorio. Ese modelo, formulado en palabras, puede concretarse de manera mucho más formal empleando un sistema de representación conocido como diagramas de paso (*path diagrams*).

La Figura 12.1. presenta el diagrama del modelo. En la construcción se han utilizado las convenciones habituales. Las elipses contienen las variables latentes y los rectángulos las variables superficiales. Las flechas unidireccionales desde las latentes a las superficiales indican una relación causal (los factores son causas posibles de las variables superficiales), mientras que la flecha bidireccional entre los factores indica la correlación esperada entre las dos variables latentes. Cada flecha tiene asociada una etiqueta, que es un parámetro del modelo.

Figura 12.1. Un modelo de dos factores con la notación habitual de análisis factorial confirmatorio



Así,  $\lambda_{11}$  indica el grado de relación causal entre el factor Ejecutivo Central ( $\xi_1$ ) y la variable superficial Dígitos hacia atrás, mientras que  $\Phi_{21}$  hace referencia a la correlación entre los dos factores. Nótese, además, que se emplea el alfabeto griego para etiquetar las variables latentes y el latino para las variables superficiales. La letra  $\delta$  se reserva para indicar la variabilidad de una variable superficial no explicada por las variables latentes. Es importante caer en la cuenta de que estos diagramas facilitan en gran medida la construcción de un modelo matemático claro de las relaciones supuestas entre variables. Concretamente, asumiendo, lo que es bastante conveniente para simplificar las ecuaciones, que las variables superficiales están expresadas en puntuaciones diferenciales (la puntuación de cada sujeto en la variable menos la media de la variable), tendríamos que:

$$\begin{aligned} X_1 &= \lambda_{11}\xi_1 + \delta_1 \\ X_2 &= \lambda_{12}\xi_1 + \delta_2 \\ X_3 &= \lambda_{21}\xi_2 + \delta_3 \\ X_4 &= \lambda_{22}\xi_2 + \delta_4 \end{aligned}$$

lo que expresamos en forma resumida como:

$$X = \Lambda_x \xi + \delta \quad (\text{Ec. 12.1})$$

Y en forma más explícita, de la siguiente forma:

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ 0 & \lambda_{21} \\ 0 & \lambda_{22} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \end{bmatrix}$$

que es idéntica al conjunto de cuatro ecuaciones estructurales anterior. Sin embargo, debemos tener en cuenta que en el modelo habíamos incluido la correlación entre los dos factores, que no está presente en las ecuaciones. La forma de incluirla consiste en definir una nueva matriz, que llamaremos  $\Phi$ , y que contendrá las varianzas de las variables y la covarianza entre ambas. Evidentemente las dimensiones de la matriz serán, genéricamente,  $f_x f$ , siendo  $f$  el número de factores que se hayan especificado en el modelo. En nuestro caso, la matriz será:

$$\Phi = \begin{bmatrix} \sigma_{\xi_1}^2 & \sigma_{\xi_1 \xi_2} \\ \sigma_{\xi_1 \xi_2} & \sigma_{\xi_2}^2 \end{bmatrix}$$

siempre que los factores, como ahora, se consideren correlacionados. En caso de que no se postule correlación, los elementos fuera de la diagonal serán 0. Por último, en el modelo también habíamos incluido términos de error, puesto que siempre asumimos que las variables superficiales son una medida imperfecta de las variables latentes, por tanto, necesitamos una matriz más que contenga las varianzas-covarianzas error. Que las covarianzas error sean nulas o no depende de manera fundamental de qué supuestos se adopten. En el caso del análisis factorial exploratorio lo común es asumir que los errores no están correlacionados. Sin embargo, en el análisis factorial confirmatorio puede asumirse correlación si, por ejemplo, un factor afecta a más de una variable superficial. En caso de que se adopte este supuesto deberían incluirse en el modelo líneas curvadas con flechas bidireccionales que conecten los términos error relacionados. Suponiendo, por ahora, que las covarianzas son nulas, la matriz de errores, que denotaremos como  $\Theta_\delta$ , léase zeta delta, tendría el siguiente aspecto:

$$\Theta_\delta = \begin{bmatrix} \sigma_{\delta_1}^2 & 0 & 0 & 0 \\ 0 & \sigma_{\delta_1}^2 & 0 & 0 \\ 0 & 0 & \sigma_{\delta_1}^2 & 0 \\ 0 & 0 & 0 & \sigma_{\delta_1}^2 \end{bmatrix}$$

y ahora sí, el modelo está completamente determinado.

En conclusión, un modelo causal queda perfectamente determinado cuando se han especificado tres aspectos: 1) las relaciones estructurales entre las variables superficiales y las variables latentes, definidas en la matriz de coefi-

cientes estructurales  $\Lambda_X$ ; 2) se han especificado las correlaciones no estructurales entre variables latentes, definidas en la matriz  $\Phi$ ; y 3) se han especificado las varianzas-covarianzas entre los errores de medida, definidas en la matriz  $\Theta_\delta$ . Sin embargo, quedan todavía por establecer las relaciones entre esas tres matrices, o si se prefiere entre los componentes del modelo. Para comenzar, hay un aspecto sustancial del modelo que, quizás por obvio, hemos demorado hasta este punto. Los datos que el investigador tiene disponibles y que somete a análisis son relativos a las variables superficiales, mejor aún, a las varianzas-covarianzas entre estas variables. Siguiendo las convenciones al uso denotaremos esta matriz, que el lector ya sabe puede obtener fácilmente multiplicando la matriz transpuesta de las puntuaciones diferenciales por la matriz de puntuaciones, también diferenciales, como  $\Sigma$ . A partir de aquí la derivación es sencilla, aunque innecesaria para nuestros propósitos:

$$\Sigma(\theta) = E(X'X) = \Lambda_X \Phi \Lambda_X + \Theta_\delta \quad (\text{Ec. 12.2})$$

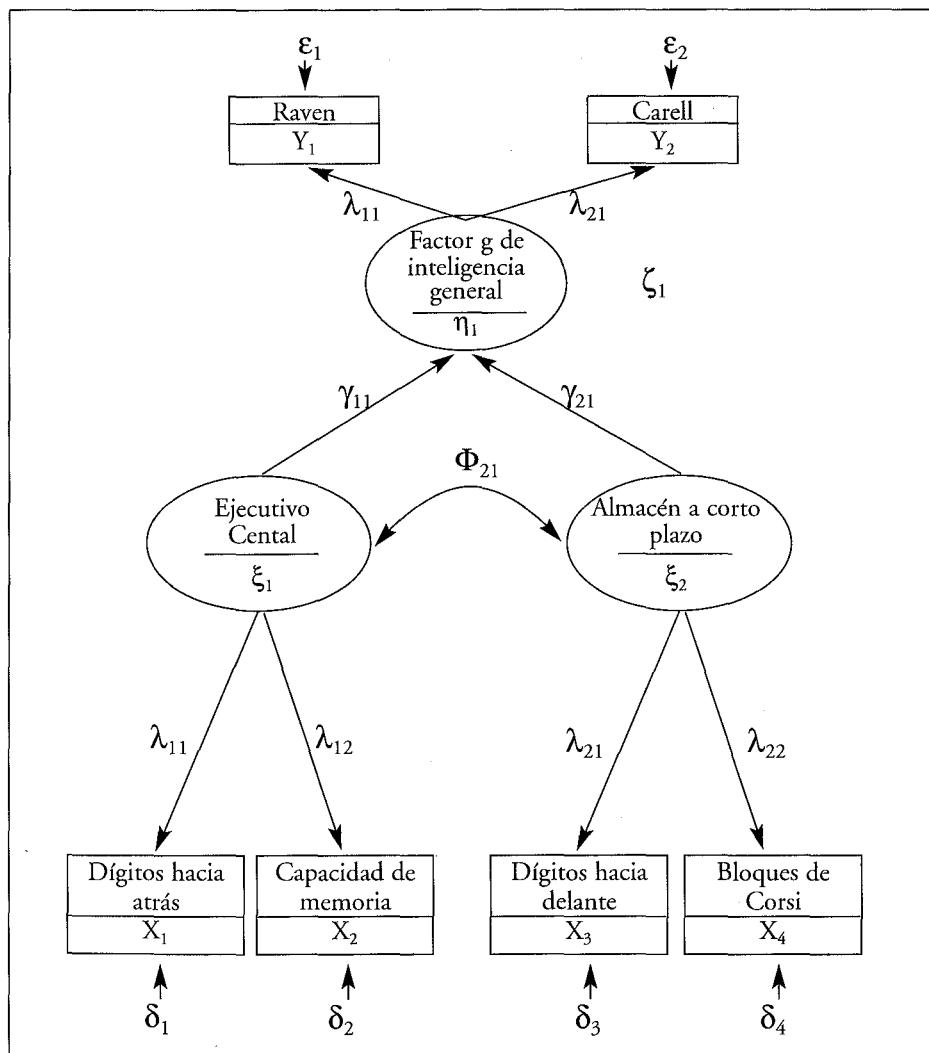
o, lo que es lo mismo, la matriz estimada de varianzas-covarianzas de las variables superficiales es una función de la matriz de coeficientes estructurales, de la matriz de covarianzas entre los factores y de la matriz de covarianzas error. Esta matriz de varianzas-covarianzas se llama matriz implicada por el modelo. Cuando un modelo es adecuado la matriz implicada por el modelo será igual a la matriz poblacional. Por tanto, las discrepancias entre la matriz poblacional y la estimada a partir del modelo constituirán el indicador adecuado para evaluar a éste. Esto es, cuando el modelo es adecuado  $\Sigma = \Sigma(\theta)$

### 3.1. EL MODELO GENERAL DE ECUACIONES ESTRUCTURALES

Hasta aquí hemos considerado la construcción de modelos relativamente sencillos en los que tenemos solamente dos niveles, variables superficiales o indicadores y variables latentes o factores. Éstos son los modelos que se emplean de manera típica en el análisis factorial confirmatorio. Sin embargo, podemos expandir esta lógica para abrazar un conjunto más amplio de modelos cuya complejidad puede llegar a ser extraordinaria. Por ejemplo, consideremos ahora la posibilidad de que los dos factores que hemos definido, ejecutivo central y almacen a corto plazo contribuyan de manera importante al factor  $g$  de inteligencia general. Por otra parte este factor podemos evaluarlo mediante test como las matrices progresivas de Raven y el test de inteligencia fluida de Cattell. En este caso el modelo debería incorporar una variable latente adicional que estaría relacionada causalmente con los dos factores que ya hemos definido. Examinemos con detalle el modelo de la Figura 12.2., porque muestra las características más relevantes de los modelos contrastables mediante SEM. En primer lugar, recordemos que las flechas direccionales representan relaciones

causales mientras las bidireccionales representan correlaciones. En segundo lugar, observemos que tenemos dos tipos de factores. Unos que no tienen nexos más que con variables superficiales, a los que denominaremos factores o variables latentes exógenas, y otros que están conectados causalmente con los anteriores, aunque también son causa de variables superficiales, son las variables latentes endógenas. Los factores exógenos son etiquetados con la letra griega  $\xi$ , y los endógenos con la letra  $\eta$ . En cada caso se añade un subíndice para indicar el número de factor.

Figura 12.2. Un modelo estructural completo de las relaciones entre memoria operativa e inteligencia general (Adaptado de Engle y cols., 1999)



Las relaciones entre los factores constituyen el modelo estructural, en el que se asume normalmente, aunque no necesariamente, que las relaciones son lineales. Las relaciones causales (o de regresión) están indicadas en el modelo por una flecha direccional etiquetada con la letra  $\gamma$  cuando conecta un factor exógeno con uno endógeno, o con la letra  $\beta$  cuando un factor endógeno regresa sobre otro endógeno. Cuando las relaciones son de covariación, las flechas son bidireccionales y se etiquetan con la letra  $\phi$ . Los factores endógenos sólo pueden tener entre sí relaciones causales, no de correlación.

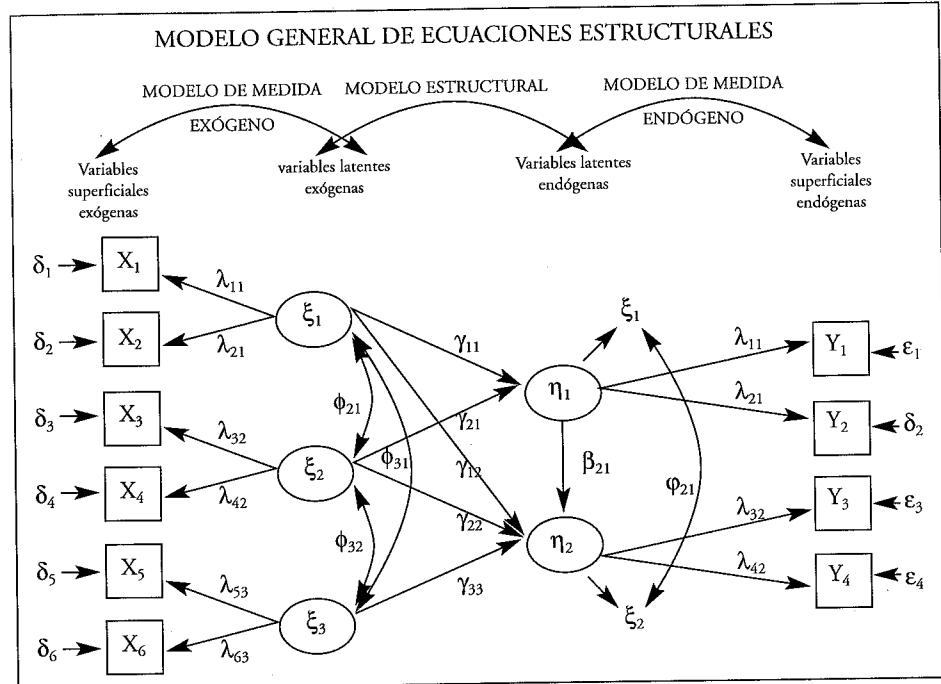
En la mayoría de las ocasiones los investigadores no esperan poder predecir un factor endógeno de forma perfecta, de modo que suelen incluir errores de predicción que se denominan errores estructurales y se etiquetan con la letra  $\zeta$ . Estos errores no están correlacionados con los cometidos en la predicción de factores exógenos. Si esto ocurre debe considerarse seriamente la exclusión de las variables superficiales que producen la correlación. Sin embargo, cuando hay más de un factor endógeno es posible que sus errores estructurales covarien entre sí, lo que se expresa mediante los parámetros etiquetados como  $\varphi$ . Esta relación indicaría que los factores endógenos comparten varianza que no es explicada por los factores exógenos.

Las variables superficiales pueden predecir factores exógenos, entonces se etiquetan como  $X_i$ , o pueden ser explicadas por factores exógenos o endógenos, y entonces se etiquetan como  $Y_i$ . El papel de unas y otras pueden cambiar de un modelo a otro. Estas variables, junto con los factores con las que están conectadas constituyen el modelo de medida. El modelo más frecuente es conocido como congenérico, caracterizado por la simplicidad de las conexiones, puesto que cada variable superficial está conectada solamente con un factor, y, además las covariaciones se deben a las relaciones entre factores y variables superficiales. Las relaciones de ambas con sus factores respectivos están etiquetadas con la letra  $\lambda$ , de modo que pueden tenerse dos matrices de parámetros,  $\Lambda_X$  y  $\Lambda_Y$ , respectivamente para las conectadas con factores exógenos y endógenos. Finalmente, se asume que cada variable  $X$  e  $Y$  puede tener asociado un error de medida. Los asociados a variables  $X$  se etiquetan con la letra  $\delta$ , y los asociados a variables  $Y$  mediante  $\varepsilon$ .

Un modelo más completo, incluyendo todos los parámetros, y los tipos de modelos incluidos, estructural y de medida aparece en la Figura 12.3. Esta distinción de tipos es importante, porque nos pone sobre la pista de cómo se puede identificar y especificar un modelo general de ecuaciones estructurales. La construcción del modelo general, como el lector habrá notado ya, es aparentemente compleja. Sin embargo, no debe dejarse llevar por la apariencia, en realidad la cuestión es bastante simple. Para empezar, caigamos en la cuenta de que el modelo tendrá que especificar tres sistemas de ecuaciones, dos correspondientes a los modelos de medida, uno exógeno y otro endógeno, y el tercero correspondiente al modelo estructural. Recordemos que el modelo de medidas exógeno especifica las relaciones entre variables superficiales y factores exógenos, mientras que el modelo de medidas endógeno especifica las relacio-

nes entre los factores endógenos (que sólo reciben entradas de otros factores) y variables superficiales endógenas (que son explicadas por factores endógenos). Veamos cada uno de ellos.

Figura 12.3. Un modelo completo de ecuaciones estructurales



La parte estructural del modelo general debe indicar cómo están relacionados los factores endógenos,  $\eta$ , entre sí, y con respecto a los factores exógenos,  $\xi$ . Lógicamente estas relaciones dependen de dos grupos de parámetros, los que ligan a los factores endógenos,  $\beta$ , y los que ligan a éstos con los exógenos,  $\gamma$ . Así, por ejemplo, examinando la figura veremos que el factor  $\eta_2$  se obtendría, siguiendo el árbol de flechas, como:

$$\eta_2 = \beta_{21} \eta_1 + \gamma_{22} \xi_2 + \gamma_{32} \xi_3 + \zeta_2$$

y generalizando tendríamos que

$$\eta = B\eta + F\xi + \zeta \quad (\text{Ec. 12.3})$$

es decir:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \beta_{21} & 0 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} \gamma_{11} & \gamma_{21} & 0 \\ 0 & \gamma_{22} & \gamma_{32} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix}$$

Nótese que algunos parámetros son fijos (tienen valor cero) mientras que otros son libres, tienen que ser estimados a partir de los datos observados.

El modelo de medidas exógeno puede obtenerse de manera similar. Examinando el modelo observaremos que se trata de predecir las variables superficiales desde los factores que están conectados con ellas. Por ejemplo, la variable  $X_1$ , se obtendrá como:

$$X_1 = \lambda_{11} \xi_1 + \delta_1$$

y generalizando tendremos que:

$$X = \Lambda_X \xi + \delta \quad (\text{Ec. 12.4})$$

que ya hemos desarrollado antes.

Finalmente, el modelo de medidas endógeno puede resolverse de manera semejante al exógeno, teniendo en cuenta que ahora se trata de predecir las variables superficiales endógenas desde los factores endógenos que las explican. Por ejemplo, la variable  $Y_1$  será:

$$Y_1 = \lambda_{11} \eta_1 + \varepsilon_1 ,$$

lo que generalizado quedará de la siguiente forma:

$$Y = \Lambda_Y \eta + \varepsilon \quad (\text{Ec. 12.5})$$

y cuya expansión es semejante a la del modelo de medidas exógeno.

Así pues, necesitamos cuatro matrices estructurales,  $B$ ,  $F$ ,  $\Lambda_X$  y  $\Lambda_Y$ , más un conjunto de matrices de varianzas covarianzas,  $\Phi$ , para los factores exógenos,  $\Psi$ , para los errores de los factores estructurales,  $\Theta_\delta$ , para los errores de medida de las variables superficiales exógenas, y  $\Theta_\varepsilon$ , para los errores de medida de las variables superficiales endógenas. El orden en que estas matrices deben ser estimadas es paralelo a la secuencia de la Figura 12.3. Es relativamente fácil comprobar que este modelo general puede reducirse a un modelo de análisis factorial confirmatorio. Para ello basta con hacer  $B=0$ ,  $F=0$ ,  $\Psi=0$ ,  $\Lambda_Y=0$  y  $\Theta_\varepsilon=0$ . Esto es lo mismo que eliminar los factores endógenos y las variables superficiales endógenas.

### 3.2. ESTIMACIÓN DE PARÁMETROS

La estimación de los parámetros del modelo puede realizarse mediante dos procedimientos, no incompatibles entre sí: mínimos cuadrados y máxima verosimilitud. El procedimiento de mínimos cuadrados trata de hacer mínima la suma de los cuadrados de error (las diferencias entre la matriz de varianzas-covarianzas poblacional y la implicada por el modelo). Su empleo suele limitarse, aunque no necesariamente, a la estimación de un conjunto de valores iniciales para los parámetros, que luego será ajustado mediante el procedimiento de máxima verosimilitud. Este es un método iterativo, en el cual los valores de los parámetros se van ajustando de manera que las dos matrices, estimada y observada sean lo más semejantes posible. La semejanza entre las dos matrices se evalúa en cada paso, de manera que si la diferencia es menor que un criterio el proceso termina y se considera que se han obtenido los mejores estimados posible de los parámetros, y en caso contrario, éstos vuelven a ser modificados. La diferencia fundamental entre los dos procedimientos reside, pues, no en el objetivo, sino en la función que se emplea para valorar si el objetivo se ha alcanzado. Esto hace que, asintóticamente, los dos métodos coincidan en sus resultados. Sin embargo, la función empleada por el método de máxima verosimilitud, implica un supuesto importante, que las observaciones han sido extraídas de una distribución normal multivariada.

### 3.3. IDENTIFICACIÓN DEL MODELO

Para que un modelo de estas características pueda ser estimado, es decir, para que el conjunto de parámetros que implica puedan ser estimados es preciso, antes de nada, que su número no sea superior al número de varianzas y covarianzas entre las variables superficiales (incluyendo exógenas, X, y endógenas, Y). En otras palabras, el requisito establece que los grados de libertad del modelo deben ser mayores o iguales que cero. El concepto de grados de libertad en el contexto de SEM hace referencia a la diferencia entre el número de ecuaciones y el número de parámetros estructurales del modelo. Si  $p+q$  es el número de variables observadas ( $p$  exógenas y  $q$  endógenas), entonces, el número de ecuaciones es

$$\text{ecuaciones} = \frac{(p+q)(p+q+1)}{2}$$

y si llamamos  $\pi$  al número de parámetros, entonces los grados de libertad serán igual a ecuaciones - $\pi$ .

Sin embargo, no es suficiente que los grados de libertad sean mayores o iguales que cero para que el modelo sea estimable. Además es necesario que todos los factores (exógenos y endógenos) tengan asignada una unidad de medida, lo que no ocurre en el modelo de la Figura 12.3. Para dotar de unidad de medida a una variable latente podemos proceder de dos formas diferentes. Primero, considerarla como una variable estandarizada, lo que implicaría que su varianza sería igual a la unidad. Segundo, podemos fijar uno de sus coeficientes estructurales asociados ( $\lambda$  o  $\beta$ ) a uno, lo que hace que su escala sea igual a la de una de las variables superficiales (X o Y, respectivamente). Finalmente, cuando una variable latente tiene asociada sólo una variable superficial, lo que no ocurre en el modelo de la Figura 12.3., es preciso suponer que esa variable ha sido medida sin error. Por ello, es preferible medir más de dos variables superficiales por cada variable latente.

### 3.4. LOS EFECTOS DIRECTOS, INDIRECTOS Y TOTALES

En el contexto de SEM, se distinguen tres clases diferentes de efectos. Los efectos directos son los que produce una variable sobre otra que está conectada de forma inmediata con ella. En el diagrama del modelo se producen estas conexiones directas entre variables latentes endógenas, entre variables latentes exógenas y variables latentes endógenas, entre variables superficiales exógenas y latentes exógenas y entre latentes endógenas y superficiales endógenas. Por tanto, en el primer caso, los efectos están definidos en los coeficientes de B, en el segundo en los coeficientes de F, y en los dos casos restantes en  $\Lambda_X$  y  $\Lambda_Y$ , respectivamente.

Los efectos indirectos son los que se producen entre variables que no están conectadas directamente, sino a través de una tercera variable. Por ejemplo, en el diagrama de la Figura 12.3. puede observarse que la variable  $\eta_2$  sufre un efecto indirecto de  $\xi_1$  a través de  $\eta_1$ , además del efecto directo. Los efectos indirectos se calculan mediante el producto de las conexiones que median entre las dos variables relacionadas. Así, el efecto indirecto sufrido por  $\eta_2$  desde  $\xi_1$  será igual al producto  $\gamma_{11}\beta_{21}$ . Finalmente, los efectos totales entre dos variables son la suma de los efectos directos y los indirectos. Lo interesante del análisis de la significatividad de estos efectos es que nos pone sobre la pista de cuáles son las explicaciones de la variabilidad de un factor.

### 4. AJUSTE DEL MODELO

Una vez que se ha especificado el modelo y se han estimado los parámetros la cuestión más importante a resolver concierne a la valoración de su ajuste global. El ajuste hace referencia a la capacidad que tiene el modelo de predecir la matriz de varianzas-covarianzas entre variables superficiales. Naturalmente ésta es, además, la primera etapa de la interpretación de los resultados obtenidos. Si

el modelo ajusta de manera aceptable, entonces puede procederse a evaluar e interpretar los parámetros. Por el contrario, si el ajuste no es adecuado, puede ser conveniente modificar el modelo, por ejemplo, suprimiendo alguno de los factores o cambiando algunas de las conexiones hipotetizadas. El ajuste global puede realizarse mediante un test bien conocido como la Chi-cuadrado, pero no es frecuente utilizar una sola medida para valorarlo, sino un conjunto amplio, en el que se incluyen varios estadísticos heurísticos. Sin embargo, es aconsejable, antes de proceder a una evaluación estadística, considerar la plausibilidad de los valores estimados de los parámetros. La razón es simple, valores superiores a uno son imposibles, por lo que inmediatamente cabe descartar un modelo en el que alguno de los parámetros sobreponga ese límite superior. De la misma forma, la aparición de varianzas estimadas negativas indicaría que el modelo no es adecuado, por lo que es innecesario evaluarlo más y sencillamente debe dedicarse el esfuerzo a modificarlo.

#### 4.1. VALORACIÓN MEDIANTE CHI-CUADRADO

La chi-cuadrado se emplea para contrastar la hipótesis de que la matriz de varianzas-covarianzas poblacional ( $\Sigma$ , cuyo mejor estimador es la de varianzas-covarianzas observada,  $S$ ) es igual a la matriz implicada por el modelo ( $\Sigma(\theta)$ ). Puesto que la hipótesis que se contrasta es la de igualdad, que las matrices sean iguales, si la chi-cuadrado es significativa, entonces cabe concluir que el modelo no se ajusta. El problema de este estadístico reside en su sensibilidad al tamaño muestral. Con muestras grandes, pequeñas diferencias entre ambas matrices pueden producir valores significativos de chi-cuadrado. Esto lo hace poco adecuado para tomar decisiones en SEM y PCA, donde el número de sujetos suele ser bastante elevado.

#### 4.2. VALORACIÓN MEDIANTE MEDIDAS HEURÍSTICAS

Alrededor de treinta medidas de ajuste han sido propuestas para tomar decisiones sobre el ajuste del modelo (Marsh, Balla y McDonald, 1988). Sin embargo, los paquetes estadísticos ofrecen solamente unas cuantas basadas en las diferencias entre la matriz poblacional y la matriz implicada por el modelo, esto es, sobre la matriz residual. Las más importantes son las siguientes.

##### 4.2.1. Índice de bondad de ajuste (GFI)

Mide la cantidad de variabilidad que es explicada por el modelo. Se trata, por tanto, de un estadístico que puede interpretarse de una manera similar al coeficiente de correlación múltiple. El índice se obtiene de la siguiente forma:

$$GFI = 1 - \frac{tr((\bar{\Sigma}^{-1}S - I)^2)}{tr[(\bar{\Sigma}^{-1}S)^2]} \quad (\text{Ec. 12.6})$$

GFI adopta valores en el intervalo 0-1. Es 1 cuando la matriz predicha por el modelo y la observada son idénticas y menor que 1 cuanto más diferentes son, cuanto más grande es la matriz residual. Una regla habitual para tomar la decisión de que el modelo es ajustado es exigir un GFI mayor de 0,90. Pero este un valor arbitrario sin mayor fundamento, por ello, se recomienda frecuentemente que se tome la decisión comparando los GFI de diferentes modelos. Esta estrategia es quizás la más ampliamente utilizada en la literatura actual. El problema más importante de este índice es también su dependencia del tamaño de la muestra y del número de variables superficiales, aunque en sentido contrario a chi-cuadrado. Cuando el número de sujetos es muy alto el valor de GFI tiende a situarse siempre por debajo de 1. Por otro lado, cuanto mayor es el número de variables superficiales, menor es el valor que GFI puede adoptar. Esto lo convierte en una medida conservadora del ajuste. Una alternativa, propuesta por Sharma (1996) para corregir la infraestimación del ajuste, consiste en utilizar una medida relativa de bondad de ajuste (RGFI):

$$RGFI = \frac{GFI}{EGFI} \quad (\text{Ec. 12.7})$$

donde EGFI (la bondad de ajuste esperada) es aproximadamente  $1/(1+2v/np)$ , siendo  $v$  los grados de libertad,  $p$  las variables superficiales y  $n$  el número de sujetos.

##### 4.2.2. Índice ajustado de bondad de ajuste (AGFI)

También en este caso se trata de un índice interpretable como el coeficiente de determinación, en el sentido de que mide la variabilidad explicada por el modelo. La diferencia con respecto a GFI reside en que tiene en cuenta los grados de libertad del modelo y el número de variables superficiales. Se calcula como:

$$AGFI = 1 - \left[ \frac{p(p+1)}{2v} \right] [1 - GFI] \quad (\text{Ec. 12.8})$$

Su problema más importante reside en que no hay modo objetivo de determinar qué valores indican un ajuste significativo. Puesto que el ajuste que se

realiza de GFI hace que AGFI disminuya, frecuentemente se adopta como punto de referencia un valor de 0,80. Pero de nuevo se trata de un criterio arbitrario, y, por tanto, otra vez es recomendable que se comparan los valores obtenidos para diferentes modelos.

#### 4.2.3. Raíz de la media de cuadrados residual

Este índice se computa sumando los cuadrados de las diferencias entre las varianzas-covarianzas observadas y las predichas por el modelo. Concretamente:

$$RMSR = \sqrt{\frac{\sum_{i=1}^p \sum_{j=1}^i (s_{ij} - \hat{s}_{ij})^2}{p(p+1)/2}} \quad (\text{Ec. 12.9})$$

Evidentemente, cuanto mayores sean las diferencias entre las varianzas-covarianzas observadas y las predichas, mayor será el valor de RMSR, y menor el ajuste del modelo. Su problema más importante es que las covarianzas son dependientes de la escala y pueden adoptar cualquier valor, no tienen límite superior, lo que hace prácticamente imposible fijar un criterio para tomar la decisión sobre si el modelo se ajusta o no de manera significativa. De nuevo, es recomendable que se comparan las RMSR de varios modelos alternativos para seleccionar el que mejor ajuste.

#### 4.2.4. Índices de no centralidad

Se trata de índices en los que se corrige chi-cuadrado, teniendo en cuenta los grados de libertad y el número de sujetos. El Parámetro de no centralidad reescalado (NCP) se obtiene como:

$$NCP = \frac{\chi^2 - v}{n} \quad (\text{Ec. 12.10})$$

y adopta valores entre cero e infinito. Más fácil de interpretar es la modificación de MacDonald del parámetro (MDN), que se obtiene como

$$MDN = e^{-0.5(NCP)} \quad (\text{Ec. 10.11})$$

que adopta valores entre 0 y 1. Los valores cercanos a 0 son indicativos de un mal ajuste del modelo y los cercanos a 1, lo son de un buen ajuste.

Una segunda modificación es la de Tucker-Lewis (TLI), que implica la comparación del modelo de interés contra un modelo de referencia (o de línea de base), en el que puede suponerse cualquier tipo de relación entre las variables, aunque lo más frecuente es precisamente que se asuma que todos los parámetros son cero, es decir, no relación entre variables y factores. El índice se computa como:

$$TLI = \frac{\frac{NCP_{LB}}{v_{LB}} - \frac{NCP_H}{v_H}}{\frac{NCP_{LB}}{v_{LB}}} \quad (\text{Ec. 12.12})$$

siendo  $NCP_{LB}$  el índice NCP para el modelo de referencia,  $NCP_H$  el del modelo de interés y  $v_{LB}$  y  $v_H$  los grados de libertad asociados a cada uno de ellos. El índice relativo de no centralidad (RNI) es semejante al anterior excepto en que no incluye los grados de libertad de ninguno de los modelos en la ecuación de cálculo.

En conjunto los índices de no centralidad tienen el mismo problema que hemos comentado en los índices heurísticos anteriores: no hay un criterio objetivo que permita decidir si el modelo ajusta de modo significativo. Aunque en bastantes casos se emplea un punto de corte de 0,90, pero de nuevo es recomendable que el investigador pruebe distintos modelos y compare los índices obtenidos con cada uno de ellos. El modelo que mejor ajuste puede ser el elegido para explicar los datos observados.

#### 4.2.5. ¿Por qué no ajusta un modelo?

Los paquetes estadísticos que incorporan SEM presentan varias matrices que pueden ayudar a discernir las causas por las cuales un modelo no se ajusta de manera adecuada a los datos. Como hemos visto, los índices anteriores se basan todos en un resumen de la matriz de residuos (*fitted residuals*, la diferencia entre la observada y la predicha). Esta matriz contiene los errores de predicción del modelo, por lo que su examen detenido puede proporcionar información más exhaustiva de cuáles son las partes en las que el modelo falla. Sin embargo, los residuos son dependientes de la escala (porque lo son las varianzas-covarianzas de las que se derivan), lo que hace difícil evaluar de forma absoluta cualquiera de ellos. Naturalmente siempre es posible hacer una evaluación relativa, comparando unos residuos con otros dentro de la matriz. Sin embargo, una mejor aproximación podría consistir en estandarizarla dividiendo cada ele-

mento por su error estándar correspondiente, lo que equivaldría a obtener puntuaciones z. Esto permitiría tomar una decisión individual sobre cada residuo. Por ejemplo, un valor z mayor o igual de 1,96 tiene una probabilidad asociada de 0,05. Por tanto valores mayores o iguales que 1,96 son índices de un mal ajuste en esa variable. Desafortunadamente, el error estándar depende del tamaño de la muestra, hasta el punto de que con muestras grandes errores insignificantes pueden ser significativos. Parece, pues, que no hay más remedio que recurrir a un conjunto de medidas para tomar una decisión sobre el ajuste. Por ejemplo, se puede examinar la matriz residual no estandarizada para comparar los tamaños de los errores y obtener así un índice subjetivo de ajuste. Después pueden utilizarse otros índices para corroborar esa impresión subjetiva, como la RMSR.

#### 4.2.6. Evaluación de los parámetros del modelo

Cuando el ajuste global es inadecuado no queda más remedio que abandonar el modelo y tantear otros alternativos. Sin embargo, si el ajuste global es suficiente, la próxima etapa del análisis consiste en evaluar los parámetros supuestos en el modelo. ¿De qué forma puede hacerse esta evaluación?

En primer lugar, los paquetes estadísticos proporcionan los valores estimados de los parámetros estandarizados. Es importante tener en cuenta que la estandarización no se realiza con respecto a las variables observadas, sino a las variables latentes. Como sabemos, un parámetro estandarizado no debería adoptar valores fuera del intervalo {-1, 1}. Sin embargo, la forma de estandarizar puede hacer que algunas variables tengan parámetros asociados que excedan esos límites. Una forma de evitarlo es la estandarización completa, en la que se tienen en cuenta también las variables superficiales, además de las latentes. Si en esta situación de estandarización completa, se observan parámetros que exceden los límites del intervalo, entonces el proceso de estimación no ha sido correcto y el modelo no es admisible. Otras causas de no admisibilidad son la aparición de varianzas negativas, o de varianzas error mayores que la unidad.

Una vez que se ha decidido que el modelo es admisible, el paso siguiente consiste en evaluar cada uno de los parámetros de forma individual. Para ello se emplea la t de Student. Solamente deben considerarse significativos los parámetros cuya t tenga un nivel de significación asociado de 0,05 o menor.

Si los parámetros son significativos, la cuestión siguiente a resolver es si las variables superficiales son buenos indicadores de los factores supuestos. Hay varias formas de evaluar este aspecto. En primer lugar, la varianza de una variable superficial puede considerarse compuesta de dos partes, una que es explicada por el factor asociado causalmente a ella, esta es la comunalidad, y otra que puede considerarse error. Evidentemente cuanto mayor es la comunalidad, mayor es la proporción de varianza de la variable superficial explicada por el factor. Por tanto, un buen indicador debe tener una comunalidad alta. No hay

una forma objetiva de decidir qué comunalidad es lo suficientemente alta, pero puede ser una buena idea considerar que por encima de 0,50 la variable superficial es adecuada respecto del factor. Los paquetes proporcionan la comunalidad bajo el nombre de cuadrado del coeficiente de correlación múltiple. El problema de esta aproximación es que obliga a un análisis variable a variable. Una forma más rápida de evaluar si el conjunto de las variables superficiales es un buen indicador consiste en computar el coeficiente de determinación total, según la siguiente expresión:

$$CDT = 1 - \frac{|\Theta_{\delta}|}{|S|} \quad (\text{Ec. 12.13})$$

donde como sabemos  $\Theta_{\delta}$  es la matriz de varianzas-covarianzas error de las variables superficiales exógenas y S es la matriz de varianzas-covarianzas observada. Para variar, el problema reside otra vez en decidir si el índice es lo suficientemente alto. No hay criterios objetivos, pero puede suponerse que por encima de 0,80, el conjunto de variables superficiales es un buen indicador.

#### 4.3. REESPECIFICACIÓN DEL MODELO

Hemos indicado antes en diversas ocasiones que cuando un modelo no se ajusta a los datos es preciso decidir el curso de acción futuro. En principio hay tres alternativas claramente distinguibles entre sí: rechazar el modelo, comprobar un modelo alternativo y modificar el modelo.

En primer lugar, es evidente que un modelo que no ajusta de modo suficiente puede ser descartado. Sin embargo, no es esta una opción común entre los investigadores, quizás porque es posible que el modelo no sea erróneo más que de una manera parcial. Más frecuente es que el investigador suponga que puede haber más de un modelo posible. Por ejemplo, podríamos diseñar un modelo alternativo al de la Figura 12.2., simplemente suponiendo que el almacén a corto plazo no tiene influencia sobre el factor g. La estrategia sería entonces comprobar el ajuste de los modelos alternativos y compararlos con el modelo principal propuesto. La cuestión puede reducirse entonces a utilizar el modelo que mejor ajuste presenta. La comparación entre los diferentes modelos puede realizarse comparando los índices de ajuste mencionados anteriormente, sin embargo, es bastante frecuente en este contexto el empleo del criterio de información de Akaike (AIC), que se basa en la chi-cuadrado, pero penaliza en función del número de parámetros libres en el modelo y del criterio de información consistente de Akaike (CAIC), que penaliza también según el número de sujetos.

La opción más frecuentemente empleada consiste en modificar el modelo, partiendo de la información adquirida con el modelo contrastado, principal-

mente la proporcionada por los estadísticos de ajuste y los residuos (la diferencia entre la matriz predicha y la observada). Los errores que pueden cometerse en la construcción de un modelo son de dos tipos, externos e internos (Kaplan, 1990). Por una parte pueden haberse omitido variables importantes. Estos errores externos se producen fundamentalmente cuando el dominio que se está estudiando no es bien conocido por el investigador, o simplemente se trata de un dominio nuevo en el que no hay suficiente información disponible. Lógicamente la subsanación de esta clase de errores depende más de cuestiones de conocimiento teórico general del campo que de elementos intrínsecos al estudio. Por otra parte, pueden haberse omitido relaciones importantes entre variables, aunque también pueden haberse introducido algunas innecesarias. Estos errores internos son los subsanables en el proceso de modificación de un modelo. Sin embargo, es preciso advertir que la modificación del modelo debe realizarse teniendo en cuenta que SEM y CFA son técnicas orientadas al contraste de hipótesis, y que no cualesquiera modificaciones serán válidas. En otras palabras, la modificación no es resumible a una estrategia de ensayo y error en la que se suprimen o se añaden relaciones cuyo contenido teórico puede ser dudoso. Una modificación exenta de fundamentación teórica es injustificable y debe descartarse.

Prácticamente las únicas modificaciones posibles del modelo consisten en reducir el número de parámetros fijos (iguales a cero), incrementando el número de parámetros libres (que han de ser estimados a partir de los datos). Nótese que esto es lo mismo que aceptar que una relación inexistente (dado su peso 0) se considera ahora real. Para tomar decisiones en este sentido el investigador dispone de dos índices importantes, el de modificación (MI) y el de cambio esperado del parámetro (EPC). Estos índices se denominan también multiplicadores de Lagrange. El índice MI mide el cambio que se produciría en la  $\chi^2$  si el parámetro pasa de fijo a libre. Por su parte EPC es una estimación del valor del parámetro que cambió de fijo a libre. En general, se considera que no deben pasar a libres los parámetros fijos que tienen un valor bajo de EPC; sea MI alto o bajo, no deben pasar a libres. Por el contrario, los parámetros con alto MI y alto EPC deben ser liberados, admitiéndose entonces una nueva relación entre variables en el modelo. No es ésta, sin embargo, la única estrategia posible. Una alternativa consiste en considerar también el valor de  $\chi^2$  del modelo. Si es muy bajo, puede ser indicativo de que algunos parámetros puedan pasarse de libres a fijos (eliminando la relación que implicaban) sin que el modelo sufra en su capacidad predictiva. Un valor alto sería indicativo de que la capacidad predictiva mejoraría si algunos parámetros pasan de libres a fijos. Cuáles sean estos parámetros puede decidirse examinando los índices de modificación MI y EPC.

Esta aproximación de modificación de modelos adolece de un par de problemas importantes. En primer lugar, es posible ajustar un modelo a unos datos sin que el modelo tenga relevancia teórica. Es relativamente fácil que el cambio en los parámetros de un modelo que no ajusta bien lleve a un modelo que

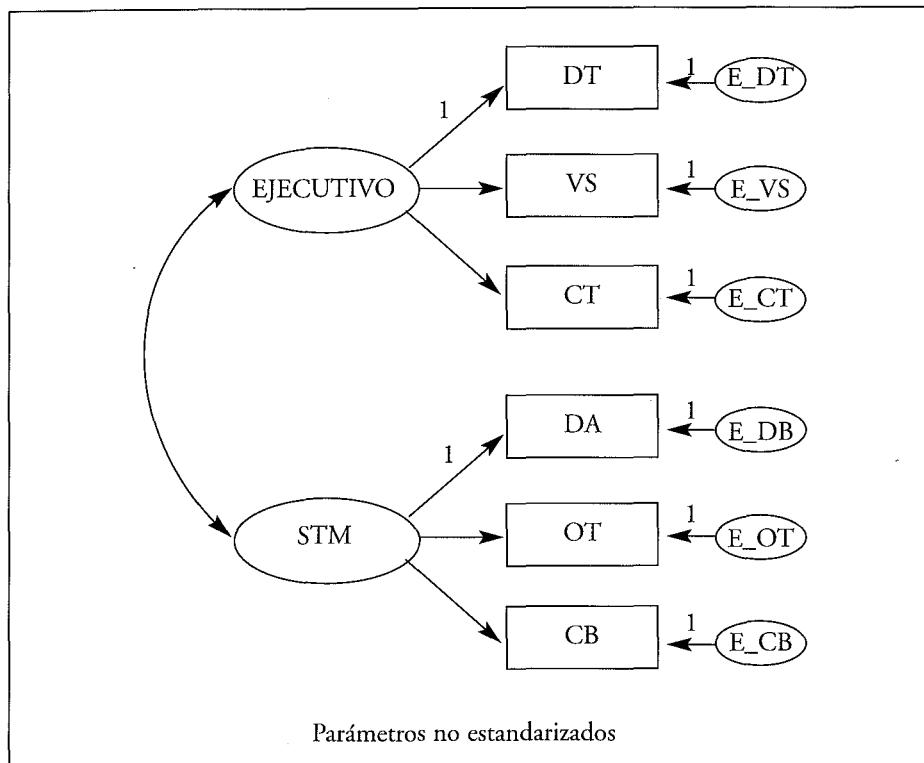
ajuste mejor, pero eso no implica que el nuevo modelo sea adecuado. En este sentido es importante tratar de validar el nuevo modelo. Como hemos visto en capítulos anteriores, la validación puede hacerse de tres formas diferentes. En primer lugar, puede evaluarse el modelo en una muestra y validararlo en otra muestra diferente medida al efecto. Esto no siempre es posible, debido al esfuerzo económico y de tiempo que puede implicar. Por ello, es más frecuente tratar de validar el modelo en la propia muestra. Cuando el número de sujetos es suficientemente grande, puede dividirse en dos mitades de manera aleatoria, en una se estima el modelo y en la otra se valida. Puede obtenerse un índice de validación cruzada (CVI) comparando la matriz de varianzas-covarianzas observada en la muestra de validación, con la predicha por el modelo en la muestra de calibración (estimación). Naturalmente, cuanto más semejantes sean ambas, mayor validez tiene el modelo. El problema es, sin embargo, que no hay un modo objetivo de definir lo que es una semejanza lo suficientemente grande entre ambas matrices. De nuevo, se impone comparar modelos alternativos. Cuando esos modelos no han sido definidos, una solución consiste en comparar el modelo que se está probando contra un modelo saturado, es decir, un modelo en el que se estiman tantos parámetros como varianzas y covarianzas tiene la matriz, o contra un modelo de independencia, en el que todas las relaciones se asumen nulas (una estrategia empleada por el paquete AMOS).

Pero ¿qué hacer cuando la muestra no es suficientemente grande? Una solución sugerida por Browne y Cudeck (1989) consiste en emplear un estadístico de validación cruzada basado en una sola muestra (ECVI), que tiene en cuenta el número de parámetros implicados por el modelo y el número de sujetos de la muestra.

## 5. APLICACIÓN

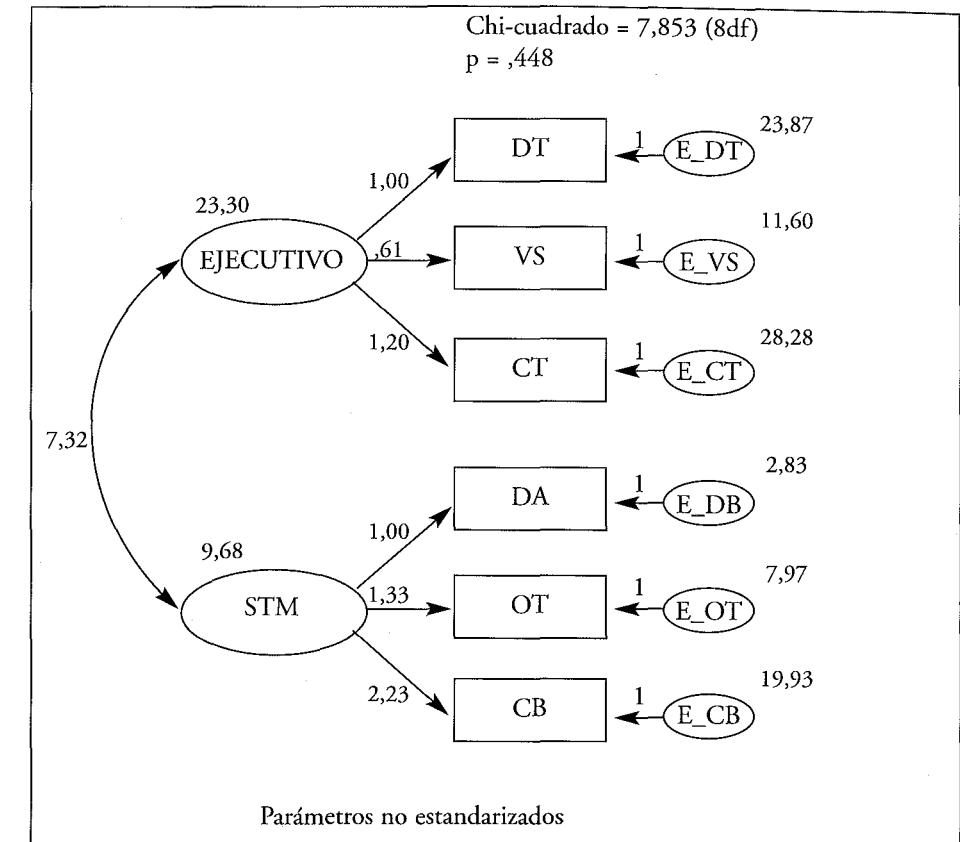
Ejemplificaremos el análisis mediante ecuaciones estructurales asumiendo que hemos recopilado datos de un grupo de 73 estudiantes de bachillerato a los que hemos pasado seis pruebas (DA, OT, CB, DT, VS y CT), que miden diferentes aspectos de la memoria de corta duración. Nuestro interés reside en saber cuántos factores son necesarios para explicar el patrón de correlaciones observadas entre esas variables. Por consiguiente realizamos en primer lugar un análisis factorial exploratorio, que parece indicar que una solución de dos factores es la más adecuada. Concretamente las tres primeras variables presentan una alta carga (rotación Oblimin directo) en el primer factor, y las tres últimas en el segundo. Decidimos pues, realizar un análisis factorial confirmatorio, asumiendo el modelo que aparece en la Figura 12.4. Nótese que se asume que los dos factores están correlacionados y que fijamos la escala de medida de los factores haciendo que uno de sus coeficientes estructurales sea la unidad. Ahora podemos realizar el ajuste del modelo mediante un paquete estadístico.

Figura 12.4. Análisis factorial confirmatorio. Modelo de dos factores



Uno de los paquetes estadísticos más sencillos de utilizar es AMOS, que puede adquirirse como paquete complementario de SPSS. El ajuste realizado por AMOS 4.0, aparece en la Figura 12.5. Nótese que según chi-cuadrado el ajuste es significativo, puesto que no puede rechazarse la hipótesis nula. Esta idea puede confirmarse examinando la salida completa (en texto) que proporciona AMOS, y que aparece en el Cuadro 12.1. (véase página 378). Obsérvese cómo los estadísticos de ajuste (por ejemplo, GFI, AGFI, etc.) están situados por encima de los puntos de corte recomendados. Además, prácticamente todas las variables latentes son buenas medidas de su factor correspondiente, puesto que los cuadrados de los coeficientes de correlación múltiple son claramente mayores que 0,50, excepto en el caso de VS, aunque se sitúa relativamente cerca (0,42). Por otra parte, las covarianzas residuales estandarizadas entre las variables son relativamente bajas, lo que sugiere otra vez que el modelo ha considerado de forma adecuada las posibles correlaciones entre factores. El lector puede comprobar que eliminando la correlación entre los dos factores, las covarianzas residuales entre las variables observadas de un factor y las del otro incrementan notablemente. Finalmente, los pesos estandarizados indican que la solución es aceptable.

Figura 12.5. Salida gráfica de AMOS 4.0 para el modelo de dos factores correlacionados



Cuadro 12.1. Modelo de dos factores correlacionados. AMOS 4.0

Chi-square = 7,853; Degrees of freedom = 8; Probability level = 0,448					
Regression Weights:					
DT <---- EJECUTIVO		Estimate	S.E.	C.R.	
VS <---- EJECUTIVO	1,000	0,143	4,250		
CT <---- EJECUTIVO	0,610	0,272	,405		
DA <---- STM	1,198				
OT <---- STM	1,000				
CB <---- STM	1,334	0,160	8,322		
CB <---- STM	2,234	0,263	8,482		
Standardized Regression Weights:					
DT <---- EJECUTIVO	0,703	VS <---- EJECUTIVO	0,654		
CT <---- EJECUTIVO	0,736	DA <---- STM	0,880		
OT <---- STM	0,827	CB <---- STM	0,841		
Covariances:					
EJECUTIVO <----> STM		Estimate	S.E.	C.R.	
		7,315	2,571	2,846	
Squared Multiple Correlations:					
CB	0,708	OT	0,684	DA	0,774
CT	0,542	VS	0,428	DT	0,494
Standardized Residual Covariances					
CB	OT	DA	CT	VS	DT
OT	-0,002	0,000			
DA	0,020	-0,020	0,000		
CT	0,546	0,317	0,085	0,000	
VS	-0,689	-0,695	-0,567	0,085	0,000
DT	-0,475	0,683	0,341	-0,194	0,180
					-0,000
Summary of models					
Model	NPAR	CMIN	DF	P	CMIN/DF
Your model	13	7,853	8	0,448	0,982
Independence model	6	187,718	15	0,000	12,515
Model	RMR	GFI	AGFI	PGFI	
Your model	1,677	0,966	0,910	0,368	
Independence model	13,807	0,496	0,294	0,354	
Model	AIC	BCC	BIC	CAIC	
Your model	33,853	36,653	86,922	76,629	
Independence model	199,718	201,010	224,211	219,461	

## 6. LIMITACIONES Y SUPUESTOS

SEM es, en general, relajado en cuanto a las exigencias que plantea a los datos para su aplicación. Sin embargo, es importante tener presente que sirve de manera más adecuada para contrastar modelos cuando carece de algunas limitaciones y se cumplen los pocos supuestos que requiere. Veámoslos a continuación.

### 6.1. NORMALIDAD MULTIVARIADA

Este supuesto se aplica tanto a los indicadores como a las variables latentes. En ambos casos requiere que la distribución de cualquier variable sea normal con respecto a cualquier valor del resto de las variables. Como sabemos esto implica que cualquier combinación lineal de las variables sea normal. Puesto que los tests de ajuste son, en general, variaciones de chi-cuadrado, el cumplimiento de este supuesto es importante, dado que incluso pequeñas violaciones de normalidad pueden traducirse en grandes cambios en chi-cuadrado. Existen índices alternativos cuando se sospecha no normalidad, como el índice de Bentler-Satorra, que se basa en chi-cuadrado, pero trata de corregir el sesgo que se produce por la no normalidad de los datos.

Otro aspecto importante del supuesto está relacionado con la estimación de los parámetros en la que suele emplearse de modo preponderante la máxima verosimilitud. Este procedimiento asume la normalidad. Estudios de simulación recientes (Kline, 1998) han indicado que la violación del supuesto causa problemas no en los valores de los parámetros (que no suelen variar incluso en casos extremos de no normalidad), sino en sus errores asociados, que suelen ser menores cuanto mayor desviación de la normalidad. Recordemos que para tomar una decisión sobre la significación de un parámetro (la existencia de una relación entre variables), se construye un cociente entre el parámetro y su error asociado, lo que llevará aparejada una alta tasa de error tipo I, esto es, que se encontrarán significativas más relaciones de las que realmente son. Como hemos indicado ya, una solución posible consiste en utilizar transformaciones que corrijan la no normalidad.

### 6.2. LINEALIDAD

En SEM se asume que las relaciones entre las variables son lineales. Esto se aplica tanto a variables latentes como observadas. El no cumplimiento del supuesto tiene consecuencias más graves cuando el procedimiento de estimación de parámetros es máxima verosimilitud, que asume linealidad, pero no cuando se emplean mínimos cuadrados, que no la asumen. La solución, como

hemos mencionado antes consiste en realizar transformaciones, por ejemplo, exponenciales o logarítmicas.

### 6.3. AUSENCIA DE MULTICOLINEALIDAD Y SINGULARIDAD

Hemos señalado ya en capítulos anteriores que cuando es necesario invertir matrices su determinante no puede ser nulo. Si la matriz es singular no puede ser invertida y por tanto, el proceso no puede completarse. Una de las razones por las cuales esto puede producirse es que una de las columnas de la matriz sea una combinación lineal de otras columnas, en otras palabras las variables sean multicolineales, su correlación sea prácticamente perfecta.

### 6.4. INDICADORES MÚLTIPLES

Una limitación importante de SEM es que requiere que haya más de un indicador para poder medir las variables latentes del modelo y estimar los errores. Cuando solamente hay dos, la correlación entre ambos puede ser empleada como un tercer indicador. Esta exigencia se debe a la necesidad de reducir los errores de medida e incrementar la fiabilidad de la medición. Un problema asociado con éste es la pérdida de datos que puede producirse cuando algunos sujetos no pasan algunas de las pruebas. Lógicamente la solución es incrementar el número de indicadores en el primer caso y tratar de tener casos completos en el análisis (lo que puede lograrse eliminando los que tienen casos perdidos, en cuyo caso debe utilizarse la eliminación por lista, para evitar que la matriz sea singular, o estimando los valores perdidos).

### 6.5. MODELOS INFRAIDENTIFICADOS Y JUSTAMENTE IDENTIFICADOS

Un modelo justamente identificado es aquel que tiene tantos parámetros libres como varianzas y covarianzas tiene la matriz observada. El modelo puede ser estimado, pero su ajuste no puede ser evaluado (los paquetes como AMOS reportan una chi-cuadrado de 0 con 0 grados de libertad, y la probabilidad no puede ser computada). Por otro lado, un modelo es infraidentificado cuando tiene más parámetros libres que varianzas y covarianzas en la matriz observada. El problema más importante en este caso es que no hay una solución única a los parámetros, y el ajuste no puede ser evaluado. Por tanto, los únicos modelos de valor son los sobreidentificados, es decir, los que tienen más varianzas-covarianzas que parámetros libres, o, en otras palabras, grados de libertad positivos.

La cantidad de soluciones que pueden adoptarse en el caso de que el modelo sea infra o justamente identificado es grande. He aquí algunas de las más importantes. En primer lugar, pueden suprimirse relaciones haciendo que sus

parámetros sean fijos (iguales a cero). Obligar a ciertos coeficientes a adoptar determinados valores (cuando se tienen razones suficientes para utilizar esos valores). Tratar de simplificar el modelo eliminando variables latentes o variables que sean colineales con otras. Añadir más variables exógenas. Definir modelos recursivos (en los que las flechas fluyen en una dirección y no se incluyen bucles ni *feed-back*). Utilizar procedimientos de estimación alternativos, como mínimos cuadrados no ponderados o generalizados. Si se usa máxima verosimilitud, incrementar el número de iteraciones para conseguir la convergencia, cambiar los estimados iniciales de los parámetros.

### 6.6. TAMAÑO DE LA MUESTRA

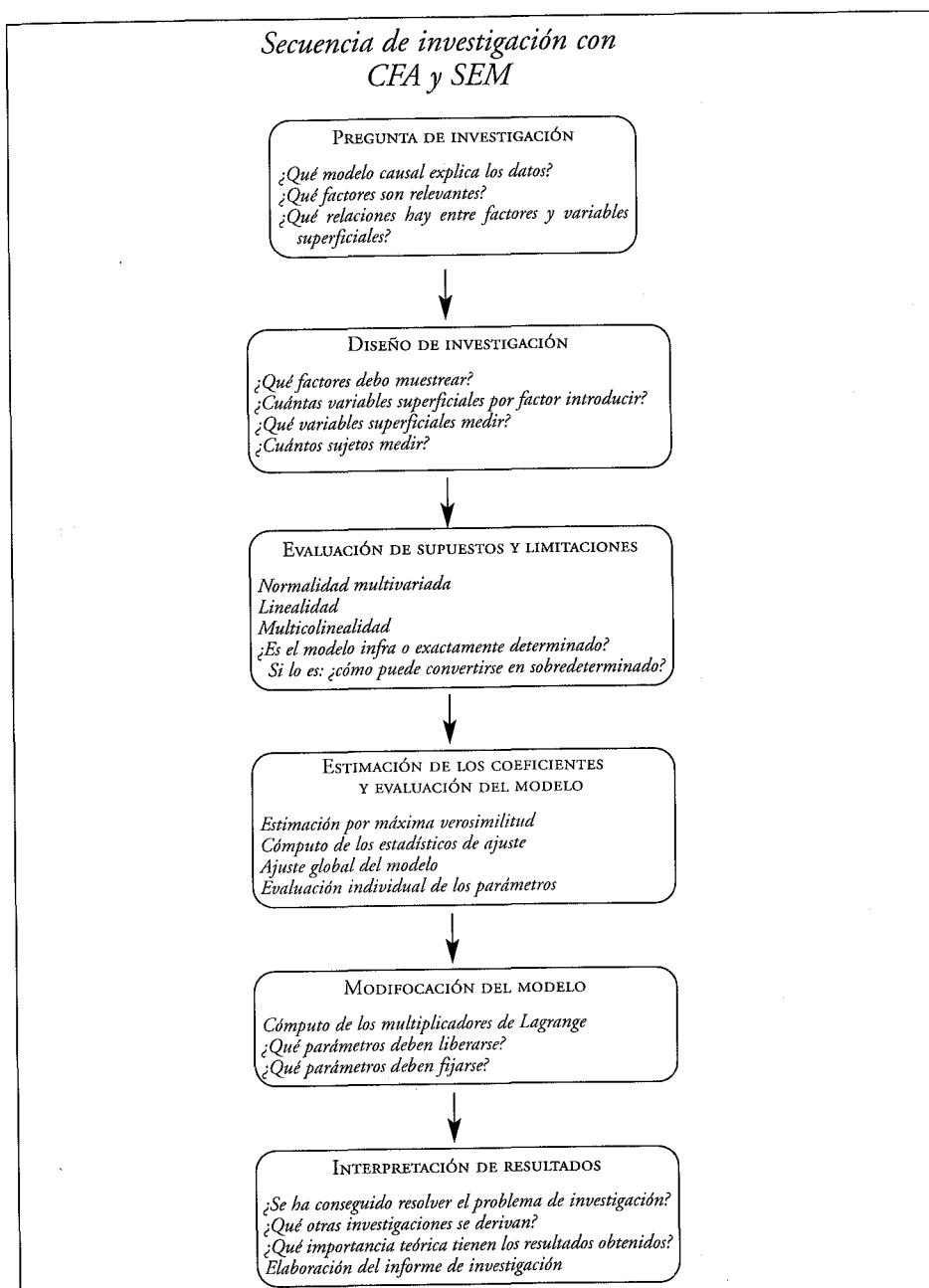
En muchos contextos de investigación éste es uno de los problemas más importantes de SEM, puesto que requiere un número bastante elevado de sujetos. Habitualmente se estima que tamaños muestrales por debajo de 200 son insuficientes para evaluar adecuadamente un modelo debido a la inestabilidad de los parámetros y a la falta de potencia de los tests de contraste. Una regla adecuada para decidir el tamaño de la muestra sería tener 8 veces más sujetos que variables (incluyendo las latentes) o bien 15 casos por variable superficial.

## 7. LA SECUENCIA DE INVESTIGACIÓN CON CFA Y SEM

El proceso de investigación orientado al análisis mediante CFA o SEM implica cinco etapas fundamentales: conceptualización y especificación del modelo, identificación, estimación de parámetros, evaluación y su posible modificación. El Cuadro 12.2. presenta el esquema de estas etapas que detallamos a continuación.

En primer lugar, debe tenerse en cuenta que CFA y SEM son técnicas orientadas a la teoría. Como hemos mencionado antes, si no hay teoría que contrastar deben emplearse técnicas alternativas como EFA. En el contexto de CFA y SEM una teoría implica un conjunto de variables y sus relaciones. Concretamente es necesario asumir que ciertas variables son causa de los cambios que se observan en otras variables, particularmente que ciertas variables latentes son las causas de ciertas variables observadas, pero también que unas variables latentes están ligadas causalmente con otras. La especificación del modelo requiere, por tanto, que se detallen las variables latentes, las variables observadas que se supone son afectadas por ellas, y las conexiones existentes entre todo el conjunto de variables. En otras palabras, la especificación del modelo implica que se defina un modelo estructural (de relaciones entre las variables latentes), y los modelos de medida necesarios (entre las variables observadas endógenas y las latentes exógenas y entre las observadas endógenas y las latentes endógenas). La identificación del modelo exige que los parámetros de los

Cuadro 12.2. Secuencia de investigación con CFA y SEM



modelos de medida y del modelo estructural estén definidos y tengan asignada una unidad de medida.

Una vez que el modelo está identificado la etapa siguiente consiste en la estimación de los parámetros que puede realizarse mediante diferentes procedimientos. El más comúnmente empleado es el máximo verosímil, aunque pueden utilizarse otros como mínimos cuadrados generalizados (GLS, ponderados) o mínimos cuadrados no ponderado (ULS). La solución obtenida debe ser evaluada en el sentido de determinar si el modelo es suficientemente consistente. En este sentido deben recordarse algunos puntos importantes que ayudan a determinar si el proceso ha funcionado de forma correcta, como la existencia de varianzas negativas, o parámetros cuyos valores estén fuera de rango. La evaluación del modelo requiere además que se computen estadísticos de ajuste global, que son indicadores de la semejanza entre la matriz de varianzas-covarianzas predicha por el modelo y la observada. Recuérdese, sin embargo, que varios de estos índices dependen del tamaño de la muestra y del supuesto de normalidad multivariada, y que no hay criterios objetivos para tomar la decisión de si el ajuste obtenido es lo suficientemente bueno. En este sentido es importante tener presente que una buena estrategia consiste en comparar nuestro modelo contra un modelo de referencia alternativo. Además es importante evaluar la significación de cada uno de los parámetros, lo que puede ayudar a comprender mejor cómo el modelo explica los resultados observados.

Cuando el ajuste se considera suficiente, puede considerarse acabado el proceso, a excepción de la interpretación de los resultados en el marco teórico adecuado. Sin embargo, cuando el ajuste no se considera satisfactorio, suele ser frecuente intentar modificarlo para tratar de mejorar su ajuste. La modificación, por lo general, implica la inclusión de nuevas relaciones y/o la exclusión de otras ya definidas en el modelo. Los índices de modificación proporcionan una ayuda valiosa en este sentido, puesto que permiten decidir cuáles son los parámetros que deben ser liberados y cuáles los que pueden ser fijados. Una buena estrategia en este sentido es proceder paso a paso, cambiando un solo parámetro cada vez, puesto que el resto de los parámetros pueden verse afectados por él. Es importante, además, tratar de validar los resultados. Como hemos visto, la mejor estrategia de validación consiste en utilizar muestras distintas, una para la estimación y otra para la validación. Cuando eso no es posible, cabe emplear, como mínimo, el índice de validación cruzada de una sola muestra de Browne y Cudeck.

## 8. LECTURAS RECOMENDADAS

BOLLEN, K. A., *Structural equations with latent variables*, Nueva York, NY, John Wiley & Sons, 1989.

HOYLE, R. H., *Structural equation modeling: Concepts, Issues, and Applications*, Thousand Oaks, CA, Sage, 1995.

SARIS, W. E., «Structural Equation Modelling», en H. Adèr y G. J. Mellenbergh (eds.), *Research methodology in the social, behavioural and life sciences*, Londres, Sage, 1999, págs. 220-239.

#### LECTURAS COMPLEMENTARIAS

- Hox, J. J. y BECHGER, T. M., «An introduction to structural equation modeling», *Family Sciences Review*, 11, 1998, págs. 354-373.  
 KLINE, R. B., *Principles and practice of structural equation modeling*, NY, Guilford Press, 1998.  
 LOEHLIN, J. C., *Latent Variable Models: an introduction to factor, path, and structural analysis*, Mahwah, N. J., Lawrence Erlbaum Associates, 1998.

#### 9. EJERCICIOS

- 1) ¿Qué diferencias existen entre EFA y CFA?
- 2) ¿Qué diferencias hay entre CFA y SEM?
- 3) ¿Qué es el modelo de medidas?
- 4) ¿Cuántos modelos de medidas tiene CFA?
- 5) ¿Cuántos modelos de medida tiene SEM?
- 6) ¿Qué es el modelo estructural?
- 7) Clasifique los siguientes modelos como infra-, justo- o sobre- identificados, indique los grados de libertad de cada modelo y dibuje el diagrama de paso:

a)

$$\begin{aligned} X_1 &= \lambda_1 \xi_1 + \delta_1 \\ X_2 &= \lambda_2 \xi_1 + \delta_2 \end{aligned}$$

b)

$$\begin{aligned} X_1 &= \lambda_1 \xi_1 + \delta_1 \\ X_2 &= \lambda_2 \xi_1 + \delta_2 \\ X_3 &= \lambda_3 \xi_1 + \delta_3 \\ X_4 &= \lambda_4 \xi_1 + \delta_4 \\ X_5 &= \lambda_5 \xi_1 + \delta_5 \end{aligned}$$

#### RESPUESTAS A LOS EJERCICIOS

1. EFA es exploratorio en el sentido de que se desconocen *a priori* tanto los posibles factores como el peso de cada variable observada en los mismos. CFA requiere un modelo en el que los factores estén bien especificados, así como las variables que tienen asociadas.
2. CFA es un caso particular de SEM.

3. El modelo de medidas hace referencia a la relación entre variables superficiales y factores exógenos. Las coeficientes están etiquetados como  $\lambda$ .
4. En CFA hay un único modelo de medidas.
5. En SEM hay dos, uno entre variables observadas y factores exógenos, y otro entre variables observadas y endógenos.
6. El modelo estructural hace referencia a la relación entre variables latentes.
7. a) Infraidentificado.  
 b) Superidentificado, si los factores no están correlacionados.

## BIBLIOGRAFÍA

- AGRESTI, A., *An Introduction to Categorical Data Analysis*, Nueva York, John Wiley, 1996.
- ALLISON, P. D., *Multiple Regression: A Primer*, Pine Forge Press, 1999.
- AMÓN, J., *Introducción al análisis multivariante (Cálculo matricial)*, Barcelona, PPU, 1991.
- ARNAU I GRAS, J., *Diseños experimentales multivariados*, Madrid, Alianza, 1990.
- ARNAU, J., *Métodos y técnicas avanzadas de análisis de datos en ciencias del comportamiento*, Barcelona, UB, 1996.
- ATO, M., *Investigación en ciencias del comportamiento. I: Fundamentos*, Barcelona, PPU, 1991.
- ATO, M. y LÓPEZ, J. J., *Ánalisis estadístico para datos categóricos*, Madrid, Síntesis, 1996.
- BIRD, K. D. y HADZI-PAVLOVIC, D., «Simultaneous test procedures and the choice of a test statistics in MANOVA», *Psychological Bulletin*, 93, 1983, págs. 167-178.
- BISQUERRA, R., *Introducción conceptual al análisis multivariante*, Barcelona, PPU, 1989.
- BISHOP, T. A., *Heteroscedastic ANOVA, MANOVA, and multiple comparisons*, Phd Thesis, Ohio State Univ. Columbus, no publicada, 1976.
- BLASHFIELD, R. K. y ALDENDERFER, M. S., *Cluster analysis*, Sage University Papers. Quantitative Applications in the Social Sciences, Beverly Hills, CA, Sage, 1987.
- BOCK, R., *Multivariate statistical methods in behavioral research*, Nueva York, McGraw Hill, 1975.
- BOEHNKE, K., «F- and H-test assumptions revisited», *Educational and Psychological Measurement*, 44, 1984, págs. 609-617.
- BOLLEN, K. A., *Structural equations with latent variables*, Nueva York, John Wiley & Sons, 1989.
- BRAY, J. H. y MAXWELL, S. E., *Multivariate analysis of variance*, Beverly Hills, CA, Sage, 1985.
- BRECKENRIDGE, J. N., «Replicating cluster analysis: Method, consistency, and validity», *Multivariate Behavioral Research*, 24, 1989, 147-161.
- BRYK, A. S. y RAUDENBUSH, S. W., «Heterogeneity of variance in experimental studies: A challenge to conventional interpretations», *Psychological Bulletin*, 104, 1988, págs. 396-404.
- CAMACHO, J., *Estadística con SPSS para Windows*, Madrid, Ra-Ma, 2000.

- CATENA, A. y VALDÉS, B., *El ejecutivo central: estructura y funciones en atención, memoria y aprendizaje*, Madrid, no publicado, 2002.
- CATTELL, R. B., *The scientific use of factor analysis in behavioural and life sciences*, Nueva York, Plenum, 1978.
- CHRISTENSEN, R., *Log-Linear Models and Logistic Regression*, N. Y., Springer-Verlag, 1997.
- COHEN, J., *Statistical power analysis for the behavioral sciences*, Nueva York, Academic Press, 1988.
- COMREY, A. L., «The minimum residual method of factor analysis», *Psychological Reports*, 11, 1962, págs. 15-18.
- COMREY, A. L. y LEE, H. B., *A first course in factor analysis*, Hillsdale, NJ, LEA, 1992.
- COOLEY, W. W. y LOHNES, P. R., *Multivariate data analysis*, Nueva York, Wiley, 1971.
- COX, D. R., *Planning of experiments*, Nueva York, Wiley, 1958.
- CRAMER, E. M. y NICEWANDER, W. A., «Some symmetric, invariant measures of multivariate association», *Psychometrika*, 44, 1979, págs. 43-54.
- DAVIDSON, M. L., «Univariate versus multivariate tests in repeated measures experiments», *Psychological Bulletin*, 77, 1972, págs. 446-452.
- DILLON, W. R. y GOLDSTEIN, M., *Multivariate Analysis: Methods and applications*, Nueva York, Wiley, 1984.
- DRAPER, N. R. y SMITH, H., *Applied Regression Analysis*, Nueva York, Wiley, 1998.
- DUNCAN, J. y cols., «A neural basis for general intelligence», *Science*, 289, 2000, páginas 457-460.
- EDWARDS, A. L., *An introduction to lineal regression and correlation*, San Francisco, Freeman, 1976.
- ENGLE, R. W. y cols., «Working memory, short-term memory, and general fluid intelligence: A latent-variable approach», *Journal of Experimental Psychology: General*, 128, 1999, págs. 309-311.
- ESTARELLES, R.; DE LA FUENTE, E. I. y OLMEDO, P., «Aplicación y valoración de diferentes algoritmos no jerárquicos en el análisis cluster y su representación gráfica», *Anuario de Psicología*, 55, 1992, 63-90.
- EVERITT, B. S., *Cluster analysis*, Londres, Edward Arnold, 1993.
- FOX, J. D., *Linear Statistical Models and Related Methods*, Nueva York, Wiley, 1984.
- FOX, J., *Applied Regression Analysis, Linear Models, and Related Methods*, CA, Sage, 1997.
- GABRIEL, K. R., «Simultaneous test procedures in multivariate analysis of variance», *Biometrika*, 55, 1968, págs. 389-504.
- GEISER, S. y GREENHOUSE, S. W., «An extension of Box's results on the use of the F distribution in multivariate analysis», *Annals of Mathematical Statistics*, 29, 1958, págs. 885-891.
- GIL, J.; GARCÍA, E. y RODRÍGUEZ, G., *Análisis discriminante*, Madrid, La Muralla, 2001.
- GREEN, S. B., «How many subjects does it take to do a regression analysis?», *Multivariate Behavioral Research*, 27, 1991, págs. 346-354.
- GREENHOUSE, S. W. y GEISSER, S., «On methods in the analysis of profile data», *Psychometrika*, 24, 1959, págs. 95-102.
- GRIMM, L. G. y YARNOLD, P. R., *Reading and Understanding Multivariate Statistics*, Washington: APA, 1995.
- GUILLÉN, M. F., *Análisis de regresión múltiple*, Madrid, Centro de Estudios Políticos y Constitucionales, 1992.
- HAASE, R. F.; ELLIS, M. V. y LADANY, N., «Multiple criteria for evaluating the magnitude of experimental effects», *Journal of Counseling Psychology*, 36, 1989, págs. 511-516.

- HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L. y BLACK, W. C., *Análisis multivariante*, Madrid, Prentice-Hall, 1998-1999.
- HAND, D. J. y TAYLOR, C. C., *Multivariate analysis of variance and repeated measures*, Londres, Chapman and Hall, 1987.
- HARMAN, H. H., *Modern factor analysis*, Chicago, IL, The University of Chicago Press, 1976.
- HENDRICKSON, A. E. y WHITE, P. O., «Promax: A quick method for rotation to oblique simple structure», *British Journal of Statistical Psychology*, 17, 1964, páginas 65-70.
- HERNÁNDEZ, M., *Técnicas de clasificación: medidas de proximidad, métodos de agrupamiento y validación de una solución Cluster*, La Laguna, Universidad de La Laguna, 1998.
- HOSMER, D. W. y LEMESHOW, S., *Applied Logistic Regression*, Nueva York, Wiley, 1989.
- HOTELING, H., «Analysis of a complex of statistical variables into principal components», *Journal of Educational Psychology*, 24, 1933, págs. 417-441, 498-520.
- HOX, J. J. y BECHGER, T. M., «An introduction to structural equation modeling», *Family Sciences Review*, 11, 1998, págs. 354-373.
- HOYLE, R. H., *Structural equation modeling: Concepts, Issues, and Applications*, Thousand Oaks, CA, Sage, 1995.
- HOSMER, D. W. y LEMESHOW, S., *Applied Survival Analysis: Regression Modeling of Time to Event Data*, Nueva York, Wiley, 1999.
- HOWELL, D. C., *Statistical Methods for Psychology*, Boston, MA, Duxbury Press, 1989.
- HUANG, Z., «A Fast Clustering Algorithm to Cluster Very Large Categorical Data sets in Data Mining», en *Proceedings of SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, 1997.
- HUBERTY, C. J., «Issues in the use and interpretation of discriminant analysis», *Psychological Bulletin*, 95, 1984, págs. 156-171.
- *Applied Discriminant Analysis*, Nueva York, Wiley, 1994.
- HUBERTY, C. J. y MORRIS, J. D., «Multivariate versus multiple univariate analyses», *Psychological Bulletin*, 105, 1989, págs. 302-308.
- HUBERTY, C. J. y WIENBAKER, J. M., «Variable importance in Multivariate group comparisons», *Journal of Educational Statistics*, 17, 1992, págs. 75-91.
- JENNICH, R. I. y SAMPSON, P. F., «Rotation for simple loading», *Psychometrika*, 31, 1966, págs. 313-323.
- JOBSON, J. D., *Applied Multivariate Data Analysis, Vol. 2.: Categorical and Multivariate Methods*, Nueva York, Springer-Verlag, 1991.
- JOVELL, A. J., *Ánalisis de regresión logística*, Madrid, Centro de Estudios Políticos y Constitucionales, 1995.
- JUDD, Ch. M.; McCLELLAND, G. H. y CULHANE, S. E., «Data analysis: Continuing issues in the everyday analysis of Psychological data», *Annual reviews of Psychology*, 46, 1995, págs. 433-465.
- JOHNSON, R. A. y WICHERN, D. W., *Applied Multivariate Statistical Analysis*, Prentice Hall, 1998.
- KACHIGAN, S., *Multivariate Statistical Analysis: A Conceptual Introduction*, Nueva York, Radius Press, 1991.
- KAISER, H. F., «The Varimax criterion for analytic rotation in factor analysis», *Psychometrika*, 23, 1958, págs. 187-200.
- KAISER, H. F. y COFFREY, J., «Alpha factor analysis», *Psychometrika*, 30, 1965, págs. 1-14.
- KAPLAN, D., «Evaluation and modification of covariance structure models: A review and recommendation», *Multivariate Behavioral Research*, 25, 1990, págs. 137-155.

- KAUFMAN, L. y ROUSSEEUW, P. J., *Finding Groups in Data: An Introduction to Cluster Analysis*, Nueva York, Wiley, 1990.
- KLECKA, W. R., *Discriminant analysis*, Beverly Hills, CA, Sage, 1982.
- KLINE, R. B., *Principles and practice of structural equation modeling*, Nueva York, Guilford Press, 1998.
- KLEIN, P. J. y MOECHBERGER, M. L., *Survival Analysis*, Nueva York, Springer, 1997.
- KLEINBAUM, D. G., *Logistic regression: a self-learning text*, Nueva York, Springer-Verlag, 1994.
- *Survival Analysis: A Self-Learning Text (Statistics in the Health Sciences)*, Nueva York, Springer-Verlag, 1996.
- LANDAU, S.; LEESE, M. y EVERITT, B. S., *Cluster analysis*, Londres, Arnold, 2001.
- LE, C. T., *Applied Categorical Data Analysis*, Nueva York, Wiley, 1998.
- LOEHLIN, J. C., *Latent Variable Models: an introduction to factor, path, and structural analysis*, Mahwah, N. J., Lawrence Erlbaum Associates, 1998.
- LLOYD, C. L., *Statistical analysis of categorical data*. N. Y., Wiley, 1999.
- MARSH, H. H.; BALLA, J. R. y McDONALD, R. P., «Goodness-of-fit indexes in confirmatory factor analysis: The effect of simple size», *Psychological Bulletin*, 103, 1988, págs. 391-410.
- NESSELROADE, J. R. y CATTELL, R. B., *Handbook of multivariate experimental Psychology*, Nueva York, Plenum Press, 1988.
- O'BRIEN, R. G. y KAISER, M. K., «MANOVA method for analyzing Repeated Measures Design: An extensive primer», *Psychological Bulletin*, 97, 1985, páginas 316-333.
- OLSON, C. L., «On choosing a test statistic in multivariate analyses of variance», *Psychological Bulletin*, 83, 1976, págs. 579-586.
- «Practical considerations in choosing a MANOVA test statistics: A rejoinder to Stevens», *Psychological Bulletin*, 86, 1979, págs. 1.350-1.352.
- PARMAR, M. K. B. y MACHIN, D., *Survival analysis: a practical approach*, Chichester, Wiley, 1995.
- PASCUAL, J.; FRÍAS, D. y GARCÍA, F., *Manual de Psicología Experimental*, Barcelona, Ariel, 1996.
- PEDHAZUR, E. J., *Multiple Regression in Behavioral Research*, Wadsworth, 1997.
- RASMUSSEN, J. L., «An evaluation of parametric and nonparametric test on modified and non-modified data», *British Journal of Mathematical and Statistical Psychology*, 39, 1986, págs. 213-220.
- RASMUSSEN, J. L. y DUNLAP, W. P., «Dealing with nonnormal data: Parametric analysis of transformed data vs nonparametric analysis», *Educational and Psychological Measurement*, 51, 1991, págs. 509-520.
- «Dealing with nonnormal data: Parametric analysis of transformed data vs nonparametric analysis», *Educational and Psychological Measurement*, 51, 1991, págs. 32-43.
- RIBA, M.ª D., *Modelo lineal de análisis de la variancia*, Barcelona, Herder, 1990.
- RIVAS, M.ª J. y LÓPEZ, J., *Análisis de supervivencia*, Madrid, La Muralla, 2000.
- RIVAS, M.ª T.; MARTÍNEZ-ARIAS, M.ª R. y RIUS, F., *Análisis discriminante: una aplicación del método «stepwise»*, Málaga, Universidad de Málaga, Departamento de Didáctica y Organización Escolar, 1990.
- RUTHERFORD, A., *Introducing Anova and Ancova*, Londres, Sage, 2000.
- SÁNCHEZ, E., *Regresión logística en salud pública*, Granada, Escuela Andaluza de Salud Pública, 2000.

- SÁNCHEZ-CARRIÓN, J. J., *Introducción a las técnicas de análisis multivariable aplicadas a las ciencias sociales*, Madrid, CIS, 1984.
- SARIS, W. E., «Structural Equation Modelling», en H. Adèr y G. J. Mellenbergh (eds.), *Research methodology in the social, behavioural and life sciences*, Londres, Sage, 1999, págs. 220-239.
- SEDLMEIER, P. y GIGERENZER, G., «Do studies of statistical power have an effect on the power of studies?», *Psychological Bulletin*, 105, 1989, 309-316.
- SERLIN, R. C., «A multivariate measure of association based on the Pillai-Bartlett procedure», *Psychological Bulletin*, 91, 1982, págs. 413-417.
- SHARMA, S., *Applied multivariate techniques*, Nueva York, Wiley, 1996.
- STEVENS, J. P., «Step-Down Analysis and Simultaneous Confidence Intervals in MANOVA», *Multivariate Behavioral Research*, 8 (3), 1973, págs. 391-402.
- «Power of the multivariate analysis of variance test», *Psychological Bulletin*, 88, 1980, págs. 728-737.
- *Applied multivariate statistics for the social sciences*, Hillsdale, N. J., LEA, 1986.
- *Applied multivariate statistics for the social sciences*, Hillsdale, N. J., LEA, 1992.
- *Applied multivariate statistics for the social sciences*, 3.<sup>a</sup> ed., Mahwah, N. J., LEA, 1996.
- STEWART, D. K. y LOVE, W. A., «A general canonical correlation index», *Psychological Bulletin*, 70, 1968, págs. 160-163.
- TABACHICK, B. G. y FIDELL, L. S., *Using multivariate statistics*, Boston, Allyn y Bacon, 2001.
- TOMARKEN, A. J. y SERLIN, R. C., «Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures», *Psychological Bulletin*, 99, 1986, págs. 90-99.
- THOMPSON, B., *Canonical correlation analysis: uses and interpretation*, Newbury Park, Sage, 1989.
- THURSTONE, L. L., *Multiple Factor Analysis*, Chicago, University of Chicago Press, 1947.
- WALSH, A., «Teaching understanding and interpretation of logit regression», *Teaching Sociology*, 15, 1987, págs. 178-183.
- VALLEJO, G., *Ánalisis univariado y multivariado de los diseños de medidas repetidas de una sola muestra y de muestras divididas*, Barcelona, PPU, 1991.
- WICKENS, T. D., *Multiway contingency tables analysis for the social sciences*, Hillsdale, N. J., LEA, 1989.
- «Categorical data analysis», *Annual Review of Psychology*, 49, 1998, págs. 537-557.
- WILCOX, R. R., «New design in analysis of variance», *Annual Review of Psychology*, 38, 1987, págs. 29-60.
- *Introduction to robust estimation and hypothesis testing*, Nueva York, Academic Press, 1997.
- WOODWARD, J. A. y OVERALL, J. E. «Multivariate analysis of variance by multiple regression methods», *Psychological Bulletin*, 82, 1975, págs. 21-32.
- YOUNG, F. W., «Quantitative analysis of qualitative data», *Psicometrika*, 46, 1981, págs. 357-388.

## GLOSARIO

**Ajuste del modelo.** El ajuste en la predicción del modelo se evalúa mediante  $R^2$  y  $R^2$  ajustada, que tiene en cuenta el número de sujetos y el número de predictores.

**Cambio en R cuadrado.** Cuando se introduce o se elimina una variable del modelo su ajuste puede cambiar. Si el cambio es grande, la variable introducida o eliminada tiene asociada una  $R^2$  grande, es importante para hacer la predicción.

**Cluster.** Un *cluster* o conglomerado es un grupo de individuos o variables semejantes entre sí. El análisis de *cluster* trata de agrupar a los individuos (o variables) de forma que las semejanzas dentro de cada *cluster* sean lo más grandes posible y cada cluster difiera de los demás lo más posible.

**Centroide.** El conjunto ordenado (vector) de medias de grupo, cada una correspondiente a una variable medida. También se les denomina centros.

**Dendrograma.** Representación visual del proceso de agrupamiento de sujetos (o variables) en *clusters*. Los sujetos (o variables) agrupados se unen median-

te líneas conectadas entre sí. Especifica también la distancia entre los casos que se agrupan. En SPSS está distancia es reescalada a un máximo de 25 y un mínimo de 0.

**Estandarizar.** Algunas medidas de distancia son sensibles a la escala de medida de las variables. Una forma de eliminar la influencia de la escala consiste en estandarizarlas. SPSS ofrece varios métodos de estandarización, según se pretenda conseguir que la escala tenga una media determinada, una desviación típica determinada, un rango determinado, o una media 0 y una desviación típica 1.

**Iterar y clasificar.** El análisis de *cluster* por métodos de partición (K-Medias) es recomendable en dos etapas. En la primera los sujetos son asignados y reasignados a los *clusters*, cuyos centros se van actualizando. Es conveniente guardar los centros de la solución final en un archivo.

**Matriz de distancias.** Matriz cuadrada que contiene las distancias entre todos los pares de sujetos (o variables) posibles.

**Medida de distancia.** La forma en que se

va a medir la semejanza entre los individuos o las variables. Hay un gran número, dependiendo de si las variables medidas son métricas, frecuencias o dicotómicas (véase Capítulo V) para una descripción de las mismas.

**Método de conglomeración.** Algoritmo que se va a emplear para agrupar a los individuos (o las variables) en el análisis de *cluster* jerárquico de aglomeración. Hay una gran variedad de métodos (véase Capítulo IV) para una descripción de cada uno de ellos.

**Sólo clasificar.** Los centros obtenidos en la solución final del procedimiento Iterar y Clasificar son empleados como centros iniciales en la segunda etapa en la que los centros de los conglomerados no se actualizan a medida que los sujetos son asignados y desasignados de los *clusters*.

**Solución única y Rango de soluciones.** En el análisis de *cluster* jerárquico de aglomeración es posible especificar el número de *clusters* que se desea obtener (solución única) o el rango de números de *clusters* (rango de soluciones). Esta última permite explorar mejor cuál puede ser el número de grupos más adecuado.

**Témpanos.** Es equivalente al dendrograma en cuanto a la información que ofrece. Se trata de figuras que semejan carámbanos colgando, en las que los casos que se unen horizontalmente son los *clusters*.

**Transformar medidas.** Las medidas de semejanza pueden transformarse para ser medidas de diferencia (o desemejanza), para considerar sólo los valores absolutos de semejanza, o para que se ajusten a un rango máximo (entre 0 y 1), lo que las hará más fácilmente interpretables.

**Colinealidad.** Dos o más variables explicativas tienen una fuerte correlación entre ellas, lo que dificulta o incluso

hace imposible determinar su valor individual para explicar la variable predicha. En el caso extremo, la matriz de correlaciones es singular, y no pueden estimarse los parámetros.

**Contrastes o comparaciones.** El análisis global (ANOVA, ANCOVA, MANOVA, MANCOVA) permite tomar decisiones sobre hipótesis nulas globales. Cuando la(s) variable(s) independiente(s) están manipuladas a más de dos niveles es preciso compararlos para concretar el efecto de las manipulaciones. Los contrastes son comparaciones entre niveles (o mezclas de niveles) de las variables. Matemáticamente son sumas ponderadas de medias, en las que la suma de los coeficientes es cero. Hay dos tipos diferentes: planeados o *a priori* y no planeados o *a posteriori*.

**Contrastes no planeados:**

**Bonferroni.** Una prueba que controla el crecimiento del error tipo I a lo largo del experimento fijando el error alfa en cada contraste según la designidad de Bonferroni:  $\alpha_{\text{del contraste}} = \alpha_a \text{ lo largo del experimento} / c$ , siendo c el número de comparaciones. Por ejemplo, si se desea mantener la tasa de error a lo largo del experimento por debajo de 0,05 y se van a realizar 3 comparaciones, el alfa al que habría que tomar cada decisión sería  $0,05/3 = 0,017$ . Esta prueba puede usarse en comparaciones *a priori*, utilizando como valor de c el número de comparaciones realmente realizadas.

**DMS.** Diferencia mínima significativa. Una prueba basada en F que no controla el crecimiento del error tipo I. Debe emplearse sólo si el número de comparaciones que debe realizarse es muy bajo y desea tenerse una gran potencia en el contraste.

**Duncan.** En esta prueba de rango múltiple el error alfa por comparación

## Glosario

( $\alpha_r$ ) depende del rango de la comparación:  $\alpha_r = 1 - (1 - \alpha_E)^{r-1}$ .

**Dunnett.** Se emplea para comparar un conjunto de grupos experimentales respecto de un grupo de control.

**Gabriel.** Se basa también en el módulo máximo estudiantizado (M). Su valor crítico es:

$$DC = MC_{\text{error}} \frac{1}{2n_i} + \frac{1}{2n_j} M_{\alpha_B^{cf}},$$

siendo  $n_i$  y  $n_j$  el número de sujetos de las condiciones que se comparan.

**GT2 de Hochberg.** Esta prueba emplea el módulo máximo estudiantizado, definido como el cociente entre la media mayor en valores absolutos (M) y la  $MC_{\text{error}}$ . El valor crítico es:

$$DC = 2M_{\alpha_B^{cf}}$$

donde c es el número total de comparaciones entre pares de medias y f son los grados de libertad del error.

**Pruebas para varianzas no homogéneas y/o número desigual de sujetos.** Cuando el supuesto de homogeneidad de varianzas error no se cumple y/o cuando los diseños son no balanceados, las pruebas anteriores o no son aplicables o producen resultados no totalmente correctos. Las pruebas que pueden usarse en esas situaciones son la T2 de Tamhane, la T3 de Dunnett, la de Games-Howell y la C de Dunnett. Todas ellas estiman los grados de libertad del error de la siguiente forma:

$$v = \frac{(s_i^2 / n_i + s_j^2 / n_j)^2}{s_i^4 / n_i f_i + s_j^4 / n_j f_j}$$

donde i y j representan las dos condiciones que se comparan, los n son los sujetos de cada condición y los f

son los grados de libertad de cada una de ellas (su número de sujetos menos 1). La varianza del contraste se computa como

$$v = \frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}.$$

Los contrastes difieren en la estimación del rango.

**Ryan-Einot-Gabriel-Welsh.** Son dos pruebas, una basada en F y otra basada en el estadístico q. En ambos casos, se define el error  $g_r$  en función del valor de r, el rango de la comparación, definido como el número de medias entre las dos que se comparan una vez que los datos se han ordenado en forma creciente. Concretamente, si el rango es igual o mayor que los grados de libertad del efecto, se utiliza el error alfa a lo largo del experimento ( $\alpha_E$ ) como error alfa por comparación. En el resto de casos se emplea como error alfa por comparación el valor  $1 - (1 - \alpha_E)^{r/a}$ .

**Scheffe.** Es una prueba poco potente, basada en la distribución F. El grado de corrección del crecimiento del error alfa es una función directa del número de niveles del factor.

**Sidak.** Es similar a la prueba de Bonferroni, pero se emplea un valor más exacto para fijar el riesgo por comparación. Concretamente,  $\alpha_{\text{del contraste}} = 1 - (1 - \alpha_a \text{ lo largo del experimento})^{1/c}$ .

**Student-Newman-Keuls.** Es una prueba basada en el estadístico q. El error alfa por comparación depende de su rango (número de medias entre las dos que se comparan incluyendo ambas, una vez que se han ordenado de forma creciente). La diferencia crítica se obtiene como:

$$DC = q(\alpha, r, f) \sqrt{MC_{\text{error}} / n},$$

donde  $r$  es el rango de la comparación.

**Tukey.** Es una prueba basada en el estadístico  $q$ . El error alfa por comparación depende del número de niveles del factor. El valor crítico se calcula como:

$$DC = q(\alpha, a, f) \sqrt{MC_{\text{error}}/n}$$

**Tukey-b.** Es una mezcla de las pruebas de Tukey y Newman-Keuls. El valor crítico es el promedio de ambos valores críticos.

**Waller-Duncan.** Una prueba de contrastes múltiples con una aproximación bayesiana. La diferencia crítica se obtiene como:

$$DC = t_B(w, F, q, f) \sqrt{2MC_{\text{error}}/n}$$

donde  $t_B$  es el valor bayesiano que depende de la proporción entre errores tipo I y tipo II que se deseé cometer ( $w$ ), el cociente entre  $MC_{\text{efecto}}$  y  $MC_{\text{error}}$  ( $F$ ), y de los grados de libertad del numerador ( $q=a-1$ ) y del denominador ( $f=a(n-1)$ ) de  $F$ , siendo  $n$  el número de sujetos por nivel, y  $a$  el número de niveles del factor.

**Contrastes no planeados (post-hoc).** Son comparaciones realizadas sin que el investigador haya especificado hipótesis previas sobre diferencias de medias. El problema más importante reside en que se realizan un gran número de comparaciones, lo que obliga a utilizar procedimientos especiales para impedir el crecimiento del error alfa dentro del experimento. Los procedimientos más comunes son DMS, Bonferroni y Tukey. Todos ellos asumen homogeneidad de varianzas error. Todas las pruebas proceden del mismo modo: se computa una diferencia crítica entre medias, si la diferencia observada entre dos nive-

les de la variable es igual o supera la diferencia crítica, se rechaza la hipótesis nula (las medias son iguales).

**Contrastes planeados.** Son comparaciones en las que tratan de contrastarse hipótesis formuladas antes de realizar el experimento. Los coeficientes de los contrastes planeados son establecidos en función de las hipótesis. Suele emplearse la  $t$  de Student o la  $F$  de Snedecor. Ambos estadísticos producen los mismos resultados puesto que en este contexto  $F=t^2$ .

**Contraste polinómico.** También conocido como análisis de tendencias. Es un tipo especial de contraste planeado en el que se trata de dilucidar la relación funcional entre variable independiente y variable dependiente. Los componentes de tendencia psicológicamente más interpretables son los de bajo orden, normalmente lineal, cuadrático y cúbico.

**Correlaciones parcial y semiparcial.** La correlación parcial mide la correlación entre dos variables cuando se ha eliminado la influencia que sobre cada una de ellas tiene una tercera variable. Se calcula como:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2} \sqrt{1-r_{23}^2}}$$

La correlación semiparcial es la correlación existente entre dos variables cuando de una de ellas (pero no de la otra) se ha eliminado la influencia que tiene una tercera variable. Se calcula como:

$$r_{1(2.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{23}^2}}$$

**Covariable.** En el Modelo Lineal General se refiere a variables que están relacionadas con las variables dependientes,

pero que no son afectadas por las variables independientes. Se usan para realizar el control estadístico de la variabilidad error. También se les llama covariados o covariantes.

**Discriminante.** Técnica dirigida a determinar cuáles de las variables medidas a varios grupos de individuos ayudan a discriminar mejor entre ellos. Se usa en dos sentidos, para crear una nueva variable (discriminante) en la cual se localicen las diferencias entre los grupos, y para crear una función o regla de clasificación aplicable también a individuos cuyo grupo de pertenencia no es conocido *a priori*.

**Distancia de Mahalanobis.** Mide la distancia de un sujeto al centroide de su grupo en unidades típicas. Se emplea en varios contextos de análisis. En el análisis de regresión múltiple es la distancia entre las puntuaciones de un sujeto en las variables predictoras y el promedio de todos los sujetos. Esta medida es independiente de la escala en que se hayan medido las variables. Sirve para identificar sujetos extremos, que serán los que tengan grandes distancias.

**Clasificar dejando uno fuera.** Método para realizar un tipo de validación cruzada de la solución discriminante. Se calculan las funciones dejando, uno a uno, a todos los sujetos fuera de los cálculos, lo que evita el sesgo que se produce cuando un sujeto es tenido en cuenta para obtener las funciones que luego le serán aplicadas para clasificarlo. La discrepancia entre este método de clasificación y el estándar es un indicador de la validez de la solución alcanzada.

**Coeficientes de la función.** Coeficientes de las funciones discriminantes (no tipificados) y de clasificación (de Fisher).

**M de Box.** Comprueba que las matrices de covarianzas no son diferentes de un grupo a otro. La homogeneidad a través de grupos de estas matrices es un supuesto del análisis discriminante, que, sin embargo, es robusto al no cumplimiento. Además debe tenerse en cuenta que la prueba de Box es sensible al no cumplimiento de la normalidad multivariada.

**Probabilidades previas.** Las probabilidades *a priori* asociadas a cada grupo. Pueden estimarse a partir del número de sujetos de cada grupo. Los grupos con más sujetos tienen una probabilidad previa mayor que los grupos con menos sujetos.

**Independientes.** Son las variables que pueden servir para discriminar a unos grupos de otros y/o para clasificar a los individuos en los grupos. En el contexto del análisis multivariado de varianza son las variables dependientes.

**Introducir independientes juntas.** Es el equivalente del método simultáneo en el análisis de regresión. Todas las variables son utilizadas para realizar el análisis y se introducen de manera simultánea en la ecuación.

**Menor razón de F.** En cada paso se introduce la variable que tiene asociada una mayor razón de  $F$ . Para introducir variables se fija bien un valor de  $F$  para entrar o una probabilidad. Para excluirlas se fija igualmente ya un valor de  $F$  para salir, ya una probabilidad. Las variables que están fuera de la ecuación y tienen un valor de  $F$  mayor que el criterio son introducidas en la ecuación. Las variables dentro de la ecuación que tienen un valor de  $F$  menor que el criterio para salir, son excluidas.

**Método de inclusión por pasos.** Estos métodos son equivalentes a los presentados en el análisis de regresión

múltiple. Las variables se introducen y excluyen secuencialmente en la ecuación discriminante, en función de un criterio (método). En SPSS hay cinco criterios disponibles, que pueden producir resultados diferentes: Lambda de Wilks, V de Rao, Menor razón de F, Varianza no explicada y Distancia de Mahalanobis.

**Lambda de Wilks.** En cada paso se introduce en la ecuación la variable que más reduce la lambda de Wilks, esto es, que más contribuye a reducir la proporción de variabilidad total explicada por el error.

**V de Rao.** Es la suma de las proporciones entre suma de cuadrados de efecto y suma de cuadrados de error. En cada paso se introduce la variable que más incrementa esa suma, es decir, que más contribuye a incrementar la proporción efecto/error.

**Variable de agrupación.** Variable categórica, con dos o más niveles, que define los diferentes grupos de sujetos. Es la variable predicha en el análisis discriminante, pero funciona como variable independiente en el análisis multivariado de varianza.

**Varianza no explicada.** La varianza no explicada es la que no puede ser atribuida al efecto. En cada paso se introduce la variable que más reduce la cantidad de varianza no explicada.

**Distancia de Cook.** Se emplea para evaluar el cambio que se obtendría en los parámetros cuando se elimina un caso. Por tanto, indica la influencia que cada caso tiene en los resultados del análisis, y puede usarse como indicador de si el caso es extremo. Se calcula como

$$D_i = \frac{(\beta_i - \bar{\beta})(X'X)(\beta_i - \bar{\beta})}{(p+1)s^2}$$

donde  $i$  designa al sujeto que se elimina,

los betas son los coeficientes de regresión,  $s^2$  es la varianza residual y  $p$  el número de variables predictoras.

**Durbin-Watson.** Se emplea para comprobar la existencia de correlaciones significativas entre los residuos. Asume que los residuos han sido obtenidos en serie a través del tiempo. Compara el residuo en el período  $t$  ( $e_t = Y_t - \bar{Y}$ ) con el obtenido en el período  $t-1$  ( $e_{t-1}$ ). El estadístico se computa como:

$$DW = \frac{\sum_{t=2}^N (e_t - e_{t-1})^2}{\sum_{t=1}^N e_t^2}$$

La hipótesis nula que se contrasta es que no hay correlación serial de primer orden (de un residuo con el siguiente).

**Estadísticos de influencia.** Se emplean para evaluar la importancia que tiene un caso en los coeficientes del modelo y su ajuste. SPSS crea nuevas variables en el fichero, puesto que se computa cada índice para cada uno de los sujetos.

**Diferencia en el valor de Beta (DfBeta).** Cuando se excluye un sujeto del análisis el valor de los coeficientes de regresión puede cambiar respecto del calculado con todos los sujetos.

**DfAjuste tipificado.** Puede usarse en el mismo sentido que DfBeta tipificada. Es la diferencia en el pronóstico dividida por la desviación típica.

**Razón entre covarianzas.** Es similar a los índices anteriores, excepto en que ahora se trabaja con los determinantes de la matriz de covarianzas con todos los sujetos y con el sujeto excluido. Si el sujeto no influye, la razón entre ambas debe estar próxima a 1.

## Glosario

**DfBeta tipificada.** La diferencia en el valor de beta dividida por la desviación típica. Valores típicos más allá de  $\pm 2/\sqrt{N}$  son indicadores de que el sujeto contribuye mucho al valor de beta global, y probablemente debe ser excluido.

**Diferencia en el pronóstico (DfAjuste).** Para cada sujeto se realiza el pronóstico con todos los sujetos incluidos y con el propio sujeto excluido. DfAjuste es la diferencia entre ambos.

**Estado.** En análisis de supervivencia es la variable que contiene si el evento de interés ha ocurrido o no.

**Estimaciones de tamaño del efecto.** La forma más sencilla de medir la magnitud del efecto que produce una manipulación consiste en dividir la suma de cuadrados del efecto por la suma de cuadrados total, para obtener el cuadrado del coeficiente de correlación múltiple ( $R^2 = SC_{EFECHO}/SC_{TOTAL}$ ) o por la suma de las sumas de cuadrados de efecto y de su error correspondiente.

**Factorial, Análisis.** Técnica de reducción de datos en la que se asume que la variabilidad observada en las variables medidas, y las intercorrelaciones entre ellas, son debidas a la existencia de variables latentes, cuya existencia se presume real o matemática (componentes principales). Es, por tanto, un modo de identificar lo que comparten un conjunto de medidas. Puede usarse de forma exploratoria o confirmatoria. Se distinguen varios tipos de factores: comunes, específicos y error.

**Alfa.** Es aplicable cuando las variables medidas son una muestra aleatoria de la población general. Los autovalores representan el índice de generalizabilidad para cada factor.

**Análisis Factorial Exploratorio.** Se realiza para intentar descubrir el número

de variables latentes necesarias para explicar el patrón de correlaciones observado entre las variables medidas.

**Ánálisis Factorial Confirmatorio.** Las variables latentes son conocidas, así como su influencia causal sobre las variables medidas. Es un caso particular de los modelos de ecuaciones estructurales.

**Autovalor.** En el contexto del análisis factorial son una medida de la varianza explicada por cada variable latente.

**Carga factorial.** Peso de cada variable en cada factor. Cuando la carga es alta, la variable realiza una alta contribución al factor.

**Componentes principales.** Método de extracción factorial en el que no se asume que los factores tengan contenido teórico, sino que son simplemente combinaciones lineales de las variables observadas.

**Comunalidad.** La proporción de varianza de cada variable explicada por el conjunto de los factores comunes.

**Ejes principales.** Primero se obtiene una solución de componentes principales asumiendo que las communalidades son 1, después se reestiman las communalidades de las variables, se computa el cambio de communalidades, si el cambio supera un criterio pre establecido, la diagonal se rellena con las nuevas communalidades. Este proceso se repite hasta que el cambio de communalidades es menor que el criterio de convergencia.

**Factores comunes.** Son el objetivo principal del análisis, puesto que influyen más de una variable observada.

**Factores error.** Influencias no sistemáticas sobre las variables. Por ejemplo, el estado de ánimo de los sujetos mientras se les aplican las pruebas.

**Factores específicos.** Factores que influyen sólo sobre una variable medida.

**Factores únicos.** Son la suma de factores específicos y de error.

**Gráfico de sedimentación.** Gráfico empleado como ayuda en la decisión del número de factores que se van a retener. Los factores aparecen representados en el eje de abcisas y los autovalores asociados en la ordenada.

**Imagen.** La communalidad de una variable es el cuadrado del coeficiente de correlación múltiple entre esa variable y el resto de variables medidas.

**Matriz de configuración.** Contiene los coeficientes de configuración de cada variable correspondientes a cada factor. Los cuadrados de los coeficientes de configuración indican la cantidad de varianza de las variables observadas que es explicada por un factor.

**Matriz de correlaciones.** La matriz que contiene las correlaciones entre todos los pares posibles de variables medidas. El examen de esta matriz es fundamental para determinar si el análisis factorial es aplicable. Se realiza a través de diferentes medidas: Determinante, KMO, prueba de esfericidad de Bartlett y Anti-imagen.

**Anti-imagen.** Los negativos de los coeficientes de correlación parcial son los coeficientes anti-imagen. La matriz anti-imagen debe tener la mayoría de los elementos de bajo valor para que el análisis factorial sea adecuado.

**Determinante.** El determinante de la matriz de correlaciones indica la magnitud de las correlaciones entre variables. El análisis factorial es adecuado si el determinante es bajo (las correlaciones son altas).

**KMO.** El índice de Kayser-Meyer-Olkin compara las correlaciones entre variables con las correlaciones parciales. El análisis es adecuado cuando las correlaciones parciales son bajas y el índice se aproxima a la unidad.

**Prueba de esfericidad de Bartlett.** Una matriz de correlaciones es esférica

cuando es una matriz identidad, esto es, todos los elementos fuera de su diagonal principal son nulos. El análisis factorial es recomendable cuando la matriz de correlaciones no sigue un patrón esférico.

**Matriz de estructura.** Contiene las cargas factoriales de cada variable en cada factor. Los cuadrados de las cargas miden la varianza de las variables observadas que es explicada por cada factor, pero también incluyen efectos asociados a la interacción de factores.

**Matriz Reproducida.** La matriz estimada una vez que se ha realizado la extracción factorial. En un análisis adecuado esta matriz debe diferir poco de la matriz de correlaciones observada.

**Máxima verosimilitud.** Las communalidades de las variables se estiman de modo que se haga máxima la probabilidad de observar un patrón de correlaciones como el obtenido entre las variables medidas.

**Mínimos cuadrados generalizados.** El objetivo es el mismo que en los mínimos cuadrados no ponderados, sin embargo, las variables que comparten varianza con otras reciben más peso, tienen mayor importancia en la solución, que las que comparten poca varianza, tienen mucha varianza única.

**Mínimos cuadrados no ponderados.** Las communalidades no se estiman, sino que son una consecuencia del cómputo, cuyo objetivo es obtener una matriz reproducida de correlaciones cuyas diferencias cuadradas con respecto a la matriz observada sea mínima.

**Puntuaciones factoriales.** La suma ponderada de las variables que pesan en un factor. La ponderación que debe aplicarse a cada variable puede determinarse mediante diferentes métodos: Anderson-Rubin, Bartlett y Regresión.

**Método Anderson-Rubin.** Los coeficientes son estimados por un proce-

## Glosario

dimiento de mínimos cuadrados ponderados.

**Método de Bartlett.** Los coeficientes son estimados por un procedimiento de mínimos cuadrados ponderados.

**Método de Regresión.** Usa como coeficientes los de la matriz patrón.

**Rotación.** La interpretación de la solución factorial no rotada suele ser compleja. Una manera de simplificarla consiste en rotar los ejes factoriales. Hay varios tipos de rotación y diferentes criterios. La rotación se realiza multiplicando la matriz factorial (que contiene las cargas factoriales) por la matriz de transformación (que contiene la rotación que ha de aplicarse a los factores).

**Rotación oblicua.** Se asume que los factores están correlacionados entre sí, por lo que cada eje puede rotarse de manera diferente que los demás.

**Oblimin directo.** Método de rotación oblicua en el que se tiene control sobre el ángulo entre los factores. Las soluciones son las más oblicuas posible cuando delta es cero y a medida que se hace más negativo los factores son cada vez menos oblicuos.

**Promax.** Se aplica realizando primero una rotación Varimax, y computando después una matriz en la que cada elemento de la original es elevado a una potencia y dividido por el mayor de los elementos de la columna. De este modo los pesos mayores tienden a la unidad y los menores a cero.

**Rotación ortogonal.** Los ejes factoriales son ortogonales entre sí, esto implica que se asume que los factores no correlacionan entre ellos. Cada eje es rotado de la misma forma que los demás.

**Cuartimax.** Su objetivo es simplificar la interpretación de las variables medidas. Es un método de rotación ortogonal en el que se hace mínimo el número de factores necesarios para explicar cada variable medida.

**Ecuamax.** Método de rotación ortogonal en el que se minimiza simultáneamente el número de variables y el de factores necesarios. Es una combinación entre los métodos Varimax y Cuartimax.

**Varimax.** Es el procedimiento idóneo cuando se pretende realizar una rotación ortogonal en la que se haga mínimo el número de variables que tienen saturaciones altas en cada factor y se simplifique la interpretación de los factores.

**F de entrada y de salida.** En los análisis que pueden realizarse incluyendo o excluyendo variables por pasos pueden adoptarse varios criterios de decisión, entre los que destaca por su frecuencia de uso la F de Snedecor. La F de entrada es la F de una variable que no está incluida todavía en el modelo. Si su probabilidad es menor o igual que un criterio, la variable es incluida en el modelo. La F de salida es la F de una variable que está ya dentro del modelo. Si su probabilidad es mayor que un criterio (lo que implica que la variable explica menos de lo requerido) es eliminada de la ecuación. Es lógico, pues, que la probabilidad de la F de entrada sea menor que la de salida. Puede emplearse el valor absoluto de F en lugar de la probabilidad.

## Introducir variables de forma no simultánea

**Eliminar.** Las variables se eliminan de la ecuación por bloques (por ejemplo, todas las variables de neuroticismo medidas, todas las variables de psicotísmo medidas, etc.).

**Hacia atrás.** En el primer paso se introducen todas las variables en la ecuación de predicción. En cada paso siguiente se elimina la variable que mayor probabilidad de F para salir tenga, siempre que sea mayor que el criterio. Las variables

excluidas de la ecuación no pueden volver a entrar en ella.

**Hacia delante.** En cada paso se introduce la variable que menor probabilidad de F para entrar tenga. Las variables ya introducidas en la ecuación no son eliminadas.

**Introducir (simultáneo).** Método de introducción de variables en el que todos los predictores se incluyen a la vez, con independencia de cuál sea su importancia para hacer la predicción.

**Pasos sucesivos (por etapas).** Las variables se introducen secuencialmente en la ecuación de predicción. En cada etapa o paso se introduce la variable que mayor probabilidad de F para entrar tenga, siempre que sea menor que un criterio, esto es, que más contribuya a mejorar la predicción. Además se elimina la variable cuya probabilidad de F para salir sea mayor que el criterio de salida.

**Log-lineal.** En el contexto del análisis de tablas de contingencia se emplea para decidir qué cantidad mínima de efectos es necesaria para explicar las frecuencias observadas en la tabla.

**Delta.** En los modelos saturados a las frecuencias de cada casilla se les añade un valor, 0,5 por defecto en SPSS. Este valor puede cambiarse por cualquier otro entre 0 y 1.

**Modelo personalizado.** Los modelos saturados no suelen ser de interés, puesto que se sabe de antemano su poder explicativo. Los modelos de interés son aquéllos en los que solamente se tienen en cuenta algunos de los efectos.

**Modelo saturado.** Un modelo loglineal en el que se incluyen todos los efectos principales e interacciones. El modelo saturado siempre explica perfectamente las frecuencias observadas.

**Matriz de covarianza.** La matriz cuadrada que contiene las varianzas de cada variable en la diagonal principal (esquina superior izquierda a esquina inferior derecha) y las covarianzas fuera de la diagonal.

**Ponderación MCP.** En el contexto del análisis de regresión la ponderación MCP se aplica para tener en cuenta la heteroscedasticidad. Los pesos dependen del tipo de heteroscedasticidad. En SPSS los pesos son el inverso de la varianza, lo que implica que las observaciones con gran varianza tienen menor importancia que las que tienen varianza pequeña.

**Potencia observada.** La potencia observada se calcula una vez que han sido obtenidos los datos, es potencia *a posteriori*. Es la probabilidad de rechazar una hipótesis nula falsa.

**Regresión logística.** En muchas investigaciones la variable dependiente es categórica (dicotómica o politérmica). La regresión logística permite predecir la probabilidad de aparición de cada una de las categorías a partir de un conjunto de predictores que pueden ser métricos o categóricos.

**Bondad de ajuste de Hosmer-Lemeshow.** El test divide las puntuaciones en diez grupos (deciles) en orden creciente de probabilidad. La comparación entre las frecuencias observadas y las esperadas se usa para calcular la bondad de ajuste del modelo.

**Contraste de la razón de verosimilitud.** Se emplea para comprobar si los parámetros de los efectos incluidos en el modelo son significativos (diferentes de 0) o no.

**Desviación.** Los residuos de desviación. Son fáciles de calcular manualmente cuando la variable es dicotómica. Si la variable es 1, el residuo es:

## Glosario

$$+ \sqrt{-2 \ln(p)},$$

donde p es la probabilidad pronosticada por el modelo. Si la variable es 0, el residuo es:

$$- \sqrt{-2 \ln(p)}$$

**Estadístico de la bondad de ajuste chi-cuadrado.** Indica si el ajuste global del modelo es o no significativo. No sirve para decidir la importancia de cada efecto.

**Grupo de pertenencia.** La categoría o grupo pronosticado para el sujeto a partir del modelo.

**IC para exp(B).** Intervalo de confianza para  $e^B$ .

**Logit.** Residuos cuando se ha empleado la escala logit. El logit de una probabilidad p (la probabilidad de que ocurra un evento) es el logaritmo del cociente entre la probabilidad y su complementario:

$$\text{logit}(p) = \ln \frac{p}{1-p}$$

**Modelo de regresión logística multinomial.** En la regresión logística politérmica SPSS permite especificar un modelo que incluya sólo los efectos principales de las variables predictivas, los efectos principales y las interacciones (factorial completo) o un modelo definido por el usuario (personalizado).

**Probabilidades.** Las probabilidades de aparición de las categorías predichas por el modelo.

**Residuos.** Las diferencias entre el valor observado y el pronosticado para un sujeto. SPSS permite guardar residuos no tipificados, tipificados, estandarizados, logit y de desviación. Los tres primeros han sido descritos en el contexto del análisis de regresión lineal.

**Test de Wald.** Se emplea para decidir si los parámetros de los efectos incluidos en el modelo son significativos. Es el cociente entre el parámetro no estandarizado y su error típico.

**Regresión múltiple.** Permite predecir una variable métrica a partir de un conjunto de predictores también métricos, cada uno de los cuales tendrá asociado un peso en la predicción, su coeficiente de regresión. Los predictores pueden introducirse en la ecuación de predicción de varias formas: simultáneamente o secuencialmente.

**Residuos.** Los errores de predicción del modelo. Pueden almacenarse de tres maneras diferentes: no tipificados, tipificados, estandarizados, eliminados y eliminados estandarizados.

**Residuos eliminados.** El residuo de un caso cuando ese caso ha sido excluido del cálculo de los coeficientes de regresión. Se realiza el cálculo eliminando, uno por uno, a todos los sujetos del estudio.

**Residuos eliminados estandarizados.** Los residuos eliminados divididos por su error típico. Si un caso tiene poca importancia en la determinación de los coeficientes estos residuos deben ser iguales que los no eliminados estandarizados.

**Residuos estandarizados.** Son semejantes a los tipificados, aunque siguen la distribución t de Student. La inspección visual de una gráfica de residuos respecto de puntuaciones observadas puede ayudar también a identificar los puntos extremos.

**Residuos no tipificados.** La puntuación observada menos la predicha por el modelo.

**Residuos tipificados.** Residuos divididos por su error típico. Su media es 0 y su desviación típica 1. Son importantes para evaluar los puntos extremos.

**Tiempo.** En análisis de supervivencia, la variable que contiene el tiempo de permanencia de los sujetos en el estudio, hasta que ha ocurrido el evento de interés y lo han abandonado.

**Valores de influencia.** Expresan la importancia que un sujeto tiene en los coeficientes de regresión. La mínima es 0 y la máxima se aproxima a 1 cuando el número de sujetos tiende a infinito.

**Valores pronosticados.** Son las puntuaciones predichas por un modelo. Pueden guardarse de forma directa, o transformada. Las transformaciones más frecuentes son la tipificada y la corregida.

**Valores pronosticados corregidos.** A cada sujeto se le pronostica un valor excluyéndolo del análisis, lo que permite reducir el sesgo en el cálculo de los parámetros.

**Valores pronosticados tipificados.** A cada valor pronosticado se le sustrae la media de los todos los valores pronosticados y se divide por la desviación típica de los mismos. La tipificación permite identificar de manera más fácil puntuaciones extremas.

**Variable endógena.** Variable que es afectada por otras variables latentes. En sentido estricto es una variable cuyos valores están determinados por otras variables dentro del sistema.

**Variable exógena.** Variable que no es afectada por variables latentes, y que afecta a variables observadas.

**Variable latente.** Es una variable no observada, sino inferida. Se supone que afecta a variables observadas.

## ÍNDICE DE MATERIAS

- Agrupamiento, métodos de, 172-186.  
    Jerárquico, 172-182, 186.  
    K-Medias, 183-186.  
    Partición, 183, 187.
- Akaike, Criterio de información de, 373.
- Análisis de varianza  
    ANOVA/ANCOVA, 47-95, 258-260.  
    Efectos de interacción, 75-77.  
        Efectos simples, 83-85, 281.  
        Interacción parcial, 84-85.  
    Efectos principales, 74.  
    F de Snédecor, 52, 55, 56, 58, 230.  
    Grados de libertad, 50, 54, 67.  
    Greenhouse-Geisser, 71.  
        Huynh-Feldt, 71-72.  
    Magnitud de los efectos, 60, 276.  
    MANOVA/MANCOVA, 28, 241, 256, 304.  
    Potencia del contraste, 61, 273.  
    Supuestos  
        Aditividad, 63.  
        Esfericidad, 70, 258.  
        Homogeneidad de varianzas, 62, 246, 294, 326.  
        Independencia de los errores, 63, 248, 294.  
        Normalidad, 62, 249, 296.  
    t de Student, 49, 52.
- Asociación, medidas de, 170-171.
- Azar, proporción de, 212.
- Bartlett, test de, 131, 199.
- Bondad de ajuste  
    AGFI, 369.  
    Chi-cuadrado, 70, 368.  
    GFI, 368.  
    MacDonald, 370.  
    Parámetro de no centralidad reescalado, 370.  
    RGFI, 369.  
    Tucker-Lewis, 371.
- Cambio esperado del parámetro, 374.
- Clasificación, función de, 305, 319.
- Coeficiente de determinación total, 373.
- Comparaciones (Contrastes)  
    No planeados/*A posteriori*, 59-60.  
    Planeados/*A priori*, 56-59, 83, 277.
- Componente principal, 122, 132.
- Comunalidad, 139, 372.
- Congénérico, modelo, 363.
- Contexto de Investigación, 23-26.  
    Experimental, 23-25.  
    No Experimental, 25-26.
- Contraste de hipótesis, 37-38.  
    Multivariado, 38, 279.  
        Tasa de error tipo I, 62-63.  
        Tasa de error tipo II, 62-63, 273.  
    Univariado, 37, 56.
- Correlación, 41-42, 129, 193.
- Correlación canónica, 193, 195, 197.
- Covariable, 214.
- Covarianza, 28, 68, 69.
- Cox, regresión de, 214.
- Cruzada, validación, 324, 375.

Dendrograma, 177.  
Discriminante, función, 269, 305, 310.

Diseño experimental  
Completamente aleatorizado, 52.  
De covarianza, 86-88, 290.  
Factorial  
Entre grupos, 74, 80, 263, 280.  
De Medidas repetidas, 80, 287.  
Mixto, 81, 289.  
Medidas repetidas, 66, 284.  
Multivariado, 256.  
Unifactorial, 53.  
Univariado, 52.  
Distancia, medidas de, 165-169.

Ecuaciones estructurales, 361.  
Escala de medida, 34-35.  
Esfericidad, test de Bartlett, 131.  
Estructural, modelo, 363.

F para entrar, 234.  
F para salir, 234.  
Factor, 122, 128.  
Común, 123-125.  
Específico, 123-125.  
Único, 123-125.  
Factores principales, 150.  
Factorización alfa, 153.

Hosmer-Lemeshow, bondad de ajuste, 341.

Igualación, medidas de, 169-171.  
Imagen, 153.  
Infradeterminado, modelo, 360, 380.

Kaiser-Meyer-Olkin, 130.  
Kaplan-Meier, 213.  
Kolmogorov-Smirnov, 62.

Lagrange, multiplicadores de, 374.  
Lambda de Wilks, 271, 274.  
Latente, variable, 122, 201, 356.  
Linealidad, 154, 245, 296, 326, 379.  
Logística, función, 333.  
Logit, función, 333, 348.

Mahalanobis, distancia de, 168.  
Matriz, 35, 38.  
Anti-imagen, 131.  
Correlación, 41-42, 129, 197.  
Covarianza, 41-42, 69, 129.  
De configuración, 147-148.

De estructura, 147-148, 315.  
De transformación, 142, 147.  
Diagonal, 69.  
Operaciones con matrices, 39, 41.  
Singular, 155.  
Sumas de cuadrados y productos cruzados, 40.  
Valores propios, 133-137, 146, 267.  
Vectores propios, 133-137, 267.  
Máxima Verosimilitud, 152, 366.  
Mínimos cuadrados, 152, 366.  
Medida, modelo de  
Endógeno, 364.  
Exógeno, 364.  
Modelos logarítmico-lineales, 101.  
Ajuste del modelo, 114-116.  
Asociación  
Marginal, 103-104.  
Parcial, 103-104.  
Efectos  
Incompleto, 106.  
Interacciones, 103.  
Jerárquico, 106-107.  
Modelo de predicción, 105.  
Principales, 102.  
Saturado, 106.  
Parámetros, 105, 112.  
Razón de verosimilitud, 102.  
Modificación, índice de, 374.  
Mortalidad, tablas de, 213-214.  
Multicolinealidad, 251, 297, 349, 380.  
  
Normalidad, 62, 154.  
Multivariada, 296, 326, 379.  
  
Partición de variabilidad, 229, 266.  
Pseudo R cuadrado  
Cox-Snell, 340.  
Nagelkerke, 340.  
Puntos extremos, 154, 250, 298, 327, 350.  
Puntuación factorial, 127, 148-149.  
  
R cuadrado, 60-61, 180-182, 230.  
Ajustada, 230.  
Redundancia, índice de, 202-203.  
Regresión  
Lineal, 224, 225.  
Logística, 332, 343, 345.  
Múltiple, 224, 226, 231.  
Por etapas, 234-237.  
Secuencial, 232, 240.  
Simultánea, 232, 239.

## Índice de materias

Rotación, 126, 140-145, 158.  
Oblicua, 126, 143-145.  
Ortogonal, 126, 141-143.  
Roy, Mayor raíz de, 271, 274.  
  
Semejanza, medidas de, 165-171.  
Shapiro-Wilk, 62.  
Superficial, variable, 122, 356.  
Supervivencia, funciones de, 211.  
  
t de Student, 49, 52.  
T<sup>2</sup> de Hotelling, 277.

Témpanos, diagrama de, 178.  
Tolerancia, 251, 297.  
Traza  
De Hotelling-Lawley, 272, 274.  
De la matriz, 70.  
De Pillai, 271, 275, 276.  
  
Validación, 188, 324, 346, 375.  
Valores de influencia, 342.  
Variado, 36, 194, 196, 197, 262, 267.  
  
Wald, test de, 335, 338.

